The New Encyclopædia Britannica

Volume 21

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

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

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



THE UNIVERSITY OF CHICAGO

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

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

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

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

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

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

© 1991
By Encyclopædia Britannica, Inc.

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

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

Printed in U.S.A.

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

CONTENTS

1 India
164 Indian Ocean Islands
180 Indian Philosophy
201A Indonesia
202 Industrial Engineering and Production Management
231 Industrial Glass and Ceramics
251 Chemical Process Industries
334 Extraction and Processing Industries
467 Manufacturing Industries
508 Textile Industries
527 Infectious Diseases
552 Information Processing and Information Systems
568 Information Theory
575 Inheritance and Succession
585 Insects
678 Insurance
691 Integumentary Systems
710 Human Intelligence
717 Intelligence and Counterintelligence
724 International Law
732 20th-Century International Relations
824 International Trade
854 Iran
899 Iraq
951 Ireland

India

India (also known constitutionally in Hindi as Bhārat Varsha, or Bhāratavarsha, "Land of Bharat," after a legendary monarch and sage) is a republic in southern Asia formed of a union of states and territories. It is the seventh largest and the second most populous country in the world. It covers an area of 1,269,419 square miles (3,287,782 square kilometres), with a land frontier of 9,425 miles (15,168 kilometres) and a coastline of 3,533 miles. (Before partition and independence in 1947, it covered 1,575,107 square miles.) It has an annual population growth of around 13,000,000; the birth bulge consequently constitutes one of the country's most serious problems.

Physically, India is cut off from the greater part of Asia. The mountain ranges of the Himalayas to the north have virtually isolated the country from two of the mainstreams of Asian history—the civilization of China and the Islāmic world of the Middle East. India can claim to represent a third distinct stream of Asian history, going back to the Indus Valley civilization of Mohenjo-daro, which dates to around 2500 BC. Ancient artifacts recovered at the site of Mohenjo-daro (located in the Lārkāna district of the Pakistan province of Sind) suggest that the religion of the Indus people was the lineal progenitor of the Hindu religion. This would make the origins of Hinduism almost coincidental with the beginnings of Indian civilization.

India's neighbours to the north are the Chinese Tibetan Autonomous Region, the independent states of Nepal and Bhutan, and the Sinkiang Uighur Autonomous Region of China. To the northwest, India shares a boundary with Pakistan. To the east, India is bounded by Burma, while the territory of Bangladesh forms an enclave bounded by the Indian states of West Bengal, Assam, Meghālaya, Tripura, and Mizorām and by the Bay of Bengal. To the south, India tapers like a triangle, with its apex pointing downward. At the Tropic of Cancer the landmass of the country thins between the Bay of Bengal on the east and the Arabian Sea on the west. The southern tip of India, Cape Comorin, is washed by the Indian Ocean. Off the cape, the Gulf of Mannār and Palk Strait separate India from Sri Lanka (Ceylon). The Andaman and Nicobar Islands in the Bay of Bengal and the Lakshadweep Islands in the Arabian Sea are parts of Indian territory.

India is more a subcontinent than a country. Its regional diversities are reflected in its myriad races, languages, and religions, in the differing dress, manners, ways of life and thought, and in its wide disparities in education and illiteracy, in poverty and wealth. The overall literacy rate at the beginning of the present decade averaged barely 36 percent of the population. Since 1960, per capita real income has tended to decline, though the net national income has increased—a paradox that is largely explained by the intimidating rate of population growth.

Within the country's socio-religious framework, strong divisive influences such as caste, the status of untouchability, and linguistic chauvinism operate. The wonder is that, despite these limitations, India still functions as the world's largest working democracy.

The paradox of India lies in its political survival, despite the age-old maze of social and economic contradictions that it has inherited. Well over 80 percent of the country's total population are Hindus, and of the Hindu population some 60 percent compose orthodox Hindu society. Despite divisive internal pulls and urges, Hinduism is the one unifying factor that has kept the large mass of the peoples of India together. The only common link between the Dravidian Hindus of southern India and the Aryan Hindus of the north is, in fact, Hinduism.

Two great Indian nationalists—Mahatma Gandhi and Jawaharlal Nehru—exerted an immense influence on contemporary India. By encouraging women to participate actively in his nonviolent civil-disobedience movements, Mahatma Gandhi helped to emancipate them and to draw them into the mainstream of political and public life; one consequence has been that India has been able to accept a woman prime minister, Indira Gandhi.

This article contains the following sections:

Physical and human geography 2
  The land 2
    Relief
    Drainage
    Soils
    Climate
    Plant and animal life
    Settlement patterns
  The people 10
    Ethnolinguistic groups
    Religions
    Population
  The economy 13
    Resources
    Agriculture
    Mining and quarrying
    Industry
    Energy
    Finance
    Trade
    Management of the economy
    Transportation
  Administrative and social conditions 18
    Government
    Education
    Health and welfare
    Police
    Cultural life
History 21
  India from the Late Stone Age to the decline of the Indus civilization (1750 BC) 22
    The early prehistoric period
    The Indus civilization
  The development of Indian civilization from c. 1500 BC to c. AD 1200 30
    c. 1500–c. 600 BC
    The beginning of the historical period, c. 600–150 BC
    150 BC–AD 300
    AD 300–750
    750–c. 1200
  The early Muslim period 50
    North India under Muslim hegemony, c. 1200–1526
    The Muslim states of South India, c. 1350–1680
    The Vijayanagar empire, 1336–1646
  The Mughal Empire, 1526–1761 63
    The establishment of the Mughal Empire
    Akbar the Great
    The empire in the 17th century
    Mughal decline in the 18th century
  The Marāṭhās 71
    Emergence of the Marāṭhā Empire
    The Marāṭhā Confederacy
  India and European expansion, c. 1500–1858 77
    European activity in India, 1498–c. 1760
    The extension of British power, 1760–1856
    The mutiny and great revolt of 1857–59
  British imperial power, 1858–1920 93
    Climax of the Rāj
    Foreign policy
    Indian nationalism and the British response
    World War I and its aftermath
  Prelude to independence 102
    Indian nationalism and the end of the Rāj
    The transfer of power
  The Republic of India 107
    Nehru's premiership
    Indira Gandhi's premierships
North India 110
  Chandīgarh 110

Delhi 110
Haryāna 111
Himāchal Pradesh 112
Jammu and Kashmir 113
Punjab 116
Rājasthān 118
Uttar Pradesh 121
Northeast India 123
Arunāchal Pradesh 123
Assam 123
Bihār 126
Manipur 129
Meghālaya 129
Mizorām 130
Nāgāland 131
Orissa 133
Sikkim 135
Tripura 138
West Bengal 139
West Central India 141
Gujarāt 141
Madhya Pradesh 143
Mahārāshtra 145
South and Insular India 147
Andaman and Nicobar Islands 147
Andaman Islands
Nicobar Islands
Andhra Pradesh 149
Dādra and Nagar Haveli 151
Daman and Diu 152
Goa 152
Karnātaka 153
Kerala 154
Lakshadweep 156
Pondicherry 156
Tamil Nādu 157
Bibliography 159

PHYSICAL AND HUMAN GEOGRAPHY

The land

RELIEF

The three mainland regions

The Indian mainland comprises three broad well-defined regions. These are, first, the mountain wall of the Himalayas and associated mountains, which stretch across northern India from west to east in an irregular crescent; second, the Indo-Gangetic Plain, 1,500 miles long and 150 to 200 miles wide, which is formed by the basins of the Indian subcontinent's three great rivers—the Indus, the Ganges, and the Brahmaputra; and, third, the peninsula, or Deccan Plateau, separated from the Indo-Gangetic Plain by a mass of mountain and hill ranges varying from 1,500 to 4,000 feet (450 to 1,200 metres) in height.

While three major landforms—mountains, plains, and plateaus—are thus represented in India by distinct tracts of country, there are many topographical variations. Within the Himalayas, for example, is the Vale of Kashmir (divided between India and Pakistan), which has a plains topography.

While the geological placement of India was for long one of the puzzles of geology, it had become increasingly accepted by the late 20th century that the explanation might lie in the drifting of continental masses over the surface of the world after the breakup of the ancient continent of Gondwanaland. One of these masses—peninsular India—would, according to this theory, have collided with the southern shore of the Asian landmass, causing the uplifting of the Himalayan mountain ranges. The erosive action of rivers flowing down from the Himalayas would then, in subsequent eras, have resulted in the creation of the Indo-Gangetic Plain, which today is one of the world's longest stretches of alluvium and also one of the most densely populated areas in the world.

The Himalayas. The Himalayas constitute the highest mountain system of the world, containing most of the world's highest peaks, some ten of which rise above 25,000 feet. Mount Everest (29,028 feet [8,848 metres]), which dominates the Himalayas, also extends into Nepal and Tibet. Geologically, the Himalayas constitute the world's youngest and longest east–west mountain system, extending almost without interruption for more than 1,500 miles.

The Himalayas have variegated landscapes, consisting of snow-clad peaks, sprawling ice fields above the heads of large valley glaciers, waterfalls tumbling over high precipices, deep river gorges, and wide valleys. Such scenes are well represented in Jammu and Kashmir and in Himāchal Pradesh, as well as in the northern regions of Uttar Pradesh, West Bengal, and Assam. The valleys of Kulu and Kāngra in Himāchal Pradesh are popular vacation resorts. In Uttar Pradesh are the important peaks of the Himalayas—Nanda Devi (25,646 feet [7,817 metres]), Rāmet (25,446 feet [7,756 metres]), and Trisūl (23,359 feet [7,120 metres]).

The Himalayas contain some of the largest mountain glaciers in the world, giving India the world's largest area under snow and glaciers outside subpolar regions. Extending from Kashmir in the west to Assam in the east, the snowfields cover about 15,500 square miles. The largest glaciers occur in the two highest ranges in the Himalayan system, the Great Himalaya and the Karakoram Himalaya. At their lowest limits these glaciers melt and supply water to the rivers of northern India.

Glaciers and snowfields

Plains. The plains of India cover more than 400,000 square miles, the greater part of which consists of the Indo-Gangetic Plain in the north. There is also an extensive slice of plains land in the south, mainly deltaic in character. From the standpoint of human use, the plains of India are the most important type of landform in the country. They represent the most extensively cultivated and densely populated areas, particularly where water is available. The alluvial plains covering the greater part of West Bengal, Bihār, Uttar Pradesh, Haryāna, and Punjab are most extensive in North India. They are also to be found in Assam and Rājasthān. In the east coast region, from Cape Comorin in the south to the Mahānadi Delta, the alluvial plains stretch across the deltas of India's southern rivers.

Plateaus. Geologically, the peninsular plateau is the oldest part of India's land surface. It consists largely of ancient crystalline rocks in various stages of change and deformation. The rocks generally range in age from 300,000,000 to 500,000,000 years. Until the Koyna Nagar (Mahārāshtra) earthquake in 1967, the peninsula was believed to be a region of relative stability that was rarely subject to seismic disturbances. Some geologists hold that the Koyna Nagar earthquake originated from the so-called Malabar fault, which extends from Kutch on the west coast to Cape Comorin in the south and with which is associated a series of parallel fractures; not all geologists, however, agree with this explanation.

The Deccan Plateau

The peninsula, or Deccan Plateau, is bordered to the north by the Vindhya Mountains, which separate it from the Indo-Gangetic Plain; other prominent mountain and hill ranges associated with it are the Arāvalli, Sātpura, Maikala, and Ajanta ranges. The plateau is flanked on the east by the Eastern Ghāts, which have an average elevation of about 2,000 feet. A narrow coastal strip separates the Western Ghāts from the Arabian Sea. Between the Eastern Ghāts and the Bay of Bengal is a broader coastal area. At the southern tip of the plateau, the Eastern and Western Ghāts meet at a point formed by the Nīlgiri Hills.

Plateaus occur in other parts of India as well. Among the most notable, the Ladākh Plateau has an average elevation of almost 17,400 feet and is situated in the northeast part of the Kashmir Himalayas. It is the highest plateau in India and one of the most inaccessible. North of the Chāng Chenmo Range, which runs from west to east and divides Ladākh, the Chāng Chenmo River flows westward into a flat-floored valley. Farther north is a typical interior drainage basin, carved out of limestone and shale, containing some salt lakes with centripetal drainage.

DRAINAGE

Groundwater basins. India contains at least three major groundwater basins. These are the Ganges Basin, which is the largest; the Punjab alluvial basin in the northwest, which runs from Ludhiāna to Amritsar; and the western basin, covering a portion of Rājasthān and curving southward toward the Gujarāt plains as far as Ahmadābād (Ahmedabad). Because of the flat terrain, the flow of groundwater in these basins is not substantial. The underlying structure of the greater part of India is not favourable for storing up a substantial portion of the rains precipitated during the monsoon months.

Conditions for artesian water are most favourable among the porous sandstones of the Himalayan foothills. Another favourable artesian area occurs along the edge of the Narmada Valley to the north of the Sātpura Range in the northwest part of the peninsula; here, water-bearing conglomerate beds (formed of agglomerated fragments of sedimentary rock) are overlain by low ridges of impervious crystalline rocks. Aquifers (water-bearing beds of rock) also occur in such areas as the Mahārāshtra lava plateau on the west of the peninsula and in some artesian zones in Gujarāt in the westernmost part of India.

Groundwater, like the surface streams, also flows in hilly areas where the rocks are pervious. Such underground streams occur near Cherrapunji in the East Khāsi Hills district in Meghālaya and near Dehra Dūn in Uttar Pradesh. In the Rāmgarh Hills in Madhya Pradesh there are natural caves carved out of sandstone. Sinkholes and short, blind valleys are common along the southern fringe of Meghālaya in the east, where the rock is limestone.

Springs and streams appearing on the surface from the groundwater below are to be found mainly in three regions—the Kumāūn Himalaya, in Uttar Pradesh; the low hills and uplands of southern Bihār; and the foot of the Western Ghāts, in the Konkan region. The majority of these are cold springs, but a few are hot springs, either with or without sulfur.

India's lakes

For such a large country, India has very few lakes. Most of the lakes in the Himalayas are basins scooped out by glaciers or dammed by moraines and subsequently filled with water. The Wular Lake, with an area varying from 12 to 100 square miles, is the largest lake in the Kashmir Valley. An example of a crater lake of volcanic origin is the Lonār Lake, in the Buldāna district of Mahārāshtra, in the Deccan. The Sāmbhar Salt Lake in east central Rājasthān is the largest salt lake in India.

On the coastal plains a totally different type of landscape occurs, typified by the backwaters of Bombay harbour, on the Konkan Coast; farther south is Asthamudi Lake.

Rivers. India's rivers may be grouped into the Himalayan rivers, the Deccan rivers, the coastal rivers, and the rivers of the inland drainage basin.

The total available volume of flow from India's rivers is assessed at 2,200,000,000,000 cubic yards per year (1,680,000,000,000 cubic metres per year). The Indo-Gangetic Plain is formed from deposits made by the eastern tributaries of the Indus, by the Ganges and its affluents, and by the Brahmaputra. The east coast deltas are formed from deposits made by the Mahānadi, Godāvari, Krishna, and Cauvery rivers.

The three principal watersheds are the Great Himalaya, with its Karakoram branch, in the north; the Vindhya Range, with the Sātpura–Maikala ranges, in central India; and the Western Ghāts, directing the courses of most of the Deccan rivers. The Himalayan rivers are generally snow fed, as well as rain fed, and therefore have an almost continuous flow. The Deccan rivers are subject to great fluctuations in volume; many of them shrink into streamlets in the hot season. The coastal rivers are short and have limited catchment areas. The few rivers of western Rājasthān mostly drain toward salt lakes or are lost in the sands.

The Ganges Basin receives waters from about a quarter of India's total area, the Godāvari Basin, from about 10 percent. The basins of the Brahmaputra, in the east, and of the Indus within Indian territory are nearly equal in area. The Krishna Basin is the second largest in peninsular India, with the Mahānadi Basin ranking third. The basins of the Narmada and the Cauvery are about equal in area but of different character and shape.

SOILS

India has practically all types of soils, ranging from alluvial soils to arid desert soil, mountain and hill soils, black and red soils, laterite, and swampy lowland soil. With the exception of Himāchal Pradesh, Delhi, and Jammu and Kashmir, alluvial soil is found in almost every state in India. It is most frequently found on the coastal fringes of states like Kerala, Andhra Pradesh, Karhātaka (formerly Mysore), Gujarāt, West Bengal, and Tamil Nādu. It is also found in Assam and Madhya Pradesh. Andhra Pradesh is rich in black soil, on which the farmers depend, and in red soil, which is spread throughout the state. Mahārāshtra is another reservoir of black soil, as also is Manipur. Tamil Nādu has much red soil.

Laterite soil exists in parts of Kerala, Tamil Nādu, Mahārāshtra, and West Bengal. Rājasthān is one of few areas in India where arid desert soil covers the greater part of the state. The eastern strip of Kerala and the central part of Bihār have mountain and hill soils. *Terai* soil is associated with the lowland belt of India, including the region that dips between plain and hills in Uttar Pradesh. Soil charged with peaty, carbonized, and other organic matter exists in some parts of Kerala and Bihār.

CLIMATE

Broadly speaking, the climate of India is governed by the tropical monsoon (rain-bearing winds) regime. The Indian Meteorological Department recognizes four seasons. These are the cold-weather season, from December to March; the hot-weather season, in April and May; the rainy (monsoon) season, from June to September; and the season of the retreating southwest monsoon, from October to November. In effect, however, there are only three noticeable seasons—the cold season, from about November to the end of February; the hot season (when it can be intensely warm, particularly in central and northern India), from about March to June; and the rainy season, from about June until October, by which time the southwest monsoon expends itself.

Three seasons

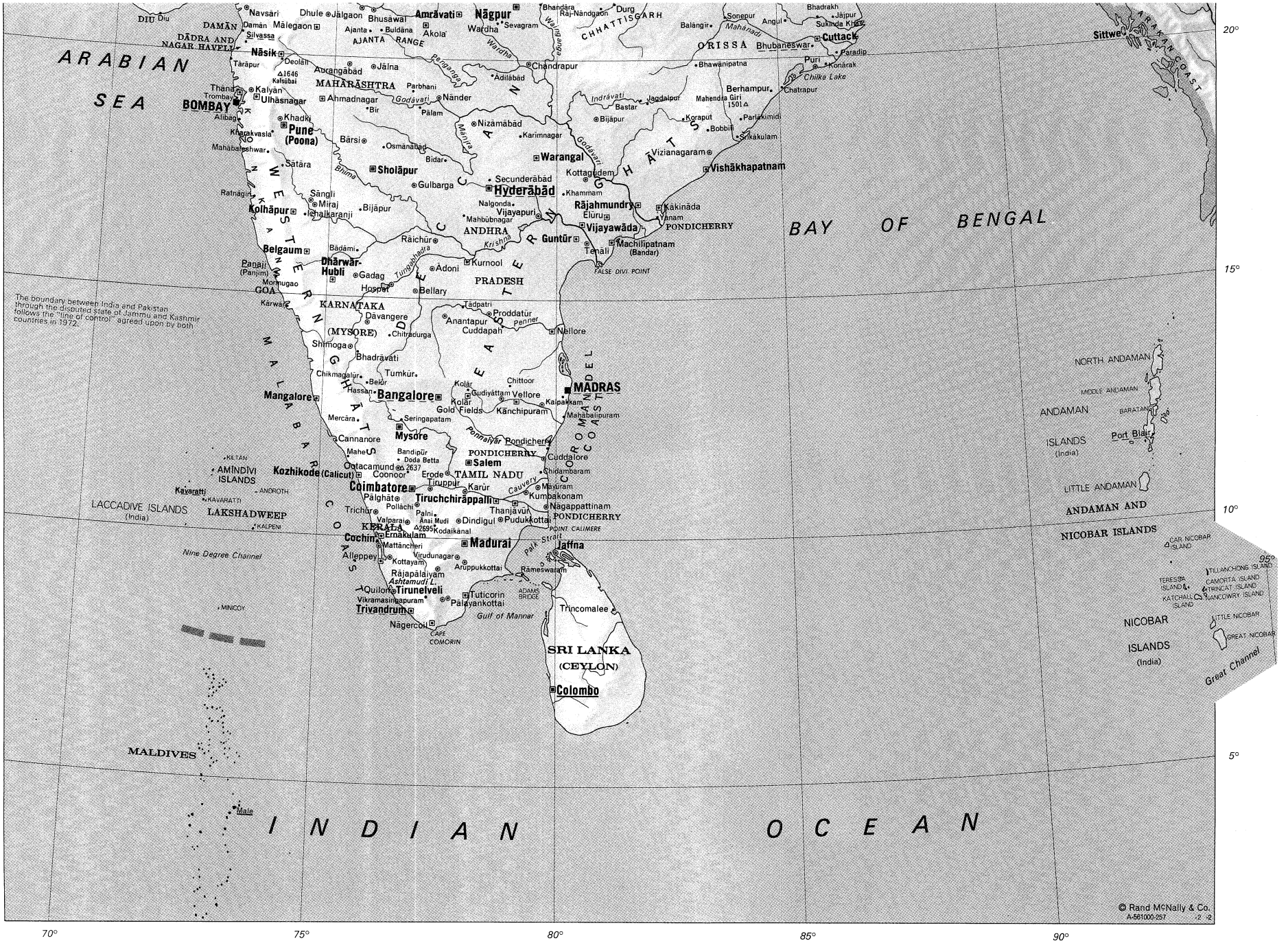
Rainfall and temperature show wide variations in India, which experiences two rain-bearing winds, the northeast monsoon, which blows from December to February, and the southwest monsoon, which blows from June to September.

There are four broad climatic regions based on rainfall. Practically the whole of Assam, Meghālaya, and the west coast of India lying at the foot of the Western Ghāts and extending from the north of Bombay to Trivandrum are areas of very heavy rainfall. Cherrapunji, in Meghālaya, averaged 450 inches (11,430 millimetres) annually over a 74-year period—the largest amount ever recorded in Asia and the second largest in the world. In contrast to these tracts, the Rājasthān desert, extending to Kutch, as well as the high Ladākh Plateau of Kashmir, extending westward to Gilgit, are regions of low precipitation; the Thar Desert, in Rājasthān, for example, receives a mere four inches of rain annually.

In between these regions at the extreme end of the rainfall range are two areas, one of moderately high and one of low rainfall. The former consists of a broad belt in the eastern part of the peninsular plateau, which seasonally moves northward to the Indo-Gangetic Plain and then southward toward the coastal plains; in this belt the annual rainfall averages around 60 inches. The latter constitutes an area extending from the Punjab plains across the Vindhya Mountains into the western part of the Deccan and fanning out into the Karnātaka Plateau; this belt receives about 30 inches of rainfall a year.

Temperature varies as much as rainfall in different parts of India. Hill stations in the Himalayan region, such as Darjeeling and Simla, have the lowest temperatures, with annual averages of between 54° and 57° F (about 12° and 14° C). In the Indo-Gangetic Plain, Delhi and Allahābād register an average of 79° F (about 26° C). In the foothills of the Himalayas, Dehra Dūn has an average of about 71° F (22° C), while Pālayankottai in Tamil Nādu, records


70°
75°
80°
85°
90°
95°
40°
35°
30°
25°
SOVIET UNION
UZBEK S.S.R.
TADZHIK S.S.R.
TURKMEN S.S.R.
KIRGIZ S.S.R.
AFGHANISTAN
PAKISTAN
CHINA
TIBET
SINKIANG
KANSU
TSAIDAM BASIN
TSINGHAI
TARIM BASIN
TAKLA MAKAN
TIEN SHAN
KUNLUN MOUNTAINS
HINDU KUSH
KARAKORAM
HIMALAYAS
Samarkand
Kashgar
Dushanbe
Kābul
Srīnagar
Islāmābād
Rāwalpindi
Jammu
Lahore
Lyallpur
Amritsar
Jullundur
Ludhiāna
Multān
Sukkur
Bīkaner
Jaipur
Jodhpur
Ajmer
Agra
Gwalior
Jhānsi
Kānpur
Lucknow
Allahābād
Varānasi (Benares)
Patna
Gaya
Bareilly
Moradābād
Meerut
DELHI
New Delhi
Ambāla
Patiāla
Sahāranpur
Dehra Dūn
Aligarh
Farrukhābād cum Fatehgarh
Gorakhpur
Lhasa
Kātmāndu
Thimbu
BHUTAN
Darjeeling
Gangtok
Kalimpong
Siliguri
Jalpaiguri
Dacca
Chittagong
CALCUTTA
Howrah
Asansol
Jamshedpur
Khulna
BANGLADESH
BURMA
KARĀCHI
Hyderābād
Ahmadābād
Bhopāl
Indore
Ujjain
Vadodara (Baroda)
Rājkot
Bhaunagar
Jāmnagar
Surat
Bombay
Nāgpur
Jabalpur
Raipur
Bilāspur
Rourkela
Sambalpur
Shillong
Imphāl
Aizawl
Kohima
Dimāpur
Gauhāti
Tezpur
Nowgong
Lakhimpur
Sibsāgar
Dibrugarh
Digboi
Sadiya
Lohit
Subansiri
Brahmaputra
Ganges
Jumna
Indus
Sutlej
Ravi
Chenab
Jhelum
Narmada
Tapti
Godāvari
Mahānadi
Kosi
Gomati
Gandak
Damodar
Son
Betwa
Chambal
Luni
Mahi
Sabarmati
Barda Hills
Aravalli Range
Satpura Range
Vindhyachal
Tropic of Cancer
THAR DESERT (GREAT INDIAN DESERT)
RANN OF KUTCH
Gulf of Kutch
Gulf of Cambay
KATHIAWAR
GUJARAT
RĀJASTHĀN
HARYANA
PUNJAB
CHANDĪGARH
HIMACHAL PRADESH
UTTAR PRADESH
MADHYA PRADESH
BIHAR
WEST BENGAL
ASSAM
MEGHALAYA
KHĀSI HILLS
NĀGĀLAND
MANIPUR
TRIPURA
MIZORAM
ARUNACHAL PRADESH
MISHMI HILLS
PATKAI RANGE
Mount Everest 8848
Kanchenjunga 8598
Nanga Parbat 8126
K2 (Godwin Austen) 8611
Nanda Devi 7817
Dhaulagiri 8172
Annapurna I 8091
Makalu 8481
Cho Oyu 8201
Shishapangma 8013
Gurla Mandhata 7728
Kamet 7756
Chang Chenmo Range
Ladakh Range
Zaskar Range
Pir Panjal Range
Dhauladhar Range
Karakoram Pass 5575
Zoji La 3528
Shipki La 5029
Nathu La 4392
Jelep La 4267
Bangotri 3155
Badrinath 3133
Kedarnath 3581
Gangotri 3140
Yamunotri 3235
Mussoorie
Nainital
Almora
Ranikhet
Hardwar
Rishikesh
Roorkee
Muzaffarnagar
Karnāl
Pānipat
Ambāla
Kurukshetra
Jagādhri
Bhatinda
Sri Gangānagar
Hanumangarh
Ferozpur
Firozpur
Pālampur
Kangra
Mandi
Kulu
Simla
Solan
Nāhan
Chandigarh
Dalhousie
Pathānkot
Batāla
Hoshiārpur
Bilāspur
Jhunjhunu
Churu
Khetri
Nārnaul
Sikar
Alwar
Vrindāvan
Mathura
Dig
Bharatpur
Amber
Jaisalmer
Nāgaur
Sāmbhar
Fatehpur Sikri
Firozabad
Etawah
Bhind
Orai
Hamīrpur
Datia
Jhānsi
Shivpuri
Kota
Bundi
Bhilwāra
Chittorgarh
Udaipur
Nāthdwāra
Rana Pratap Sagar
Jhālawar
Guna
Lalitpur
Khajuraho
Chhatarpur
Panna
Satna
Rewa
Sidhi
Sāsarām
Dehri
Buddh Gaya
Jangipur
Berhampore
Agartala
Aijal
Silchar
Pālanpur
Pātan
Mehsāna
Vijayapuri
Himatnagar
Gāndhīnagar
Dungarpur
Bānswāra
Ratlām
Dewas
Vidisha
Sāgar
Damoh
Murwāra
Jabalpur
Shahdol
Ambikapur
Amarkantak
Narsimhapur
Hoshangābād
Itārsi
Mandla
Pachmarhi
Seoni
Chhindwāra
Betūl
Khandwa
Khargon
Burhānpur
Achalpur
Bālāghāt
Bilāspur
Raigarh
Gondia
Bhilai
Baripada
Balasore
Haldia
Kharagpur
Midnapore
Bankura
Burdwan
Durgapur
Krishnanagar
Nabadwip
Bhatpara
Barrackpur
Barasat
Basirhat
Ranchi
Hazāribāgh
Dhanbad
Bokaro
Chittaranjan
Rānchi
Sindri
Rānchi
Jamshedpur
Jamalpur
Bhāgalpur
Monghyr
Muzaffarpur
Darbhanga
Chapra
Ballia
Arrah
Bihār
Nālanda
Purnea
Katihār
Jalpaiguri
Siliguri
Cooch Behār
Bardeta
Dhubri
Tura
Chirapunji
Sylhet
Comilla
Narāyanganj
Khulna
Jessore
Faridpur
Barisāl
Patuakhali
Sundarbans
Mouths of the Ganges
Andaman Islands
Nicobar Islands
Port Blair
Car Nicobar
Great Nicobar
Little Andaman
Middle Andaman
North Andaman
Ten Degree Channel
Duncan Passage
Coco Channel
Andaman Sea
Bay of Bengal
Arabian Sea
Laccadive Islands
Minicoy
Amindivi Islands
Kavaratti
Agatti
Amini
Kalpeni
Androth
Malpeni
Suheli Par
Minicoy
Kadmat
Kiltan
Chetlat
Bitra
Bangaram
Tinnakara
Pamban
Mannar
Palk Strait
Gulf of Mannar
Adam's Bridge
Rameswaram
Dhanushkodi
Talaimannar
Jaffna
Point Pedro
Kankesanturai
Trincomalee
Batticaloa
Hambantota
Galle
Matara
Colombo
Kandy
Anuradhapura
Polonnaruwa
Negombo
Puttalam
Mannar
Vavuniya
Jaffna
Kayts
Delft
Nedunkerni
Mullaittivu
Kilinochchi
Elephant Pass
Point Pedro
Kankesanturai
Trincomalee
Batticaloa
Hambantota
Galle
Matara
Colombo
Kandy
Anuradhapura
Polonnaruwa
Negombo
Puttalam
Mannar
Vavuniya
Jaffna
Kayts
Delft
Nedunkerni
Mullaittivu
Kilinochchi
Elephant Pass



INDIA

Size of symbol indicates relative size of town • ○ ◎ ⊡ ▣ ■

Elevations in metres

0 100 200 400 600 km

0 100 200 400 mi

MAP INDEX

Political subdivisions

Andaman and Nicobar Islands. 10·30n 93·00e
Andhra Pradesh. 16·00n 79·00e
Arunachal Pradesh. 28·30n 95·00e
Assam. 26·00n 92·00e
Bihār. 25·00n 86·00e
Chandīgarh. 30·45n 76·45e
Dādra and Nagar Haveli. 20·05n 73·00e
Delhi. 28·37n 77·10e
Goa, Daman and Diu. 15·35n 74·00e
20·25n 72·50e
Gujarāt. 22·30n 71·30e
Haryana. 29·20n 76·20e
Himachal Pradesh. 32·00n 77·00e
Jammu and Kashmir. 34·00n 76·00e
Karnataka (Mysore). 14·00n 76·00e
Kerala. 10·00n 76·30e
Lakshadweep. 10·00n 73·00e
Madhya Pradesh. 23·00n 79·00e
Mahārāshtra. 19·00n 76·00e
Manipur. 25·00n 94·00e
Meghalaya. 25·30n 91·15e
Mizoram. 23·30n 93·00e
Nāgāland. 26·00n 95·00e
Orissa. 20·00n 84·00e
Pondicherry. 11·56n 79·50e
Punjab. 31·00n 72·30e
Rājasthān. 27·00n 74·00e
Sikkim. 27·35n 88·35e
Tamil Nadu. 11·00n 78·15e
Tripura. 27·00n 92·00e
Uttar Pradesh. 27·00n 80·00e
West Bengal. 24·00n 88·00e

Cities and towns

Abu Road. 24·29n 72·47e
Achalpur. 21·16n 77·31e
Ādilābād. 19·40n 78·32e
Ādoni. 15·38n 77·17e
Agartala. 23·50n 91·16e
Āgra. 27·12n 77·59e
Ahmadābād. 23·02n 72·35e
Ahmadnagar. 19·05n 74·45e
Aijal. 23·44n 92·43e
Ajanta. 20·30n 75·48e
Ajmer. 26·27n 74·40e
Akola. 20·42n 77·00e
Alībāg. 18·39n 72·53e
Alīgarh. 23·59n 87·20e
Allahābād. 25·27n 81·50e
Alleppey. 9·30n 76·20e
Almora. 29·36n 79·41e
Alwar. 27·46n 76·37e
Amarkantak. 22·40n 81·45e
Ambāla. 30·23n 76·46e
Amber. 27·00n 75·51e
Ambikāpur. 23·08n 83·11e
Amrāvati. 20·56n 77·47e
Amreli. 21·36n 71·12e
Amritsar. 31·38n 74·53e
Amroha. 28·54n 78·38e
Anantapur. 14·41n 77·37e
Anantnāg. 33·44n 75·10e
Angul. 20·51n 85·06e
Anjār. 23·07n 70·02e
Arrah. 25·33n 84·40e
Aruppukkottai. 9·31n 78·07e
Asansol. 23·41n 86·59e
Aurangābād. 19·53n 75·20e
Bādāmi. 15·52n 75·46e
Badrīnāth. 30·44n 79·30e
Bahraich. 27·36n 81·36e
Bālāghāt. 21·49n 80·11e
Balāngīr. 20·43n 83·30e
Balasore. 21·29n 86·56e
Ballia. 25·45n 84·20e
Bānda. 25·29n 80·21e
Bandar, see Machilīpatnam
Bandipūr. 11·40n 76·38e
Bangalore. 12·58n 77·36e
Bānkura. 23·14n 87·04e
Bānswāra. 23·32n 74·26e
Baranagar. 22·38n 88·22e
Bareilly. 28·20n 79·25e
Baripāda. 21·56n 86·44e
Barmer. 25·45n 71·23e
Baroda, see Vadodara
Barpeta. 26·19n 91·00e
Bārsi. 18·15n 75·42e
Bāruni. 25·29n 85·59e
Basīrhāt. 22·39n 88·52e
Bastar. 19·12n 81·56e
Batāla. 31·48n 75·13e
Beāwar. 26·06n 74·20e
Belgaum. 15·52n 74·31e
Bellary. 15·09n 76·56e
Belūr. 22·38n 88·20e
Benares, see Vārānasi
Berhampore. 24·06n 88·15e
Berhampur. 19·19n 84·48e
Betūl. 21·54n 77·54e
Bhadrakh. 21·03n 86·30e
Bhadrāvati. 13·52n 75·43e
Bhāgalpur. 25·15n 86·58e
Bhandāra. 21·10n 79·39e
Bharatpur. 27·26n 77·30e
Bhatinda. 30·13n 74·56e
Bhātpāra. 22·52n 88·24e
Bhaunagar. 21·47n 72·09e
Bhawānipatna. 19·55n 83·10e
Bhilai. 21·11n 81·20e
Bhīlwāra. 25·21n 74·40e
Bhind. 26·35n 78·48e
Bhiwāni. 28·48n 76·08e
Bhopāl. 23·15n 77·25e
Bhubaneswar. 20·15n 85·50e
Bhuj. 23·17n 69·41e
Bhusāwal. 21·03n 75·46e
Bīdar. 17·55n 77·33e
Bihār. 25·12n 85·33e
Bijāpur. 16·50n 75·42e
Bījāpur. 18·47n 80·49e
Bīkaner. 28·01n 73·18e
Bilāspur. 22·05n 82·08e
Bilāspur. 31·19n 76·45e
Bīr. 18·59n 75·46e
Bobbili. 18·35n 83·22e
Bokāro. 23·51n 86·02e
Bombay. 18·58n 72·50e
Broach. 21·42n 72·58e
Budaun. 28·03n 79·07e
Buddh Gaya. 24·35n 84·58e
Buldāna. 20·32n 76·11e
Būndi. 25·26n 75·41e
Burdwān. 23·14n 87·52e
Burhānpur. 21·19n 76·14e
Calcutta. 22·34n 88·20e
Calicut, see Kozhikode
Cambay. 22·19n 72·37e
Cannanore. 11·58n 75·21e
Chandīgarh. 30·44n 76·47e
Chandrapur. 19·57n 79·18e
Chāpra. 25·46n 84·45e
Chatrapur. 19·21n 85·00e
Cherrapunji. 25·17n 91·44e
Chhatarpur. 24·55n 79·36e
Chhindwāra. 22·04n 78·56e
Chidambaram. 11·24n 79·42e
Chikmagalūr. 13·20n 75·47e
Chitorgarh. 24·54n 74·39e
Chitradurga. 14·14n 76·24e
Chittaranjan. 23·52n 86·52e
Chittoor. 13·14n 79·07e
Churu. 28·18n 74·58e
Cochin. 9·58n 76·15e
Coimbatore. 11·00n 76·57e
Cooch Behār. 26·20n 89·27e
Coonoor. 11·21n 76·49e
Cuddalore. 11·45n 79·46e
Cuddapah. 14·29n 78·50e
Cuttack. 20·26n 85·53e
Dabhoi. 22·08n 73·26e
Dalhousie. 32·32n 75·59e
Daltonganj. 24·03n 84·04e
Damān. 20·25n 72·50e
Damoh. 23·50n 79·27e
Darbhanga. 26·10n 85·56e
Darjeeling. 27·02n 88·16e
Datia. 25·41n 78·28e
Dāvangere. 14·28n 75·56e
Dehra Dūn. 30·20n 78·02e
Dehri. 24·53n 84·14e
Delhi. 28·37n 77·12e
Deolāli. 19·56n 73·50e
Dewās. 22·57n 76·04e
Dhānbād. 23·47n 86·26e
Dhār. 22·36n 75·18e
Dhārwār-Hubli. 15·20n 75·08e
Dhubri. 26·01n 89·59e
Dhule. 20·54n 74·47e
Dibrugarh. 27·29n 94·55e
Dīg. 27·28n 77·20e
Digboi. 27·23n 95·38e
Dimāpur. 25·54n 93·45e
Dindigul. 10·21n 77·58e
Dispur. 26·09n 91·46e
Diu. 20·43n 70·59e
Dum Duma. 27·33n 95·33e
Dūngarpur. 23·50n 73·43e
Durg. 21·11n 81·17e
Durgāpur. 23·29n 87·20e
Dwārka. 22·15n 68·58e
Elūru. 16·42n 81·07e
English Bāzār. 25·00n 88·09e
Ernākulam. 9·59n 76·17e
Erode. 11·21n 77·43e
Etāwah. 25·33n 76·22e
Faizābād. 26·47n 82·08e
Farrukhābād-*cum*-Fatehgarh. 27·24n 79·35e
Fatehpur Sīkri. 27·06n 77·40e
Fīrozābād. 27·09n 78·24e
Fīrozpur. 30·55n 74·38e
Gadag. 15·25n 75·38e
Gāndhīnagar. 23·13n 72·41e
Gangotri. 30·56n 79·02e
Gangtok. 27·20n 88·37e
Gauhāti. 26·10n 91·45e
Gaya. 24·48n 85·03e
Geleki. 27·51n 94·40e
Ghāziābād. 28·40n 77·26e
Goālpāra. 26·11n 90·37e
Godhra. 22·46n 73·36e
Gonda. 27·08n 81·58e
Gondia. 21·27n 80·12e
Gorakhpur. 26·45n 83·22e
Gudiyāttam. 12·57n 78·53e
Gulbarga. 17·20n 76·50e
Guna. 24·39n 77·19e
Guntūr. 16·17n 80·28e
Gwalior. 26·13n 78·10e
Haldia. 22·00n 88·00e
Hamīrpur. 25·57n 80·09e
Hāpur. 28·44n 77·46e
Hardoi. 27.24n 80·07e
Hardwār. 29·52n 78·10e
Hassan. 13·01n 76·06e
Hāthras. 27·36n 78·03e
Hazāribāgh. 23·59n 85·21e
Himatnagar. 23·36n 72·58e
Hisār. 29·09n 75·44e
Hoshangābād. 22·45n 77·44e
Hoshiārpur. 28·35n 77·22e
Hospet. 15·16n 76·24e
Howrah. 22·35n 88·20e
Hyderābād. 17·23n 78·29e
Ichalkaranji. 16·41n 74·28e
Imphāl. 24·48n 93·56e
Indore. 22·43n 75·52e
Itanagar. 27·20n 93·50e
Itārsi. 22·37n 77·46e
Jabalpur. 23·10n 79·57e
Jagādhri. 30·10n 77·18e
Jagdalpur. 19·04n 82·02e
Jaipur. 26·56n 75·50e
Jaisalmer. 26·55n 70·55e
Jajpur. 20·49n 86·37e
Jālgaon. 21·03n 76·32e
Jālna. 19·50n 75·53e
Jālor. 25·21n 72·37e
Jalpaiguri. 26·32n 88·44e
Jamālpur. 25·25n 86·40e
Jammu. 32·44n 74·52e
Jāmnagar. 22·28n 70·05e
Jamnotri. 30·57n 78·30e
Jamshedpur. 22·48n 86·11e
Jangipur. 24·28n 88·03e
Jaunpur. 25·45n 82·42e
Jhābua. 22·46n 74·35e
Jhālawār. 24·36n 76·11e
Jhānsi. 25·27n 78·35e
Jhūnjhunu. 28·08n 75·24e
Jodhpur. 26·17n 73·01e
Jorhāt. 26·46n 94·13e
Jullundur. 31·20n 75·35e
Junāgadh. 21·31n 70·27e
Kākināda. 16·57n 82·14e
Kālimpong. 27·05n 88·29e
Kalpakkam. 12·50n 80·15e
Kalyān. 19·15n 73·08e
Kānchipuram. 12·50n 79·43e
Kandla. 23·02n 70·14e
Kāngra. 32·06n 76·16e
Kānpur. 26·28n 80·20e
Karīmnagar. 18·26n 79·08e
Karnāl. 29·41n 76·59e
Karūr. 10·57n 78·05e
Kārwār. 14·48n 74·08e
Kasia. 26·44n 83·54e
Katihār. 25·32n 87·35e
Kavaratti. 10·33n 72·38e
Keshod. 21·20n 70·19e
Khadki. 18·34n 73·52e
Khajrāho. 24·50n 79·58e
Khammam. 17·15n 80·09e
Khandwa. 21·50n 76·20e
Kharagpur. 22·20n 87·19e
Kharakvasla (Khadakvasla). 18·26n 73·46e
Khargon. 21·50n 75·37e
Khetri. 27·59n 75·48e
Kodaikānal. 10·14n 77·29e
Kohīma. 25·40n 94·06e
Kolār. 13·08n 78·08e
Kolār Gold Fields. 12·55n 78·17e
Kolhāpur. 16·42n 74·13e
Konārak. 19·53n 86·10e
Koraput. 18·48n 82·41e
Kota. 25·11n 75·50e
Kottagūdem. 17·33n 80·39e
Kottayam. 9·35n 76·31e
Kozhikode (Calicut). 11·16n 75·48e
Krishnanagar. 23·24n 88·30e
Kulu. 31·58n 77·06e
Kumbakonam. 10·57n 79·23e
Kurnool. 15·50n 78·03e
Kusijan. 27·33n 95·35e
Lakawa. 27·50n 94·40e
Lakhīmpur. 27·57n 80·47e
Lalitpur. 24·42n 78·25e
Ledo. 27·18n 95·44e
Leh. 34·10n 77·36e
Lucknow. 26·50n 80·52e
Ludhiāna. 30·55n 75·51e
Machilīpatnam (Bandar). 16·11n 81·08e
Madras. 13·04n 80·16e
Madurai. 9·56n 78·08e
Mahābaleshwar. 17·56n 73·42e
Mahābalipuram. 12·37n 80·13e
Mahbūbnagar. 16·44n 77·59e
Mahe. 11·42n 75·32e
Mālegaon. 20·33n 74·32e
Manāli. 32·15n 77·10e
Mandasor. 24·04n 75·04e
Mandla. 22·36n 80·23e
Mangalore. 12·52n 74·52e
Mathura. 27·30n 77·41e
Mattāncheri. 9·57n 76·16e
Māyūram. 11·06n 79·39e
Meerut. 29·00n 77·43e
Mehsāna. 23·36n 72·24e
Mercāra. 12·26n 75·45e
Midnapore. 22·25n 87·20e
Miraj. 16·50n 74·39e
Mirzāpur-*cum*-Vindhyāchal. 25·09n 82·35e
Monghyr. 25·23n 86·29e
Morādābād. 28·50n 78·47e
Mormugao. 15·26n 73·50e
Morvi. 22·50n 70·50e
Murwāra. 23·50n 80·24e
Mussoorie. 30·27n 78·05e
Muzaffarnagar. 29·28n 77·42e
Muzaffarpur. 26·08n 85·24e
Mysore. 12·18n 76·39e
Nabadwīp. 23·25n 88·22e
Nadiād. 22·42n 72·52e
Nāgappattinam. 10·46n 79·50e
Nāgaur. 27·12n 73·45e
Nāgercoil. 8·11n 77·26e
Nāgpur. 21·08n 79·04e
Nāhan. 30·33n 77·13e
Naini Tāl. 29·23n 79·28e
Nālanda. 25·07n 85·25e
Nalgonda. 17·03n 79·16e
Nānder. 19·10n 77·21e
Nārnaul. 28·03n 76·07e
Narsimhapur. 23·00n 79·10e
Nāsik. 19·59n 73·47e
Nāthdwāra. 24·56n 73·51e
Navsāri. 20·57n 42·56e
Nellore. 14·26n 79·59e
New Delhi. 28·36n 77·15e
Nizāmābād. 18·40n 78·06e
North Lakhimpur. 27·14n 94·07e
Nowgong. 25·04n 79·27e
Ootacamund. 11·25n 76·43e
Orai. 25·59n 79·32e
Osmānābād. 18·10n 76·02e
Pachmarhi. 22·28n 78·26e
Pālam. 19·01n 76·57e
Pālampur. 32·07n 76·32e
Pālanpur. 24·10n 72·26e
Pālayankottai. 8·43n 77·44e
Pālghāt. 10·46n 76·40e
Pāli. 25·46n 73·20e
Pālitāna. 21·31n 71·49e
Palni. 10·27n 77·31e
Panaji (Panjim). 15·29n 73·50e
Pānipat. 29·25n 76·59e
Panjim, see Panaji
Panna. 24·43n 80·12e
Paradip. 20·15n 86·41e
Parbhani. 19·16n 76·47e
Parlākimidi. 18·47n 84·06e
Pātan. 23·51n 72·06e
Pathānkot. 32·17n 75·40e
Patiāla. 30·19n 76·23e
Patna. 25·36n 85·06e
Pilībhīt. 28·37n 79·48e
Pollāchi. 10·40n 77·01e
Pondicherry. 11·56n 79·50e
Poona, see Pune
Porbandar. 21·38n 69·36e
Port Blair. 11·39n 92·45e
Proddatūr. 14·45n 78·33e
Pudukkottai. 10·24n 78·49e
Pune (Poona). 18·31n 73·54e
Puri. 19·48n 85·50e
Purnea. 25·47n 87·31e
Pushkar. 26·28n 74·36e
Quilon. 8·53n 76·36e
Rāe Bareli. 26·14n 81·15e
Rāichūr. 16·12n 77·21e
Raigarh. 21·55n 83·24e
Raipur. 22·48n 86·57e
Rājahmundry. 17·01n 81·48e
Rājapālaiyam. 9·27n 77·33e
Rājgarh. 24·02n 76·43e
Rājkot. 22·18n 70·47e
Rāj-Nāndgaon. 21·06n 81·03e
Rāmeswaram. 9·17n 79·19e
Rāmpur. 29·48n 77·27e
Rānāghāt. 23·11n 88·35e
Rana Pratap Sagar. 25·00n 76·00e
Rānchī. 23·22n 85·19e
Rānīkhet. 29·38n 79·26e

MAP INDEX (continued)

Ratlām.....23·28n 75·00e
Ratnāgiri.....16·59n 73·17e
Raurkela.....22·14n 84·54e
Raxaul.....26·59n 84·51e
Rewa.....24·32n 81·17e
Rohtak.....28·54n 76·35e
Roorkee.....29·53n 77·52e
Sadiya.....27·50n 95·40e
Sāgar.....14·10n 75·02e
Sahāranpur.....29·57n 77·48e
Salem.....11·39n 78·10e
Sambalpur.....21·28n 83·59e
Sambhal.....28·35n 78·34e
Sāmbhar.....26·55n 75·12e
Sāngli.....16·52n 74·34e
Sasarām.....24·57n 84·02e
Sātāra.....17·41n 73·59e
Satna.....24·35n 80·51e
Sawai Mādhopur.26·00n 76·39e
Secunderābād...17·27n 78·29e
Seoni.....22·06n 79·32e
Seringapatam....12·25n 76·42e
Sevagram.....20·45n 78·30e
Shahdol.....23·20n 81·21e
Shāhjahānpur....27·53n 79·55e
Shillong.....25·35n 91·53e
Shimoga.....13·56n 75·35e
Shivpuri.....25·25n 77·40e
Sholāpur.....17·40n 75·55e
Sibsāgar.....26·59n 94·39e
Sidhi.....24·24n 81·53e
Sīkar.....27·36n 75·08e
Silchar.....24·49n 92·48e
Silīguri.....26·43n 88·27e
Silvassa.....20·17n 73·00e
Simla.....31·06n 77·09e
Sindri.....23·45n 86·42e
Sirohi.....24·54n 72·51e
Sītāpur.....27·35n 80·40e
Sonepur.....20·50n 83·54e
Sri Gangānagar..29·55n 73·52e
Srīkākulam.....18·18n 83·55e
Srīnagar.....34·05n 74·49e
Sukinda Khās....20·58n 85·55e
Surat.....21·12n 72·50e
Surendranagar..22·43n 71·38e
Tādpatri.....14·55n 78·01e
Tārāpur.....19·51n 72·42e
Tenāli.....16·14n 80·40e
Tezpur.....26·37n 92·48e
Thāna.....19·12n 72·59e
Thanjāvūr.....10·47n 79·09e
Tiruchchirāp-
palli.....10·48n 78·41e
Tirunelveli..... 8·44n 77·41e
Tiruppur.....11·06n 77·21e
Tonk.....26·11n 75·47e
Trichūr.....10·31n 76·13e
Trivandrum..... 8·28n 76·57e
Trombay.....19·02n 72·54e
Tumkūr.....13·21n 77·06e
Tura.....25·31n 90·13e
Tuticorin..... 8·50n 78·09e
Udaipur.....24·35n 73·41e
Ujjain.....23·12n 75·46e
Ulhāsnagar.....19·13n 73·07e
Vadodara
(Baroda).....22·18n 73·12e
Valparai.....10·19n 76·58e
Vārānasi
(Benares).....25·20n 83·00e
Vellore.....12·56n 79·09e
Verāval.....20·54n 70·20e
Vidisha.....23·32n 77·49e
Vijayapuri.....16·52n 79·35e
Vijayawāda.....16·31n 80·37e
Vikramasinga-
puram..... 8·43n 77·24e
Virudunagar..... 9·35n 77·57e
Vishākhapatnam.17·43n 83·19e
Vizianagaram....18·07n 83·25e
Vrindāvan.....27·35n 77·42e
Warangal.....18·00n 79·33e
Wardha.....20·45n 78·36e
Yanam.....16·35n 82·10e
Ziro.....27·40n 92·40e

Physical features and points of interest

Adams Bridge,
shoal..... 9·04n 79·37e
Agnigundala,
physical region..24·00n 83·50e
Ajanta Range,
mountain range.20·30n 76·00e
Alia Bet, *island*...21·36n 72·40e
Amīndīvi Islands.11·23n 72·23e
Amjore,
physical region..24·00n 83·30e
Anai Mudi,
mountain.....10·10n 77·04e
Andaman
Islands.....12·00n 93·00e
Androth, *island*..10·49n 73·40e
Arabian Sea.....19·00n 69·00e
Arāvalli Range,
mountain range.25·00n 73·30e
Ashtamudi Lake.. 8·59n 76·36e
Banās, *river*.....25·55n 76·45e
Baratang, *island*..12·15n 92·45e
Bengal, Bay of...15·00n 88·00e
Betwa, *river*.....25·55n 80·12e
Bhīma, *river*.....16·24n 77·18e
Brāhmani, *river*..20·45n 87·00e
Brahmaputra,
river.....26·00n 89·55e
Calimere, Point..10·20n 79·52e
Cambay, Gulf of..21·30n 72·30e
Camorta Island... 8·10n 93·30e
Car Nicobar
Island..... 9·10n 92·45e
Cauvery, *river*....11·10n 79·51e
Chambal, *river*...26·30n 79·15e
Chenāb, *river*....33·00n 74·30e
Chhattīsgarh,
plain.....21·15n 82·00e
Chilka Lake.....19·45n 85·25e
Chota Nāgpur
plateau.....23·30n 84·30e
Comorin, Cape... 8·06n 77·33e
Coromandel
Coast.....13·30n 80·30e
Dāmodar, *river*...23·14n 87·39e
Deccan, *plateau*..17·00n 78·00e
Dhaula Dhār,
mountains.....32·12n 76·33e
Doda Betta,
mountain.....11·26n 76·44e
Eastern Ghāts,
mountains.....16·00n 79·00e
False Divi Point..15·45n 80·50e
Gandak, *river*....25·35n 85·12e
Ganges (Ganga),
river.....24·00n 88·45e
Gīr, *forest*.....21·00n 70·40e
Godāvari, *river*...16·37n 82·18e
Gomati, *river*.....25·32n 83·11e
Great Channel... 6·25n 92·40e
Great Indian
Desert, see
Thar Desert
Great Nicobar,
island..... 7·00n 93·50e
Gujarāt,
physical region..22·51n 71·30e
Gumti, *river*.....25·31n 83·10e
Himalayas,
mountains.....32·30n 77·30e
Indrāvati, *river*...18·43n 80·17e
Indus, *river*.....34·30n 76·30e
Kalpeni *island*...10·05n 73·38e
Kalsūbai,
mountain.....19·35n 73·42e
Kāmet, *mountain*.30·54n 79·37e
Kānchenjunga,
mountain.....27·42n 88·09e
Kangto,
mountain.....27·50n 92·25e
Kashmir,
physical region..34·30n 76·30e
Katchall Island... 7·55n 93·25e
Kāthiāwār,
peninsula.....22·00n 71·00e
Kavaratti, *island*..10·34n 72·39e
Khāsi Hills.....25·45n 91·30e
Kiltān, *island*.....11·30n 73·00e
Konkan,
physical region..17·00n 73·00e
Krishna, *river*....15·43n 80·55e
Kutch, Gulf of....22·30n 69·30e
Laccadive
Islands.....10·00n 73·00e
Ladākh, *physical*
region.....35·00n 78·00e
Ladākh Range,
mountain range.34·00n 77·30e
Little Andaman,
island.....10·40n 92·30e
Little Nicobar,
island..... 7·20n 93·40e
Lūni, *river*.....24·40n 71·15e
Mahānadi, *river*..20·00n 86·25e
Mahendra Giri,
mountain.....18·58n 84·25e
Mahi, *river*.....22·15n 73·00e
Maikala Range,
mountains.....22·30n 81·30e
Malabar Coast...11·00n 75·00e
Mānjra, *river*.....18·49n 77·52e
Mannar, Gulf of.. 8·30n 79·00e
Middle Andaman,
island.....12·30n 92·53e
Minicoy, *island*... 8·17n 73·04e
Mishmi Hills.....29·00n 96·00e
Mizo Hills.....23·30n 93·00e
Nancowry Island. 7·58n 93·33e
Nanda Devi,
mountain.....30·22n 79·59e
Narmada, *river*...21·32n 72·35e
Nicobar Islands.. 8·00n 93·50e
Nine Degree
Channel..... 9·00n 73·00e
North Andaman,
island.....13·20n 93·00e
Nunkun,
mountain.....33·59n 76·02e
Palk Strait..... 9·30n 79·30e
Pārbati, *river*.....25·51n 76·34e
Pātkai Range,
mountain range.27·00n 96·00e
Penganga, *river*..19·53n 79·09e
Penner, *river*.....14·35n 80·11e
Pīr Panjāl Range,
mountain range.32·37n 74·32e
Ponnaiyār, *river*..11·46n 79·48e
Rāmgarh, *hill*....22·57n 82·52e
Rann of Kutch,
salt flat.....24·00n 70·00e
Sābarmati, *river*..22·25n 73·20e
Sātpura Range,
mountain range.22·00n 78·00e
Shyok, *river*.....35·13n 75·53e
Son, *river*.....25·40n 84·51e
Subansiri, *river*...26·48n 93·50e
Subarnarekha,
river.....21·33n 87·12e
Sundarbans,
physical region..22·00n 88·45e
Surma, *river*.....24·42n 91·24e
Sutlej, *river*.....31·30n 74·30e
Tāpti, *river*.....21·05n 72·40e
Teressa Island... 8·15n 93·10e
Thar Desert,
(Great Indian
Desert).....27·30n 72·00e
Tillanchong
Island..... 8·30n 93·38e
Trincat Island.... 8·05n 93·35e
Tungabhadra,
river.....15·57n 78·15e
Vindhya Range,
mountain range.23·00n 77·00e
Wainganga, *river*.18·50n 79·55e
Wardha, *river*....19·36n 79·48e
Western Ghāts,
mountains.....14·00n 75·00e
Wular Lake.....34·30n 74·30e
Yamuna, *river*....25·25n 81·50e
Zāskār
Mountains.....33·00n 77·30e

85° F (29.5° C), the highest average for all India. In the coastal area, Bombay, on the west coast, has an average temperature of about 82° F (28° C), while Madras, on the east coast, records 83.5° F (29° C). Because of India's great climatic diversity, it is practicable to talk only in terms of local temperature averages, rather than of national averages.

PLANT AND ANIMAL LIFE

Vegetation. With its great latitudinal spread, encompassing a wide range of temperature conditions, its varied humidity and rainfall, and its altitudinal extremes, India produces a rich and varied vegetation, which differs according to region.

Palms of different kinds are endemic to the dry, elevated tableland of the Deccan region. The Malabar area, covering the west coast and the mountains of the Western Ghāts, is rich in tropical vegetation. Commercial crops such as coconuts, betel nuts, pepper, ginger, rubber, and bananas of all varieties thrive in the midlands and coastal areas of this region. At a higher elevation are plantations of coffee, tea, and cardamom. The forest areas abound in such hardwoods as rosewood, ironwood, and teak. Numerous varieties of softwood and bamboos of different kinds also grow here.

River valley vegetation

The Indus region, covering the plains of Punjab, Rājasthān, and northern Gujarāt, is very poor in endemic plants; in the Ganges region, however, which spreads from the Yamuna River to West Bengal and Orissa, forests of widely differing types occur, with forests of sal (*Shorea robusta*) predominating. The vegetation in the Brahmaputra and Surma valleys of the Assam region and of the intervening hill ranges is particularly luxuriant and is characterized by tall grass, broad-leaved forests, and thick clumps of bamboo.

About 4,000 species of flowering plants along with 20 variants of palms flourish in the eastern Himalayan region, which stretches eastward from Sikkim. Many laurels, maples, alders, birches, conifers, and junipers grow here. Rhododendrons, dwarf willows, and bamboos abound. The western Himalayan region, which extends from the Kumaun hills to Kashmir, consists of three zones—alpine, temperate, and lower—whose vegetation differs. Characteristic trees in the alpine zone are the high-level silver fir, the silver birch, and junipers. Forests of conifers are to be found in the temperate zone, which also contains deodar (an East Indian cedar), spruce, and silver fir. A belt of sal forest stretches across the lower zone, where a few species of palms also grow. In the outlying islands of Andaman and Nicobar are a variety of forests—mangrove, beech, evergreen, semi-evergreen, and deciduous.

India's main cultivated plants include rice, wheat, pulses (leguminous plants with edible seeds, such as peas and beans), and grains. Castor, sesamum, and peanuts (groundnuts) are among other important seeds grown. The chief fruit trees include mango, coconut, areca nut (from the betel palm) and certain citrus fruit. The Himalayan valleys produce apples, peaches, pears, apricots, and walnuts. Apart from tea, coffee, cardamom, pepper, and rubber, which are grown extensively in the south, cinchona (a tree

whose bark contains alkaloids, such as quinine, used as a specific in malaria) is cultivated in some of the damper regions.

Animals. India has a large variety of wild animals, some of which are classified within the "big game" category.

Among the big game are three large cats—the tiger, panther, and cheetah. The tiger, the monarch of almost all Indian forests, is found in various parts of India, principally in grassy plains and swamps but also in forests. A popular hunting ground for tigers is the long forest tract that runs for 1,000 miles along the foothills of the Himalayas. Indian tigers live principally on cattle, goats, deer, and wild hog; when old they may become man-eaters. There are four species of panther—the common leopard found all over India, the all-black and albino leopards, and the snow leopard, which is seen only in the Himalayas. The cheetah, or hunting leopard, renowned for its swiftness in the chase, has practically disappeared from India. So has the Asiatic lion, once widespread in North India but now found only in the Gir Forest National Park in Saurāshtra, a region of Gujarāt, where it is strictly protected. In 1880 there were barely a dozen head left in that region, but the lion population is now estimated to number more than 300.

Indian elephants

Elephants roam the forests of peninsular India. They are numerous in Orissa and are also found in fair numbers in the jungles of the Western Ghāts and in north Coimbatore in Tamil Nādu. Apart from being used for hauling logs from inaccessible forests, elephants also figure in ceremonial processions, although to a lesser degree with the passing of the princely order of society. A favourite haunt of the elephant is Assam, where only proclaimed rogue elephants may be shot. The great Indian one-horned rhinoceros was also once threatened with extinction, but is now a protected species; more than 400 roam the sanctuaries and national parks provided for them in Assam state.

There are various kinds of bear, deer, antelope, gazelle, goat, and sheep. The sloth bear, or common black bear, is found all over India. The snow, red, or brown bear, as well as the Asiatic black bear, roam the western Himalayas. There is only a single species of panda, the lesser or red panda, but there are 18 species of martens, weasels, and otters. Civets and mongooses are found throughout India. The Indian, or striped, hyena is the only extant species of this family in most of the country. There are two types of wolf—the woolly wolf of the western Himalayas and the small Indian wolf, which is common throughout the country. The jackal is common, and there are at least four species of fox, as well as several strains of Indian wild dog. India has different varieties of wildcats—the Indian desert cat, the common jungle cat, the leopard cat, and the fishing cat. The lynx's main habitat is in Kashmir; the karakul (a breed of Asian sheep) lives in northwestern India.

Two species of wild pig survive, the wild boar and the pygmy hog. The deer family is represented by the mouse deer, spotted deer, barking deer, and sambar (a kind of deer having three-pointed antlers), which are spread over various parts of India. The musk deer is confined to Kashmir, and the Sikkim stag has become rare. A disappearing species is the Kashmir stag, or hangul, now a protected animal. The Indian gazelle, the four-horned antelope, and the Indian black buck roam wide areas of the country.

To the wild buffalo family belong the gaur, also known as the Indian bison; one of the most handsome specimens of wild bull, it is preserved in the southern sanctuaries of Bandipur in Karnātaka, and Mudumalai in Tamil Nādu. The nilgai, or blue bull, ranges all over India except West Bengal and Malabār. The markhor (a kind of wild goat, having spiral horns and long, shaggy hair) is found in the Himalayas, which is also the home of the Asiatic ibex (another kind of wild goat). The wild yak inhabits the uplands of Ladākh.

Domestic animals include cattle, buffalo, sheep, goats, horses, ponies, mules, donkeys, and camels. The cobra is the best known among the Indian reptiles. There are two varieties of cobra, the common cobra and the king cobra (also known as the hamadryad), which both grow up to lengths of more than 14 feet. One of the most deadly poisonous snakes is the krait, of which there are 11 species. Vipers are represented by more than 20 species, the most common among them being Russell's viper, the pit viper, and the saw-scaled viper. Common snakes include the rat snake, the grass snake, and the wolf snake. All three species of crocodile found in India are protected; there are no alligators. There are about 50 species of turtles, tortoises, and terrapins (a kind of aquatic turtle).

Birdlife

About 1,200 species of birds inhabit India; together with subspecies, the total number is about 2,000. Of these, about 1,750 are resident in India, and the rest are migratory. Birds of prey include vultures, falcons, hawks, and osprey (fish hawks). Among fish eaters are pelicans, cormorants, frigate birds, and darters. Perching birds or songbirds constitute more than half of India's birdlife. The various kinds of parrots are remarkable for their beauty. Kingfishers of various kinds, as well as herons, are notable for their plumage. The myna bird, a member of the starling family, is a familiar feature of the country's ornithological landscape and like the parrot can imitate speech. Feathered game includes duck, goose, snipe, pheasant, and both gray and black partridge. The desert areas of Rājasthān are the home of the sand grouse, including the much-prized imperial sand grouse. Jungle fowl are plentiful, and the bustard, florican (smaller species of bustard), quail, green pigeon, and several other varieties of game bird are also found. Birds are protected in the bird sanctuaries of Keoladeo Ghana in Rājasthān, Vedanthangal in Tamil Nādu, and Ranganthittoo in Kārnataka. The Indian peacock, with its showy, resplendent blue plumage, has been declared the national bird and is now a protected species.

India offers generous opportunities to the angler, since many of its rivers and lakes are well stocked with a large variety of indigenous fish. Trout is available in hill streams, particularly in Kashmir, the Kulu Valley, and Ootacamund, in Tamil Nādu. The mahseer, or Indian salmon, is found in most of the large rivers. Other edible fish include members of the carp family and catfishes; the hilsa, a type of herring, is popular on the east coast and the pomfret (a kind of East Indian fish) in Bombay.

Notorious among Indian insects are mosquitoes, which abound everywhere. Locust invasions, although infrequent, are disastrous, particularly in northwestern India. Among the more useful insects are the bee, silkworm, and coceus laca, which yields lac, from which shellac is made. There are many beautiful species of butterflies.

SETTLEMENT PATTERNS

Traditional regions. From the Himalayas southward, India consists of a mosaic of traditional regions, reflecting variegated and sometimes confusing patterns of race and language. Since Indian independence, there has been a marked tendency to make state boundaries coextensive with those of linguistic or traditional regions. Further information on many of the traditional regions mentioned below may be found in the appropriate state articles.

The Himalayan region. The Himalayan region stretches between the arm of the Indus River, which curves in a great bend near Gilgit, in Kashmir, and a similar sharp turn of the Brahmaputra in Upper Assam. Traditionally the abode of the gods, the Himalayas are the meeting ground of two cultures and civilizations, Aryan and Mongolian. The Kulu and Kāngra valleys of Himāchal Pradesh, at the foot of the Dhaula Dhār Range of the Himalayas, are also a meeting ground of Hindu and Buddhist religions. In the state of Arunāchal Pradesh (formerly the North East Frontier Agency; NEFA), Buddhism and Hinduism again meet, but animism is also represented.

At the northernmost tip of India is the trans-Himalayan region of Ladākh, which spreads over a plateau that has an average elevation of 19,000 feet and occupies the northeastern part of the Kashmir Himalaya. Below the snow line, its terrain consists of rugged barren landmasses interspersed with deep vertical valleys with sparse green belts of vegetation and occasional rivulets and streams. Ladākh is a dry arid region which might be described as a high altitude desert. Its population is Tibetan, and is Buddhist by religion. It covers an area of about 38,000 square miles; its terrain is dotted with villages of mud huts, multistoried monasteries, and stone and brick houses.

Ladākh is part of the state of Jammu and Kashmir. The state, which is claimed by both India and Pakistan, who each occupy part of it, is divided by geography, race, and religion into three areas. In the north are the Tibetan and semi-Tibetan tracts, which include Ladākh and Gilgit; the latter, since 1948, has been part of Pakistan. In the south are the large, level areas of Jammu, inhabited mainly by the Hindu Dogras. In between is the Vale of Kashmir, in which the population is well over 80 percent Muslim. As a traditional region, Jammu and Kashmir represents a microcosm of the in-between world wedged between the Himalayas and the Indo-Gangetic Plain. The relics of three civilizations and religions—Buddhist, Hindu, and Muslim—survive in this area. The Buddhist monasteries are located almost exclusively in the eastern district of Ladākh.

Another traditional region in the Himalayan area is Arunāchal Pradesh, peopled by a variety of Indo-Mongoloid tribes, including the Nishis, Apa Tanis, Mishmis, Monpas, Tangsas, and Wanchos. They occupy more than 32,000 square miles of mountainous country bounded by Bhutan to the west, the Tibetan and Sinkiang regions of China to the north, Assam and Nāgāland to the south, and Burma to the south and east.

Singularly hardy people live in this region, which is criss-crossed by countless streams and rivers, which become raging torrents during rainy periods. Almost all these tribal peoples live in pile dwellings, raised well above the ground to avoid the damp. Though the tribes differ, they have a common affinity in their way of life. Each tribe is broadly endogamous (practicing marriage within the group) and is divided into clans, which are exogamous (practicing marriage outside the clan). A breach of the clan rules is a major social calamity. Society is patrilineal. Polygamy is common. There are traces of polyandry among some tribes, such as the Gallongs and the Tibetan-influenced border tribes of the far north.

Environment for many centuries has been the real ruler of the tribal peoples; among some of them, gods are identified with the elements. Christianity and Hinduism have had little influence on the tribes, and even Tibetan or Burmese Buddhism has had little effect.

Indo-Gangetic region. Like the Himalayan belt regions, the Indo-Gangetic Plain contains some traditional regions, characterized by distinctive racial, religious, or regional factors.

The Punjab region

Before independence in 1947, the Punjab (Pañjāb) region, in the heart of the Indo-Gangetic Plain, existed as a composite mosaic of Sikhs living beside Hindu and Muslim Punjabis. After independence the Punjab was divided into east and west Punjab, the former, predominantly Sikh and Hindu, becoming a part of India, the latter, exclusively Muslim, becoming part of Pakistan. Since 1966, east Punjab has been further divided into Punjab—a predominantly Sikh region—and the largely Hindu-populated state of Haryāna.

Sikh-dominated Punjab is the traditional homeland of the Sikhs, a sect that became a people. The Sikhs are physically vigorous and provide India with some of its best farmers and soldiers. The area of Punjab they now control is some 19,450 square miles.

West of Punjab stretches Rājasthān, notable for the Rājput clans who inhabit this sprawling, largely desert region. Rājasthān is famed in Indian history as the cradle of Hindu chivalry and culture; it was the Rājputs who offered the main Hindu resistance to the early Muslim invaders. Their principles of conduct were then founded upon the heroes of the ancient Hindu epics; today, however, their social structure is innately conservative. Their princely rulers seem to belong to past ages, rather than to present-day India.

Rājasthān is largely an amalgam of the old princely states of Rājputānā. Among the more prominent of these medieval domains are Udaipur, Jaipur, Jodhpur, Bīkaner, and Jaisalmer. The principal languages of this region, which covers more than 132,000 square miles, are the various Rājasthānī dialects and Hindī.

India's post-independence political leaders come principally from the huge state of Uttar Pradesh, which extends over some 113,000 square miles. Three of the country's prime ministers—Jawaharlal Nehru, Lal Bahadur Shastri, and Indira Gandhi—have come from Uttar Pradesh, a region that stretches from the Yamuna to Bihār, in which is the heartland of modern Hinduism and the ancient centre of legendary Hinduism. It contains the holy cities of Mathura and Varānāsi (sometimes known as Banaras or Benares)—the latter renowned for its temples and for its Brahmin priesthood, which is the most conservative in the country. Hindī, the principal language of this state, has also been declared the national language.

The Bengal region

East of Uttar Pradesh, beyond Bihār, is Bengal, a traditional region peopled by the Bengalis. The Bengalis have produced a great number of literary, artistic, philosophic, religious, and political figures; these have included the poet and mystic Rabindranath Tagore, Swami (a title meaning "religious teacher") Vivekananda, and Sri Aurobindo. The Bengalis, whether Hindu or Muslim, speak the same language and wear the same dress, except in the urban areas. Bengal has been described as the land of the dhoti, a flowing white-cotton loincloth tucked up in front and fastened behind. Politically, Bengal is one of the chief centres of Maoist Communism and its Indian offshoot, the Naxalite movement, which in the past left in its trail a series of riots, political murders, and strikes.

Nāgāland, Manipur, and Tripura. Nāgāland, farther to the east, between the Brahmaputra Valley and Upper Burma, is another traditional region. The Nāgas, though constituting tribes of mixed origins and varied cultures and often differing in physical appearance, have a composite character and are identifiable. They were once notorious as headhunters, but a considerable number of them have been converted to Christianity. They speak diverse languages, but all belong to the Tibeto-Burman family. The mountainous borderland they occupy covers nearly 6,400 square miles.

Manipur, south of Nāgāland, and Tripura, west of Assam and Mizorām, are other traditional regions in the east that possess their own distinctive languages and cultures. Both the Maṇipurīs and Tripuris show physical traces of a Mongolian origin. The language of the Maṇipurīs is a branch of the Kuki-Chin family; their women enjoy a high social position and are active in trade; the dominant religion is Hindu. The region is noted for its popular Maṇipurī dances. Some hold that polo was introduced to the British from Manipur. Tripura is said to have once extended as far as to Arakan, in Burma.

Gujarāt and Mahārāshtra. These two distinctive traditional regions lie in the west, grouped around the Gujarātī and Marāṭhā peoples. The mainland of Gujarāt extends from the Rann of Kutch and the Arāvalli Hills to the Damangangā River and includes Saurāshtra and Kutch and the hilly tracts to the northeast of the state. Gujarāt's area is a little more than 75,000 square miles. The prevalent language is Gujarātī, spoken by the Hindus, Muslims, and Jains of the region. Until 1960 Gujarāt formed part of Bombay state (now Mahārāshtra), from which it was separated on linguistic grounds. The Gujarātīs are noted for their enterprise and business acumen, both at home and abroad. They are much influenced by Jain philosophy, which includes belief in the inviolability of all life; Mahatma Gandhi was himself a Gujarātī, and his doctrine of nonviolence had an instinctive appeal to his regional kinsmen.

The language of Mahārāshtra is Marāṭhī. Mahārāshtra forms a major part of peninsular India, having its coastline on the Arabian Sea, to the west. It extends over some 119,000 square miles.

The Marāṭhās, a physically small people, inhabit the western hills, from where, in the 17th century, they resisted the armies of the Mughal invaders. Later, the Marāṭhās disputed the domination of the British and—with the Sikhs—were the last to offer military resistance to British rule. The Marāṭhās are noted for their intense local patriotism and pride; for nearly 300 years they have dominated the life of western India. The Brahmins among them have produced many scholars, historians, jurists, and lawyers, although it is the non-Brahmins who today are paramount in political life.

Goa. Another traditional region is Goa (capital Panaji), on the Konkan coast of Mahārāshtra, which until 1961 was a Portuguese possession. Goa is the heartland of Catholicism in India. The Goan Catholics form a distinctive community strongly influenced by Portuguese culture, largely because of the link of religion. Though Hindus constitute the majority in Goa, the Goan Catholics, because of the political patronage of the Portuguese, acquired a special place in this region.

The Dravidian region. In the southern part of India are the Dravidian lands of the Tamil, Andhra, Kannaḍa, and Malayāli peoples. Many customs, such as polyandry and matriarchy, unknown in northern India, survive in the south. The caste system exists there but only in a modified form; the Kṣatriya (warrior) and Vaiśya (trader) castes are practically nonexistent.

The Tamils

The Tamils have the longest cultural tradition of these groups; Tamil is the oldest and principal Dravidian tongue—"Dravida" and "Tamil" being two forms of the same word. They inhabit the region and state of Tamil Nādu, at the southeastern end of the Indian peninsula.

The Andhra people, who have inhabited the region and the state of Andhra Pradesh since antiquity, have developed their own language, Telugu—one of the four literary languages of the Dravidian family. Andhras do not compose one race; they are a mixture of the Dravidians and the Aryans. The majority of the Andhra people are Hindus, although there are sizeable Muslim and Christian populations as well. The cultural heritage of the Andhra people is evident in the realms of art, architecture, music, dance, and literature. Their *Kuchipudy* dance style is unique in the Indian classical dance tradition, and their shadow plays and puppet shows are two of the earliest forms of theatre in India. The *Kirtana* style of the Carnatic (Karnātic) music was developed by Tyāgarāja (1764–1846), one of the greatest composers of South Indian music.

The Kannaḍas, of the state of Karnātaka (ancient Karnād, meaning "the land of black soil") on the southwestern coast of India, possess a rich cultural heritage that dates to the times of the Maurya dynasty (3rd century BC). The Kannaḍas are predominantly Dravidians; over the centuries, however, much intermixture with Aryans has occurred. The major religion is Hinduism, but Jainism and Buddhism, once widespread, still survive. Kannaḍa literature is enriched by the writers of the Vīraśaiva, or Liṅgāyat, sect as well as by Jaina and Brahminical writers. The *Yakṣagāna*, a musical drama mostly using Purāṇic or semi-historical themes, developed in the Karnātaka region around the 17th century. Carnatic music, named for the Kannaḍas, evolved from ancient Hindu traditions and differs from the Hindustanī music of northern India.

The Malayālis who inhabit the picturesque state of Kerala comprise two distinct communities, the Hindu Nāyars and the Syrian Christians. Kerala state—formerly known as Malabār—emerged from the integration of two former princely states, Travancore and Cochin, united by the common tie of language, which is Malayālam. Of its varied communities, two are of especial interest in the context of traditional regions—the Syrian Christians and the white and black Jews. The latter live in the Cochin region, while the former are spread throughout Malabār along the west coast. Another distinctive community is formed by the Muslim Moplahs, who also inhabit the west coast and are descendants of Arab traders.

As a traditional region, Kerala is notable for some distinct social features. Matriarchal and matrilineal influences are pronounced, and the women of Malabār probably enjoy greater freedom than women in any other part of India.

Pondicherry. Pondicherry, in the east, represents an enclave somewhat different from the rest of India because of its long association with the French. Founded by the French in 1674 on the site of a village, it was later also held at different times by the Dutch and the British before being finally restored to the French in 1816; it was held by them until 1954, when it was transferred to India, becoming a legal part of India in 1963.

Pondicherry to some extent reflects its mixed heritage, but the impress of France seems as strong in this territory as the impress of Portugal in Goa.

Urban and rural settlement. Some three-quarters of India's people live in villages. Growing industrialization in the cities and towns has, however, led to an increasing drift toward urban areas.

There are more than 3,000 towns and more than 600,000 inhabited villages in India. Several cities have a population of more than 1,000,000, including Calcutta, greater Bombay, Greater Delhi, Madras, Bangalore, Ahmadābād, Hyderābād, Kānpur, Pune, and Nāgpur.

The people

ETHNOLINGUISTIC GROUPS

The peoples of India are largely the product of successive invasions that, from immemorial times, have swept into the country. Historically, the Indian subcontinent has formed a cul-de-sac; successive migratory waves of peoples have been largely halted here—although a part of some of the migratory waves also spread into Southeast Asia—and have consequently intermingled to such an extent that racially distinct categories of peoples are hard to establish on a definitive basis. Language, rather than ethnic origin, is the primary distinction between India's peoples. It may nevertheless be said that the main ethnic strains that have mingled in various combinations are the Caucasoid, Mongoloid, Australoid, and Negroid.

A British authority, Sir George Grierson (1851–1941), who compiled a monumental *Linguistic Survey of India,* listed 723 main languages and dialects. (This list, however, included dialects that vary almost from valley to valley in the Himalayan region, as well as languages that are spoken by only a very few people; Sanskrit, for example, although designated as one of the recognized languages by the Constitution of India, was spoken by fewer than 3,000 people in the late 20th century.) Officially, 15 regional languages are recognized by the Indian government. These are: (1) Assamese, (2) Bengali, (3) Gujarātī, (4) Hindī, (5) Kannaḍa, (6) Kashmīrī, (7) Malayālam, (8) Marāṭhī, (9) Oriyā (the language of Orissa), (10) Punjābī, (11) Sanskrit, (12) Sindhī, (13) Tamil, (14) Telugu, and (15) Urdū.

The national and regional languages

Broadly, the languages of India can be reduced to four families, two of which—Dravidian and Indo-Iranian (Aryan)—are spoken by the great majority of the people. The remaining two families are Austro-Asiatic and Sino-Tibetan. The Austro-Asiatic language family is limited principally to the Muṇḍā subgroup of languages, spoken in the hills and forests of central and eastern India. The Sino-Tibetan tongues are confined to the northeastern frontier area. In the course of centuries, there have been mutual borrowings by each of the four language families, but the individual character of each remains distinct and easily identifiable. Some language scholars maintain that certain Indian languages reveal no affinity to any of the four families. An example is the Nahali dialect, spoken in Bihār, which seems to be a survival of a fifth language group, as yet unidentifiable.

The Dravidian languages spread from the borders of Orissa to Cape Comorin through eastern and southern India. Tamil, the oldest, is spoken principally in Tamil Nādu. Telugu is spoken mainly in Andhra Pradesh. Malayālam, the youngest but most rapidly growing Dravidian language, is the language of Kerala, and Kannaḍa the language of Karnātaka.

The Indo-Iranian languages, of which the Indo-Aryan subgroup is found in India, belong to the larger Indo-European language family, to which Latin, Greek, the Germanic, Slavic, Romance, and other European languages belong. (The other Indo-Iranian subgroup is Iranian, spoken in Iran, Afghanistan, and parts of the Soviet Union.) According to some scholars, an easterly group of Indo-Iranian dialects was carried to India by invaders during the 2nd millennium BC. The root of the Indo-Iranian (*i.e.,* Indo-Aryan) languages of India is to be found in Sanskrit, the classical language of India, which ceased to be a generally spoken language many centuries ago. The various spoken dialects that evolved from Sanskrit are known as Prākrits.

Roots in Sanskrit

One of the oldest Prākrits, which like Sanskrit itself is no longer a generally spoken language but is still extant as a

static and sacred literary language, is Pāli; the Buddha used Pāli to convey his teachings to the general populace so that they could understand him without the intervention of the Brahmin caste, whose sacred language was Sanskrit.

Various new spoken dialects emerged as everyday use of Sanskrit and Pāli lapsed. Indian languages were then strongly affected by the great population movements that occurred in northern and western India in the 6th and 7th centuries AD; the pattern stabilized somewhat, however, around the year 1000, as the country's population became more settled. The basic ingredients of these Indo-Aryan languages were Sanskrit and *desh* (local dialect words absorbed largely from the languages of the invaders or from local Dravidian dialects).

Hindī, the most prominent of these languages, developed from the Prākrit that was prevalent from the Yamuna River to Bihār; Rājasthāni, Gujarātī, and Punjābī also derive from the same Prākrit. Sindhī, Kashmīrī, and Marāṭhi, however, stem from other Prākrits.

With the coming of the Muslims, Persian (Fārsī), an Iranian language, became the language publicly used in northern India from the 13th century onward. The Muslim influx promoted the rise of a hybrid tongue, Hindustānī (a dialect originating in the area between Delhi, Meerut, and Sahāranpur); it is spoken today in various parts of India outside the north. From colloquial Hindustānī, two literary languages—Hindī (showing a strong Sanskrit influence) and Urdū (with a Persianized vocabulary)—emerged.

The present-day Indo-Aryan languages of northern India are nevertheless all related, though with varying degrees of kinship, with Sanskrit. The chief of these Indo-Aryan languages are Hindī, Bengali, Punjābī, Gujarātī, Marāṭhi, Sindhī, and the Rājput dialects.

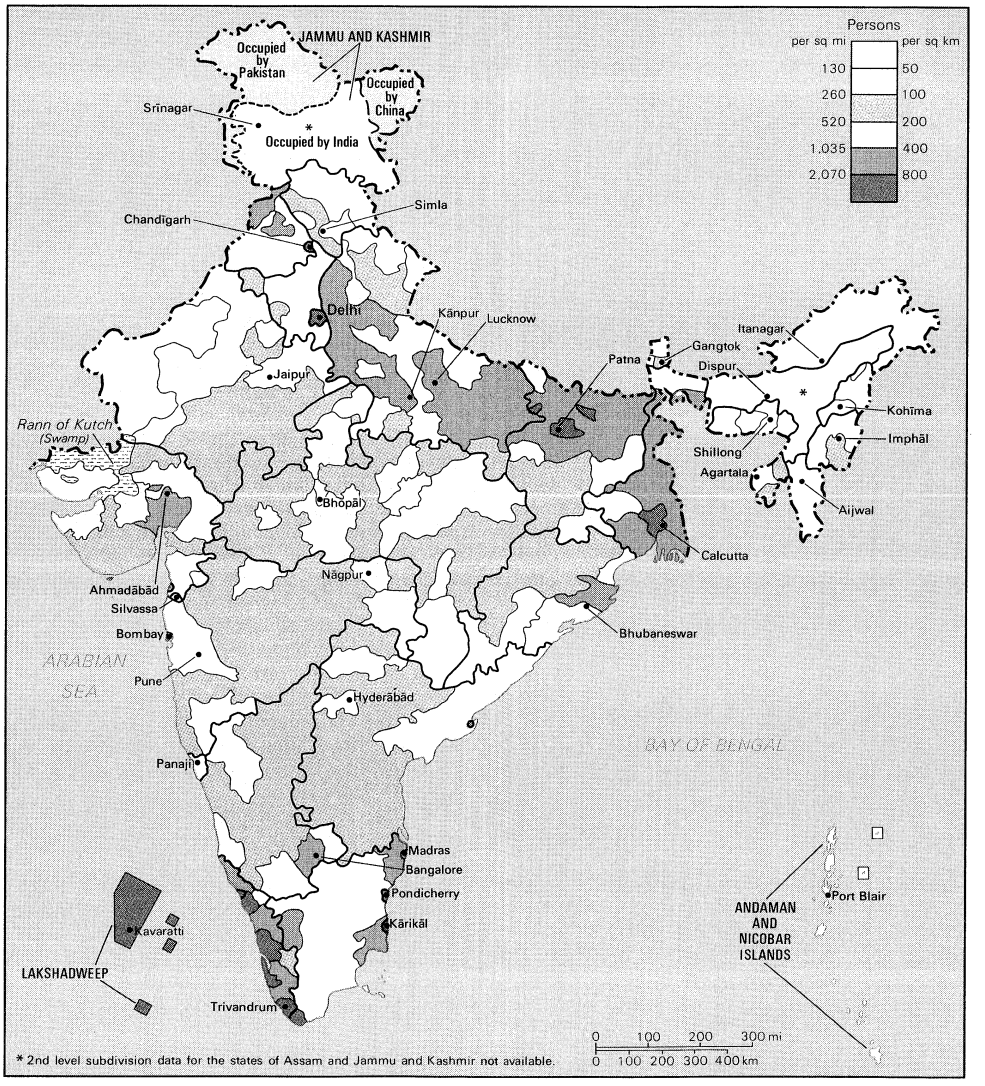
Hindī is spread over the greater part of the northern plain—east of Punjab, north of Gujarāt and Mahārāshtra, west of Orissa, and south of Nepal. It includes several languages and dialects, although strictly Hindī covers only the so-called Western Hindī dialects; these are Kanaujī, Bundelī, Braj Bhāṣā, vernacular Hindustānī, Jāṭū, and Hindustānī proper (the speech of Delhi).

Both Urdū and High Hindī (a refined form of Hindī, used in literature and by the educated) have rich literatures. Urdū, which dates to the 13th century, used the Perso-Arabic script and developed as a literary language in the 18th century under the patronage of the Mughal emperors. Hindī literature in High Hindī dates to the 18th century; the Hindī script is known as Devanāgarī.

Among other forms of Hindī are Kosalī, or Eastern Hindī, which differs from Western Hindī. It has three dialects: Awadhī (or Baiswārī), the language of medieval India, Baghelī, and Chhattīsgaṛhī.

The Magadhan language group

Yet other forms are included in the Magadhan group. The western Magadhan languages comprise Bhojpurī and Sadani (or Choṭā Nāgpurī), which is a dialect of Bhojpurī. Its speakers have adopted High Hindī as their language for education, public life, and literature. Central Magadhan consists of two tongues—Maithilī, current in northern Bihār, and Magadhī, current in southern Bihār. Maithilī has a remarkable literature, although the officially recognized regional language in both these regions is Hindī and, to some extent, Urdū. To the eastern Magadhan family belong Bengali, Assamese, and Oṛiyā. Bengali is spoken in



Population density of India.

West Bengal and has a highly developed literature. Assamese is current in the Brahmaputra Valley of Assam. Oṛiyā, though an archaic language, has quite an extensive literature.

Punjābī is divided into two linguistic areas—western Punjābī, which is spoken in the western districts of Punjab, and eastern Punjābī, spoken in and near Amritsar. The former has no noteworthy literature. Eastern Punjābī might be regarded as the standard form; it is written mostly in the Gurmukhī script, but Muslims show a preference for the Persian script, and Hindus for the Devanāgarī script. Differences of script were one of the major factors that led to the division in 1966 of the old Punjab state into the new Punjab (Sikh) state and Haryana (a Hindu state). Eastern Punjābī has developed a vigorous literature.

Rājasthānī is not a single language but a group of dialects. There are two important types, known as Ḍhuṇḍhārī (or Jaipur Rājasthānī) and Mārwaṛī, which is spoken at Jodhpur. Ḍhuṇḍhārī has not developed a literature, but Mārwaṛī (which closely resembles Rājasthānī) has. Subsidiary, in a certain sense, to these languages are Gujarātī and—more distantly—Marāṭhī. Gujarātī was akin to Rājasthānī until about AD 1500, when Rājasthānī became influenced by Western Hindī; after this, Gujarātī developed independently. Marāṭhī claims a literature that goes back to the 12th century. It has several dialects. Koṅkaṇī, spoken in and around Goa, is classified as a separate language by many linguists, as is Halbī, spoken in the Bastar district of Madhya Pradesh.

In the Indo-Aryan language group may also be included the Himalayan dialects, such as Pahāṛī and Himālī, which again are divided into various groups. These languages have little literature. Eastern Himalayan includes Nepālī, a developing language going back to the 17th century. Kashmīrī also belongs to this group and is essentially a Dardic dialect, belonging to the complex of Indic languages spoken in the upper valley of the Indus. There are two major dialects in Kashmir—Kashmīrī and Shīṇā; both are essentially branches of the Indo-Aryan group.

The Austro-Asiatic languages

The Austro-Asiatic languages are spoken by a substratum of the Indian population and are today a reminder of a largely forgotten past. These languages are spoken mainly by the tribal people and belong to the Muṇḍā, or Kolarī, group. They are still current in eastern and central India and to some extent in North Bengal and Assam. Allied dialects include Kherwāṛī, current in eastern India; Santālī, spoken largely by the Santāl group of aborigines; and Muṇḍārī.

The Austro-Asiatic languages have virtually no script, although European missionaries tried to reduce some of these languages and dialects to a form of script in the 19th century. One of the Austro-Asiatic languages is Khāsi, which has a Roman script introduced by European missionaries. The University of Calcutta recognizes three Santālī scripts—Devanāgarī, Bengali, and Roman.

The Sino-Tibetan languages are represented in India today by various offshoots of the Tai and Chinese language groups. Thus, for example, Khamti is spoken by a small tribe in the extreme northeastern frontier of India. There also exist some Tibeto-Burman dialects, such as Maṇipurī, or Meithei. These dialects were originally written in an alphabet of their own, but about the middle of the 18th century, they adopted the Bengali script. Beyond the Indian frontier, but to some extent seeping over the frontier, are Newāri, or Nepālī, the language current in Nepal and adjacent areas. Newāri uses the Devanāgarī Hindī script and has a literature that goes back to the 14th century AD. Another Tibeto-Burman language is Lepchā, spoken in eastern Nepal and Sikkim but also in Darjeeling, in India. It has a script of its own and is believed to be a branch of the Nāga group of the Tibeto-Burman family.

There are various minor languages that exist in India. Among these are Andamanese, spoken by the original inhabitants of the Andaman Islands; this seems to be a survival of an early Negroid language. Nicobarese is spoken by people of the Nicobar group of islands adjoining the Andaman group and seems to be a form of the Mon-Khmer speech of the Austro-Asiatic family of languages.

Various immigrant communities have adopted Indian languages as their mother tongue. Among these are the Parsis of Bombay, who speak Gujarātī, and the Bene-Israeli Jews of Mahārāshtra, who speak Marāṭhī.

Although it is not classified as one of the 15 officially recognized languages, English—which was introduced in the wake of the British conquest in the 18th and 19th centuries—is the only lingua franca shared by the Dravidian south and the Hindī north. The 1950 constitution of India stipulated that Hindī (in the Devanāgarī script) should be the common language of India and that Arabic numerals should be used as common numerals. It was also agreed that English should be retained as an official language until 1965, after which it was to be replaced by Hindī. In the event, however, opposition from the Dravidian south to the abandonment of English in favour of Hindī compelled another adjustment. English retained the status of an additional official language, with no definite date being set for its abandonment. The Official Languages Amendment Act of 1967 further provided that either Hindī or English should be compulsory languages for recruitment to the central government service. English is to be used in the Supreme Court and in the High courts. Correspondence in English was obligatory, however, only between the Union government and a non-Hindī-speaking state. Hindī states dealing with non-Hindī states could send their communications in Hindī with English translations.

RELIGIONS

India is the birthplace of many religions and has become the home of many others. Through the ages Indians have shown strong proclivities toward a religious outlook, resulting in the development of indigenous religions, chief among which are Hinduism, Jainism, Buddhism, and Sikhism. This religious bent has also encouraged the growth of such extraneous religions as Islām, Christianity, Judaism, and Zoroastrianism.

The oldest of all Indian religions is animism, which is still practiced by the more remote tribes, such as the Santāls, Bhīls, and Gonds. Hinduism is, however, above all the traditional religion of India; some date its origin back to the Vedas of the Aryans, between 2000 and 1500 BC. There is evidence, however, that Hinduism, as it later emerged, is not the religion preached in the Aryan Vedas but contains Dravidian elements going back to a pre-Aryan age, thus representing a fusion of two cultures. The great majority of the population of India is Hindu.

Hinduism

While Hinduism has been a great unifying force, its association with the caste system has also made for a sense of separateness. Whatever its origin—whether induced by race, colour, occupation, or simply conquest—caste has constituted a major divisive force throughout Indian history and has even permeated non-Hindu groups.

Jainism and Buddhism were the first offshoots of Hinduism, coming into being in the 6th and 5th centuries BC, respectively. Jainism survives today largely in Gujarāt and Rājasthān. Its principles of nonviolence and asceticism have left their imprint on Hindu thought; Mahatma Gandhi was deeply influenced by these two principles.

Buddhism, a younger rival of Jainism, spread over India largely as a result of the example of Aśoka, the great Mauryan emperor of India who embraced this religion. After his death, the influence of Buddhism steadily declined, being superseded by a resurgent Hinduism. Buddhism, however, spread to Tibet, China, Japan, Korea, Mongolia, and other lands, thus becoming a world religion; Jainism remains confined to India.

Of the religions that are extraneous to India, Islām, primarily introduced by conquest, struck the deepest roots. First brought into India by foreign traders, it gained strength during the Muslim invasions; after the establishment of the Mughal Empire in the 16th century, the greater part of India became Muslim dominated. Muslims today constitute more than 10 percent of India's population.

The other foreign religions are Christianity, Judaism, and Zoroastrianism. Judaism was introduced by early Jewish traders who established settlements in the coastal towns, one of their principal centres being Cochin; Jews today form a microscopic proportion of India's population. Zoroastrianism entered India between AD 700 and 800

with the influx of Persian refugees fleeing from Muslim persecution; the descendants of these refugees, the Parsis, are today concentrated largely in Bombay. Christianity claims to date back to AD 52, when Thomas, one of the Apostles of Christ, is said to have landed on the west coast of India, where he established a few churches; according to tradition, he then travelled from the west coast to the east coast, where he was martyred at Mylapore, in the city of Madras. The descendants of the Christians he converted are known today as Syrian Christians. A new era of Christian proselytization began with the advent of the Europeans in India; however, the Christians form only a small percentage of India's population and are concentrated mostly on the west coast.

POPULATION

By the late 20th century there had been a steady though slow increase in life expectancy during successive decades from 1891 onward, but it had accelerated considerably during the 30 years between 1931 and 1961. The life expectancy at birth for males in India is currently 54 years and for females is 53 years. In general, life expectancy is higher in northern India and lower in central and eastern India.

One of India's major problems—if not its major problem—is population control. A network of family-planning centres, mostly in rural areas, has been set up, giving family-planning advice and distributing free contraceptives. After extensive trials, the placing of a small loop in the uterus has been found to be the most suitable method of birth control for India. Sterilization operations have increased, but not to any marked degree. All family-planning services are provided free by the government, but among various communities there is a disinclination on the part of the more conservative elements to utilize these services. (F.R.M.)

The economy

India's economic structure

India is experiencing relatively slow economic progress. But India is not typical of most developing countries, in that it has a sophisticated administrative and political structure, a well-equipped bureaucracy, a large supply of educated manpower, and a considerable transport and communications network. Since independence in 1947, it has made great progress and undertaken large investments. India has acquired a sophisticated industrial base that has placed it among the top 20 industrial nations of the world. This has yet to be reflected in its foreign trade—its exports account for less than 1 percent of total world exports. The country suffers enormous disparities in incomes, with the rich being very rich and the poor very poor. A small elite enjoy very high standards of living, while the bulk of the population has incomes below the subsistence level. A vast, relatively affluent middle class, estimated to be the largest in the Third World, is gradually emerging, however.

India's fundamental economic problem remains one of ever-growing millions of people to feed, clothe, shelter, and educate and of ever-growing numbers of men and women who join the ranks of the unemployed. The rate of economic growth will have to be increased if India is to begin to overcome this problem. But it is apparent to most that growth alone will not help, unless investments are specifically directed to the task of creating more jobs through an emphasis on sectors making intensive use of labour, such as agriculture, and by applying technologies that make use of India's natural resources.

RESOURCES

Mineral resources. India is fairly well endowed with mineral resources. It has vast reserves of coal and iron, and it also has reserves of manganese ore. But reserves of nonferrous metals, with the exception of aluminum, are not enough to meet domestic needs. India completely lacks elemental sulfur and phosphorus.

It is clear that coal will be the country's primary fuel for years. Fears expressed that India would use up her coal reserves within a few decades are unfounded. The country has enough noncoking coal to meet its requirements, unless, of course, the initial rate of exploitation increases substantially—for example to the United States level of consumption—or the growth rate of the economy becomes increasingly rapid. Coking-coal resources are, however, insufficient to match India's iron-ore resources.

The country's coal reserves are not evenly distributed all over the country. The workable or potential coal mines are concentrated in Bihār, West Bengal, Madhya Pradesh, and Andhra Pradesh. Coal required in southern and western India has consequently had to be carried for several hundred miles.

Apart from coal, India has lignite reserves. Sizable deposits have been located in Tamil Nādu, Rājasthān, and Gujarāt. The Tamil Nādu reserves are being systematically exploited by means of an industrial complex that produces power, fertilizers, and briquettes.

India is one of the world's leading iron-ore exporters. The country's most extensive iron-ore deposits are found in Bihār, Orissa, Madhya Pradesh, Karuātaka, Goa, Andhra Pradesh, and Tamil Nādu; quality is generally very good.

India has vast manganese-ore deposits, concentrated mainly in Madhya Pradesh and Mahārāshtra; some are also scattered in Bihār, Orissa, Karuātaka, Andhra Pradesh, and Goa. There is a considerable amount of chromite in the country, but there has been no systematic exploitation of it. India produces only a small amount of titanium (a very hard metallic element used to toughen steel), in spite of ample reserves of rutile and ilmenite—both of which are sources of titanium.

India is desperately short of nonferrous metals, excluding aluminum. The scarcity is not because of a deficiency in resources. Investigations by the Geological Survey of India have revealed important nonferrous metal deposits in the Agnigundala region, in Andhra Pradesh; in the Khetri region, in Rājasthān; in Gujarāt, Mahārāshtra, and Madhya Pradesh; and at Sukinda Khās, in Orissa. But, until the estimates of these reserves are known and exploitation begins, India will have to depend largely on imports to meet her needs. Moreover, scientists at the National Metallurgical Laboratory are experimenting with the possibility of substituting aluminum for copper in the electrical industry, for zinc in galvanization, and for nickel and copper in coinage.

India does not have deposits of elemental sulfur, but sulfur can be extracted from pyrites found at Amjore, Bihār. Phosphorus has been found in the form of rock-phosphate deposits in a long belt near Udaipur, in Rājasthān, near Jhabua, in Madhya Pradesh, and near Mussoorie, in Uttar Pradesh.

India has few known petroleum resources. Existing oil fields in Gujarāt may eventually be supplemented or replaced by shallow offshore drilling near Aliabet in the Gulf of Cambay and deep offshore drilling over a geological structure named "Bombay High," off the Gujarāt coast. Some deposits of oil have also been discovered in Lokawa, Galeki Naharhabi, and Kusijan, in Assam and in Sisodra in Gujarāt.

Because of low oil output, the country has to import about half of its crude-oil requirements.

Biological resources. About half of India is cultivated, and about one-fifth is forested, while the rest is uncultivated. The quality of the soil is generally poor, and soil erosion is common in some parts of the country. Efforts at soil conservation and improvement have been woefully inadequate to meet the need that exists. About three-fourths of the total land area under forests is "reserved" or "protected." The remaining area consists of unclassified forests not generally under systematic management. The forest area is unevenly distributed; forests are most scarce in areas where most needed—for instance, in the densely populated and intensively cultivated Gangetic Basin. Only the Himalayan belt and a few mountainous areas have thick forests. The fact that most of the plains are bereft of any worthwhile forest cover has led to some desiccation of the climate and vegetation; this in turn has had an adverse effect on precipitation from the monsoon clouds.

Forest land acreage

Among the economically important species of trees are teak, sal, bamboo, and the coniferous group. A total of about 95 percent of the forests are owned by the govern-

ment, about 3 percent belong to *pañcāyat*s, while only about 2 percent are in private hands.

Although India has the largest livestock population in the world, milk yields are extremely low. The buffalo is the main milk producer; cows produce another large percentage of the marketed milk, while only about 5 percent is produced by goats, sheep, and camels. The low yield of milk is due to the poor quality of the stock, highly inadequate fodder resources, the limited grassland area, and old-fashioned methods of livestock management. It has been estimated that the present feed and fodder resources are sufficient only for about two-thirds of India's livestock strength, which is widely distributed throughout the country.

India produces quantities of wool every year from about 40,000,000 sheep, but the country's sheep are poor wool yielders. The average annual yield of greased (uncleaned) wool from a sheep is less than two pounds (0.9 kilograms), with the range extending between 0.75 and five pounds. This output is very poor compared with that in Australia and New Zealand, where the annual yield per sheep is from eight to 12 pounds. The bulk of Indian wool is coarse, containing a high proportion of hair and kempy (short, coarse, and brittle) fibres that are mainly suitable for making carpets. Most wool-yielding sheep are in the arid region of the plains of Rājasthān, Gujarāt, and Haryāna and the Himalayan region. The sheep of peninsular India have a low yield of wool that is of poor quality.

Hydroelectric and other power resources. India has made considerable progress in the development of power since 1950. Hydroelectric power constitutes the most economic source of electricity production in the country. The capital outlay on hydroelectric stations is comparable with (and in some cases lower than) that on thermal power stations and is very much lower than that on nuclear stations. The cost of energy from hydroelectric stations is significantly lower than the cost of energy from other sources.

AGRICULTURE

Farming, forestry, and fishing account for roughly one-half of India's gross domestic product at current prices and are sources of livelihood for about four-fifths of the entire population.

Rice production

Crops. Rice is India's most important food-grain crop. The area under rice production is approximately one-third of all the area under food grains. Rice production has not been rising very rapidly, largely because of the slow rate of acceptance of new high-yielding varieties. Wheat, India's second food-grain crop, has shown a remarkable increase in production but accounts for about one-fifth of the total acreage under food grains; hence, India is still not quite self-sufficient in food. The increase in wheat output is largely due to the successful evolution of new wheat varieties, both foreign and domestic strains, and to the rapid expansion of minor irrigation in the form of tube (driven) wells in the major wheat-growing areas. In the northwestern state of Punjab, the proliferation of tube wells, largely reduced the dependence on the monsoon. Other food-grain crops include *jowār* (sorghum), corn (maize), *bājrā* (a type of millet), *ragi* (a cereal grass), barley, and small millets. Of these, *jowār,* corn, and *bājrā* are by far the most important.

In the 1950s, non-food-grain production had been rising faster than food-grain production, but, in subsequent decades, it began to lag behind. The most important non-food-grain crop is oilseed. Other major non-food-grain crops are sugarcane, raw cotton, jute, tea, coffee, rubber, and tobacco. India, together with Bangladesh, accounts for the bulk of the world's jute production and, with Sri Lanka (Ceylon), accounts for the bulk of the world's tea production.

Indian agriculture is still largely traditional; it is dependent on the vagaries of the weather, and the land under cultivation is extremely unevenly distributed. Only one-fourth of the net sown area is irrigated, and less than two-thirds of the potential arable area can be irrigated from potential surface and ground-water resources. A very large percentage of farmers are concentrated on a very small percentage of the total area under cultivation; agriculture typically suffers from lack of water and credit, lack of incentives because of the land-tenure system, and fragmentation of holdings. Smallholders, defined as those with less than five acres (two hectares), account for a majority of the rural households and a fifth of the landholdings. The Green Revolution (a slogan referring to spectacular increases in agricultural production) is largely confined to land with assured water supply and to farmers with large holdings. This is, in fact, a direct consequence of the government's Intensive Agricultural District Programme started in 1960, in which districts adjudged the most responsive to development programs were chosen for concentrated development. This strategy has, in fact, worked in terms of production. But, because the beneficiaries are a small number of large farmers, it has also accentuated disparities in rural incomes, resulting in new social and economic problems that have begun to produce new political responses. To quote the Indian Planning Commission, there is danger of a "sharp polarisation between the more privileged and less privileged classes in the rural sector, the privilege in this instance relating to the resources and tools of development."

Fishing. With a long coastline and many major river systems, there is an enormous potential for the development of fisheries, which remains largely untapped. Little has been done so far to develop either inland fishery or deep-sea fishing in the Indian Ocean.

Forestry. Despite the fact that forests occupy about one-fifth of the land surface of the country, the contribution of forestry and logging to net domestic production has been very low.

MINING AND QUARRYING

India is an important producer of iron ore and coal, but its mineral development has yet to reach a level that would meet domestic and export demand. As a consequence, the mining sector still contributes very little to gross national product. Imports of petroleum, nonferrous metals, and nonmetallic minerals are costly for India. With some exceptions, the mineral industry is in the public sector. Responsibility for exploration largely rests with the Geological Survey of India, the Mineral Exploration Corporation, and the Indian Bureau of Mines and—in the case of petroleum—with the Oil and Natural Gas Commission, Oil India Ltd., and the Assarn Oil Company Ltd. A little over one-third of the output of India's refineries is from indigenous crude oil, the rest being imported.

There is a well-developed aluminum industry, which is distributed in both the public and private sectors. Attempts to develop India's copper and zinc resources are also being made, but progress is slow. It is expected that iron ore might become India's most promising metal.

INDUSTRY

In the 1950s and 1960s—the first two decades of economic planning—India succeeded in acquiring a sizable and diversified industrial base. In the late 20th century the industrial structure was still dominated by consumer-goods industries; in this sector the most important industries were cotton weaving, tea, sugar, and pharmaceuticals, and cotton spinning was also significant. Capital-goods industries accounted for just over one-tenth of production by weight, and, of these, railroad equipment and motor vehicles were the most important.

India has a relatively sophisticated industrial structure, with major investment in steel, machine tools, fertilizers, cement, aluminum manufacturing, industrial machinery and components, and electrical and transport equipment. Progress in industries making intermediate goods (such as tires and synthetic fibres) and consumer goods has been slower, but India's requirements for consumer goods are met almost totally by domestic production, and the government's policy is to encourage self-sufficiency in as many areas as possible. Consumer-goods industries are almost entirely privately owned, but most of the investment in the capital-goods and basic-industries sector is the responsibility of the state. Most of the investment in public-sector enterprises is in steel, engineering, chemicals, petroleum, mines, and minerals.

Industrial production in the public sector is approximately one-tenth of the total production of organized industry. Although the Indian government is committed to the concept of a mixed economy, private industry is viewed with some suspicion, and relations between government and industry are not always smooth. The private sector has been given considerable encouragement; but, as a result of the operation of the industrial licensing system and prevailing tax laws and import controls, considerable distortions have been introduced into the ownership pattern of Indian industry. Private industry is still largely in the hands of a few family business houses, the most notable of which are the Birlas and Tatas. These business houses formerly dominated Indian industry through a system of managing agencies that have now been closed down by the government. But the pace of introduction of professional standards of management is slow, and there is considerable inefficiency in production. Excessive concern over concentration of industry has led to disregard for the benefits offered by a mass-production economy. High tariff walls also protect considerable inefficiencies in production. The vast expansion of industry has meant relying on a first-generation labour force that is extremely vulnerable to political influence, so that labour unrest and low productivity are often manifest. Despite these teething troubles characteristic of a predominantly agricultural society that is trying to industrialize, there has been considerable progress; this is because of the availability of a plentiful supply of educated manpower, the existence of a highly developed administrative system, and large investments in transport and power.

ENERGY

India's energy resources are adequate but not abundant. The Himalayas and smaller mountain ranges in other parts of the country have an estimated potential of between 80,000,000 and 100,000,000 kilowatts of commercially exploitable hydroelectric power. Coal is abundantly available, but most of it is located in Bihār and West Bengal, and its high ash content makes it difficult to use. Known indigenous crude-oil resources are few, and the rising oil prices on the world market add to the foreign-exchange constraint inhibiting imports. The growth of road transport and the electrification policy of the railways, as well as the introduction of diesel engines, are partly responsible for the decrease in coal consumption in the transport sector, which altogether accounts for one-fourth of India's total energy consumption. In rural India, cow dung is still the basic household fuel, closely followed by wood—a fast-depleting source. Biogas (methane gas produced from cow dung) is fast becoming an important source of household fuel in rural areas. Kerosene and electric power are gaining wide acceptance. In cities kerosene, natural (cooking) gas, and electric power are the major sources of household fuel.

India has a well-organized system of generating, transmitting, and distributing power. The sources of electrical power are coal, oil, natural gas, atomic fuel, and water. The most important source was low-grade coal. Thermal-production capacity was a little more than half the total power capacity and was followed closely by hydroelectrical power. There are nuclear power stations at Tārāpur (Mahārāshtra), and Rānā Pratāp Sāgar (Rājasthān), and Kalpak-kam (Tamil Nādu).

FINANCE

Banking

India has a relatively highly developed banking and financial system. The Reserve Bank of India, as the central bank, strictly regulates all banking activities through both general and selective control measures. The important commercial banks in India were until recently owned by the large industrial houses and supplied all the short-term credit needs of trade and industry to the neglect of agriculture and small-scale industry. It was with a view to diversifying and enlarging the scope of banking operations that the government nationalized the 14 largest commercial banks in the country in 1969, six more commercial banks were nationalized in 1980. Foreign banks and banks with deposits of less than $66,700,000 were not nationalized. About 90 percent of the banking industry is in the public sector. Ever since nationalization, credit to agriculture and small-scale industry has grown rapidly, but industry and trade have faced tight credit conditions. Bank branches expanded rapidly after nationalization.

Apart from bank credit, industry depends for its short-term requirements on public deposits at high rates of interest. For medium- and long-term capital investment, industry draws on several public industrial financial institutions, at both the central governmental and state level, and the Life Insurance Corporation. For their credit requirements, agriculturists still largely rely on the village moneylender, who, although he charges exorbitant interest rates, demands no security and is easy about repayments. To combat him, the government has encouraged the growth of cooperative credit societies. But these societies are poorly run, are dominated by large farmers and, in most states, offer little help to the small farmer or tenant.

Supplementary credit is available from Central and State Cooperative banks, and long-term credit from Land-Development banks. Commercial banks are now becoming an important source of credit, which is also available from the Agricultural Refinance and Development Corporation and the Rural Electrification Corporation.

India has a relatively underdeveloped capital market. Although there are several stock exchanges in the major cities (of which Bombay, Calcutta, Madras, Delhi, and Ahmadābād are the most important), companies have been reluctant to come forth to raise money. Tax laws encourage reinvestment and discourage declaration of dividends. The existence of high tax rates, which is responsible for the large money black market, deprives the capital market of adequate funds. The excess of demand far over the supply of reliable stocks accounts for the considerable capital appreciation enjoyed by some major stocks.

A few reliable companies are over-subscribed several times, while the majority of issues are under-subscribed. Institutional buyers such as the Life Insurance Corporation dominate the market. Both central and state governments rely heavily on market borrowing for their budgetary needs. The central government meets its public debt through market loans and the sale of treasury bills and long-term securities. Small savings are also an important source of funds, obtained through national savings certificates and post-office deposits.

The growth of savings and banking operations has been checked by the pursuance of a low-interest-rate policy that is designed to encourage investment but that is handicapped because it ignores the scarcity of capital and discourages the movement of funds from the larger unorganized money market to the organized money market.

TRADE

Exports

Foreign trade constitutes a very small proportion of India's gross national product: India's share in world exports is less than 1 percent. In the 1950s, traditional items such as jute manufactures, tea, and cotton textiles dominated India's exports. Thereafter, the pattern of Indian exports changed rapidly. While jute and textile exports remained stagnant, iron ore, iron and steel, and engineering goods became major export items. Cooperation between government and industry is more noticeable where exports are concerned than it is in other fields.

India's imports consist largely of raw materials, intermediate manufactures, and capital goods. Frequently, India has imported large quantities of cereals and cereal preparations. Of the raw-material and intermediate imports, the most important are raw cotton, petroleum, oils and lubricants, fertilizers and fertilizer materials, iron and steel, and nonferrous metals. In the capital-goods sector, the biggest imports are of nonelectrical machinery, apparatus, and appliances.

India's balance of payments is often in deficit because of the slow growth of exports and the large imports necessitated by a growing economy requiring substantial imports of petroleum (more than one-third of the total imports), fertilizers, and finished iron and steel products. As a consequence, the government follows a policy of strict import controls, relying on foreign aid to keep its payments po-

sition in equilibrium. But net foreign aid has declined sharply in the late 20th century, partly because less aid is available and partly because the debt burden has grown.

The largest amount of aid—much of it for food—has been received from the United States. The United Kingdom, West Germany, the Soviet Union, Canada, and Japan are other important donors.

MANAGEMENT OF THE ECONOMY

The private sector. The government regulates and controls large-scale industry by means of specific price and allocation controls and general capital issues, its industrial licensing policy, and the Monopolies and Restrictive Practices Act. Thus government, in effect, affects the private decision-making process at every level, determining such matters as the salaries paid to a company's directors, the nature of the company's imports, the location of its plant, and the amount of capital it can raise. These controls are reinforced by elaborate taxation laws that discriminate against companies with very few shareholders as well as against intercorporate investment. All this has not succeeded in checking the growth in the concentration of economic power in the hands of a few large business families.

Relations between government and private industry

The compulsions of populist politics have forced the government to take a strict line with private industry and to discourage its growth in key and basic industries. Government policy broadly encourages small- and medium-sized industry and confines the commanding heights of the economy to the public sector. Relations between government and industry have frequently been strained. This has been partly because of an excessive concern with profits in the private sector and because of a failure on the part of government to recognize the importance of the profit motive and of the economic benefits resulting from large-scale enterprises. The reality is that government regulations have worked to the advantage of a certain breed of industrialists who have thrived in a regime of controls that they soon learned to manipulate to their advantage. Such industrialists dominate the private sector; they dislike competition and fight against any measures of import liberalization or the removal of licensing controls that would introduce an element of competitiveness. Since the bureaucracy also thrives on controls, there is now a strong vested interest in favour of controls.

The public sector. The public sector is expected eventually to emerge as a dominant area of the economy. Rail, air, and sea transport, power generation, banking, insurance, petroleum, steel, mines, minerals, and heavy engineering are predominantly in the public sector. The vast increase in state investment in industry has not been without its problems. The rate of return on capital employed is tiny and well below a reasonable economic return under Indian conditions. Considerable public-sector capacity is under-utilized. Expectations of a significant contribution to planned resources from the initial resources of these enterprises are being disappointed. In fact, several enterprises are relying on budgetary support to meet losses. There are several reasons for this poor performance. First, because much of the investment was in heavy industry, returns were necessarily slow to come. Second, public-sector companies were highly prone to political interference. This led to location decisions geared to political rather than economic considerations, to huge investments in worker's welfare amenities, to an inflated labour force, and to unhappy labour relations. Third, management of public enterprises was in the hands of bureaucrats seeking sinecures, rather than of professional managers. Fourth, these enterprises were subjected to close parliamentary supervision, which grossly violated their autonomy.

Taxation. In the late 20th century, tax revenue produced about one-seventh of the national income. This low burden of taxation was largely due to the fact that the agricultural sector bore no direct taxation, and that commodities that entered into agricultural consumption were only lightly taxed, if at all.

Direct taxation of nonagricultural incomes was in the hands of the central government. Indirect taxes were at both central and state levels. A formula laid down by the Finance Commission laid down the state's share in central taxation. Tax revenue accounted for less than four-fifths of revenue of all central and state governments.

The tax revenue of the central government was roughly three-fourths of total revenue. Only one-fourth of this came from direct taxes, of which income and corporation taxes accounted for the bulk. Union excise duties accounted for more than three-fourths of all indirect tax revenue and more than 60 percent of all tax revenue. Of these, the excise duties on petroleum products, tobacco, and sugar accounted for half the revenue from excise duties.

The burden of direct taxation bears somewhat heavily on individuals and, with the combination of wealth and other taxes, can exceed 100 percent. Marginal income tax rates have risen to as high as 97 percent. There is consequently considerable tax evasion. To counter income tax evasion and to check the circulation of black market money, two Income Tax Amendment acts came into force in 1981. Since agricultural incomes are not taxed, there are roughly 5,000,000 income-tax payers. The burden of taxation on corporations is theoretically very high, but the effective burden is low because of various incentives or rebates.

Trade unions and employer associations. Workers in organized industry are protected under elaborate labour laws that conform closely to International Labour Organisation (ILO) models, and, in both the public and private sectors, comprehensive labour-welfare measures are becoming the rule rather than the exception. The principle of collective bargaining is widely accepted. Disputes between employers and workers are increasingly being settled through adjudication by tribunals and courts and by recourse to the legal machinery—thus greatly delaying settlements. The existence of a large small-industry sector has resulted in unfair labour practices by employers, who take advantage of the vast amount of unemployment in the country and the inability of workers to organize themselves into unions.

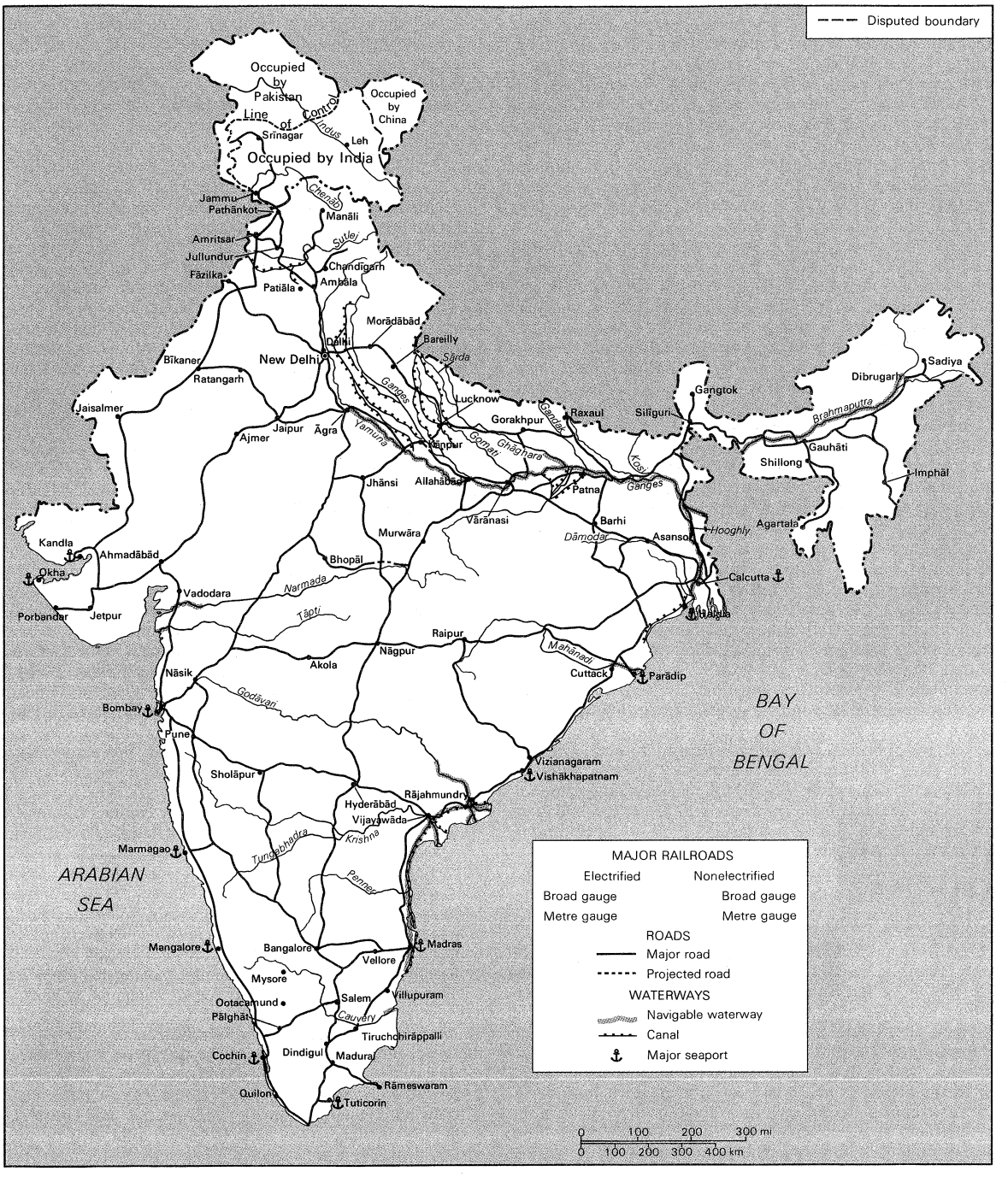
Relations between management and labour

The All-India Trade Union Congress was founded as far back as 1920, and a Trade Unions Act was passed in 1926 to give formal recognition to the right of workers to be recognized. The unions increased rapidly after independence as political parties sought to organize labour for their own ends. The percentage of union membership varies enormously from industry to industry. Unionization is highest in the coal, tobacco, cotton textiles, and iron and steel industries and in banks. Since any seven persons can constitute a union and since the prescribed membership fee is small, there are more than 7,000 registered trade unions with an average membership of more than 500. Approximately one in every four industrial workers is unionized.

There are five major central trade-union organizations, of which the two most important are the Indian National Trade Union Congress, affiliated to the ruling Congress Party, and the All-India Trade Union Congress, dominated by the pro-Moscow Communist Party of India. A third growing force is the Centre of Indian Trade Unions allied to the pro-Peking Communist Party of India (Marxist). Political rivalry has prevented a united labour position on national issues. Unions generally cover all workers, without distinction of craft or category. There are all-India unions covering several important industries. Some craft unions exist in air transport, ports, and docks. Interunion rivalry, the primary cause of most industrial disputes, generally revolves around the issue of recognition of the union as the official and sole bargaining agent.

Employers are generally organized through industrial associations at regional and national levels. The regional bodies are affiliated to the Central Industrial Organisation. The development of these organizations has been encouraged by the government in order to help policy makers consult industry on various issues. The three major employers' organizations are the Employers' Federation of India, the All-India Organisation of Employers, and the Standing Conference of Public Enterprises. Multiplicity of organizations has not been a problem for employers, and the three employers' organizations coordinate their activities through the Council of Indian Employers.

The government in India has always played an interventionist role in industrial relations by enacting both procedural and substantial laws and by setting up conciliation, arbitration, and adjudication machinery. (E.I.U./Ed.)



Surface transportation system of India.

TRANSPORTATION

Little attention was paid to road development until the 1920s, mainly because the government had previously focussed its attention on railways. In 1927 it was decided that road development was beyond the capacity of the provincial governments and that the central government should consequently bear a substantial part of the cost. In the British era, roads were developed largely for strategic or commercial reasons. In 1943, the road mileage in India totalled roughly 220,000 miles, and it was then decided that this total should be increased to 331,000 miles, of which 123,000 miles would be paved. This project, known as the Nagpur Plan, formed the basis for future road-development plans.

Since independence there has been considerable road-building activity. Following the Nagpur Plan, roads have been classified as (1) national highways, (2) state highways, (3) major district roads, and (4) minor district roads and village roads. The national highways connect capitals, ports, and important towns and constitute the main lines of communication; strategic roads for defense are also included in this category. State highways are major roads, connecting each state capital with major towns in the state, as well as with national highways. District roads connect district headquarters with important areas within the region and link up with the national and state highways or the railways that pass through the district. Minor district roads and village roads connect various villages to one another.

Railways did more than anything else to unify and modernize India. The Indian railway system, with a total length of a little more than 37,000 miles, is the fourth largest railway network in the world, ranking after the United States, the Soviet Union, and Canada. It is also the largest public sector enterprise in the country. India's railways are the direct responsibility of the minister of railways; he is assisted by the Railway Board, which is responsible for day-to-day administration.

Water transport

Transport by road and rail is largely integrated, as also is transport by inland waterway (until recently the cheapest form of transport in the country). The expansion of the railway system and the development of road transport have, however, reduced the importance of water transport. Even so, in some parts of India, notably West Bengal and Kerala, water transport is still important. There are about 9,000 miles of navigable waterways, of which the most important include the Ganges and the Brahmaputra in West Bengal, the Godāvari and the Krishna in Andra Pradesh, the Cauvery in Tamil Nādu, and the backwaters and canals of Kerala and Orissa.

Airmail service in India was started on an experimental basis in 1920, after which interest in aviation increased considerably. Between 1930 and 1945, Indian civil aviation expanded rapidly; by 1945 there were as many as 11 air companies in operation, carrying passengers and freight. In 1953 the government nationalized air services, establishing two corporations—Indian Airlines (for domestic service) and Air India (for international service). Since 1981 a third airline, Vayudoot, has been providing service to previously inaccessible places in northeastern India.

Shipping supplements these national transport facilities. As a step toward securing a major share of the Indian shipping trade for Indian ships, the coastal trade of India was reserved for Indian vessels by two acts—the Control of Shipping Act (1947) and the Merchant Shipping Act (1958). There are several prominent Indian shipping lines; since independence, shipbuilding has been increased, both for commercial and for military reasons.

The major ports

India's major ports are Calcutta, Haldia, Paradip (Orissa), Vishākhapatnam, Madras, and Tuticorin on the east coast, and Kandla (Gujarāt), Bombay, Mormugao (Goa), New Mangalore, and Cochin on the west coast. These ports are administered by statutory autonomous port trusts except for New Mangalore, which is administered directly by the union (central) government. The Indian coastline is also served by about 225 intermediate and minor ports, of which about 180 are in active use. The administration of these ports is the responsibility of the state governments. The National Harbour Board advises the central and state governments on the coordination of port development, with special emphasis on that of minor ports.

Administrative and social conditions

GOVERNMENT

Form of government

Constitutional framework. India is a sovereign democratic republic. It is variously called the Republic of India, the Union of India, and the Indian Union. The constitution of India, which came into force on January 26, 1950, is the lengthiest written constitution in the world, containing much that in other countries is to be found in statutory rather than in constitutional law. It was promulgated on the day that India was proclaimed a republic—the first republic to achieve membership in the Commonwealth and to acknowledge the British monarch as head of that organization. This arrangement has undergone several revisions, the amendments generally following certain decisions of India's Supreme Court. The constitution has been described as a quasi-federal document providing for a unitary state and having subsidiary federal features rather than a federal constitution with subsidiary unitary features.

India's constitution derives from several sources but was mainly influenced by the British concept of parliamentary government and—in its definition of certain presidential powers and in its adoption of the constitution of the federal Supreme Court—by the United States government. The Indian president, however, is more a constitutional sovereign than a chief executive. Other sources were the Canadian constitution, upon which India based its quasi-federal structure, and the Irish constitution, to which its directing principles owe much.

The broad framework on which the Indian constitution is largely based is the Government of India Act (1935; enacted by the British government), which contains no constitutional principles but a great number of administrative details. Broadly, the distribution of powers between the centre and the states follows the accepted principle that those matters which, for convenience and efficiency, ought to be administered on an all-India basis are vested in the union (central) government, while other matters—mainly those that fall within the nation-building category—can conveniently be administered by the state governments. Thus, defense, foreign affairs, transport and communication, currency and coinage, and the administration of justice at the higher level are all central subjects. State powers include powers over the police, public health, education, forests, and similar matters of primarily local concern.

Certain fundamental rights are guaranteed to citizens, although those relating to property have recently been whittled down. Freedom of speech and expression is guaranteed, and preventive detention is forbidden except in certain cases (primarily concerned with attempts to overthrow the government by armed force). India is a secular state and, as such, upholds freedom of conscience. Untouchability (the state of belonging to the hereditary "untouchable" group, having in traditional Hindu belief and practice the quality of defiling members of a higher caste by contact) is constitutionally abolished, although in practice it still prevails.

The 25 states (with the capital towns in parentheses) are (1) Andhra Pradesh (Hyderābād); (2) Arunāchal Pradesh (Itānagar) (3) Assam (Prāgjyotiṣapura); (4) Bihār (Patna); (5) Goa (Panaji); (6) Gujarāt (Gāndhīnagar); (7) Haryāna (Chandīgarh, which is the capital of both Punjab and Haryāna); (8) Himāchal Pradesh (Simla); (9) Karnātaka (Bangalore); (10) Kerala (Trivandrum); (11) Madhya Pradesh (Bhopāl); (12) Mahārāshtra (Bombay); (13) Manipur (Imphāl); (14) Meghālaya (Shillong); (15) Mizorām (Aizawl); (16) Nāgāland (Kohīma); (17) Orissa (Bhubaneswar); (18) Punjab (Chandīgarh); (19) Rājasthān (Jaipur); (20) Tamil Nādu (Madras); (21) Sikkim (Gangtok); (22) Tripura (Agartala); (23) Uttar Pradesh (Lucknow); (24) West Bengal (Calcutta); and (25) Jammu and Kashmir (Srīnagar). Control over the disputed state of Jammu and Kashmir is divided between India and Pakistan on either side of a "line of control" agreed upon by both countries in 1972.

The seven union territories, with their capitals in parentheses, are (1) Andaman and Nicobar Islands (Port Blair); (2) Chandīgarh (Chandīgarh); (3) Dādra and Nagar Haveli (Silvassa); (4) Daman and Diu; (5) Delhi (Delhi); (6) Lakshadweep (Kavaratti); and (7) Pondicherry (Pondicherry).

The union executive

The centre, or union executive, consists of the president, vice president, and the council of ministers, headed by the prime minister. The president is elected by an electoral college, consisting of elected members of both houses of parliament and of the legislative assemblies of the states. The president must be a citizen of India, not less than 35 years old, and eligible for election as a member of the lower house of parliament. The president, whose term of office is five years, is eligible for reelection. The vice president is also elected by an electoral college, consisting of both houses of parliament. The Council of Ministers, responsible to the House of the People, is headed by the prime minister; it comprises ministers who are members of the cabinet, ministers of state who are not members, and deputy ministers.

The legislature of the union government, which is called Parliament, consists of two houses—the Council of States (Hindī: Rājya Sabhā) and the House of the People (Lok Sabhā).

The House of the People consists of not more than 525 members directly elected from territorial constituencies in the states and of not more than 20 members representing the union territories, who are chosen in such manner as parliament by law provides. The president may nominate two members to the House of the People to represent the Anglo-Indian community, if in his opinion it is not adequately represented. The term of the house is five years from the date of its first meeting.

The Council of States consists of not more than 250 members, of whom 12 are nominated by the president and the rest indirectly elected. The Council of States is not subject to dissolution; one-third of its members retire at the end of every second year. Elections to the Council are indirect, the allotted quota of representatives from

each state, as provided by the constitution, being elected by the elected legislators of that state through a system of proportional representation.

State government

In the states, the system of government closely resembles that of the union. The state executive consists of the governor—appointed by the president for a term of five years—and the council of ministers, with a chief minister at its head. Every state has a legislature, which normally consists of one house, known as the Legislative Assembly, except for Bihār, Jammu and Kashmir, Karnātaka, Mahārāshtra, Tamil Nādu, and Uttar Pradesh, which each have two houses, known as the Legislative Assembly and the Legislative Council. Each state has a high court at the head of its judicial administration.

Elections are held under the supervision of the chief election commissioner of the central government. Article 324 of the constitution empowers this official with the superintendence, direction and control of the preparation of the electoral rolls for, and the conduct of, "all elections to Parliament and to the Legislature of every State." The commissioner is an independent officer who cannot be removed from office "except in like manner and on the like grounds as a Judge of the Supreme Court."

Political parties. A political party that contests in more than one state and, in a general election, secures more than 4 percent of the votes cast in the country as a whole, is classified as a recognized multistate party for the next general election. Each recognized party is allotted a reserved (or specific) symbol, and the candidates it sponsors in different states use the symbol on behalf of the party. Among the major parties are (1) the Indian National Congress (I); (2) the Lok Dal; (3) the Janata Party; (4) the Communist Party of India (Marxist); (5) the Indian National Congress (S); (6) the Communist Party; (7) the Bhartiya Janata Party; and (8) the Rāštriya Sanjay Manch.

In addition to the recognized multistate parties, there were also recognized parties in different states. The more important among them are the All-India Dravida Munnetra Kazhagam in Tamil Nādu; the Shiromani Akāli Dal, a Sikh group in Punjab; the National Conference in Jammu and Kashmir; the Nāgā National Democratic Party in Nāgāland; the Sikkim Jamata Parishad in Sikkim; and the All-India Forward Bloc in West Bengal.

Justice. At the head of the judicial establishment—which consists of the federal high court, the state high court, and local and district courts—is the Supreme Court, which consists of the chief justice of India and not more than 17 other judges. Every state except Assam, Nāgāland, Punjab, and Haryāna has a high court; Assam and Nāgāland share one, as do Punjab and Haryāna. The union territory of Delhi has a high court of its own. The judges of the high court are appointed by the president after consultation with the chief justice of India, the governor of the state, and the chief justice of the state high court. Each judge of the Supreme Court is appointed by the president after consultation with the chief justice of the Supreme Court. A judge of the Supreme Court holds office until the age of 65 and is not liable to be removed from his office except by an order by the president issued after an address to each house of parliament that is supported by the majority of the total membership of that house and by not less than two-thirds of the members present and voting.

Armed forces. Supreme command of all armed forces is vested in the president. Responsibility for their administrative and operational control lies with the Ministry of Defence and with the army, navy, and air force headquarters. The main functions of the ministry are to ensure that the development and the activities of the three services are coordinated; that decisions on policy matters are obtained from the government, transmitted to the three service headquarters, and implemented; and that financial sanction for defense expenditure is obtained from parliament. Although the overall control of the three services is vested in the Ministry of Defence, each service functions under its respective chief of staff.

Role of the Ministry of Defense

In coordinating the activities of the three services, the Ministry of Defence has established a network of committees at different levels. At the highest level is the political affairs committee of the cabinet; all important policy matters relating to defense are decided by this committee, which is presided over by the prime minister.

EDUCATION

Education is primarily the responsibility of the state governments; the union government, for its part, is primarily concerned with the coordination of educational facilities, with the determination of standards for higher education, with research, and with scientific and technical education. Coordination in other sectors is the responsibility of a standing committee of the Central Advisory Board of Education. The union government also administers the Alīgarh Muslim University at Alīgarh; the Banaras Hindu University at Vārānasi; the Viśva-Bhārati University at Śantiniketan; the Jawaharlal Nehru University at Delhi; the University of Hyderābād at Hyderābād; and the North Eastern Hill University at Shillong. A Central Advisory Board of Education lays down general educational policy; it is assisted by four standing committees which deal with elementary, secondary, university, and social education.

About a third of India's population can read and write—about twice as many as in 1947, when the country attained independence. For nine out of ten children there is a primary school within walking distance (*i.e.*, less than a mile from home). Primary education up to the fifth standard (ages 6 to 11) is free throughout the country. Middle education up to the eighth standard (11 to 14) is free in almost all union territories and states. Secondary education (14 to 17) is also free in most of the states and union territories.

Apart from difficulties arising from the growing pressure of numbers, difficulty is encountered in imparting high standards of education in the various languages. Progress has, however, been encouraging in some spheres, notably in women's education, social education, adult education, and rural higher education. Some impressive advances have also been made in the fields of science and technology. For a country in which bullock carts outnumber automobiles by 12 to one, it is a notable achievement that India is able to build jet engines, construct a plutonium plant, design computers, develop a new rocket fuel, and manufacture transformers, steel-mill machinery, oil rigs, electronic equipment, and television receivers.

Since 1951, the five-year plans have been designed to reduce dependence on imported machinery and to build up India's strength in technological experience and research. The atomic research centre at Trombay, near Bombay, is producing enough radio isotopes and radioactive compounds not only to meet domestic needs but also to permit some exports to countries in Asia, Africa, and Europe. The plant genetics centre in New Delhi is evolving new varieties of seeds, while the hydrological research station in Pune has carried out studies on which the designs of most of India's dams have been based. Progress has also been achieved in research in tropical medicine and in the manufacture of drugs and pharmaceuticals. Some 600 scientific periodicals are published annually in India, and a considerable number of scientific and technical books are also published.

HEALTH AND WELFARE

Public health is primarily a state responsibility, although the union government plays a vital role in health programs and provides substantial financial grants to the union territories of Delhi, Andaman and the Nicobar islands, and Lakshadweep.

The union government has shown particular interest in sponsoring and supporting major national schemes for the prevention and control of disease. These include programs to combat malaria, tuberculosis, leprosy, venereal diseases, smallpox, trachoma (a contagious inflammation of parts of the eyes), and cancer. The union government is also concerned with schemes relating to sanitation, water supply, and nutrition.

Medical education is a state responsibility, but the Central Health Education Bureau also plays a coordinating role.

Family planning

Family planning, on the other hand, is centrally sponsored but locally implemented. The broad aim of the family-planning program is to reduce the birth rate and

to stabilize the population at a level consistent with the requirements of national economy. This involves planning on a vast scale, because of the millions of couples in the reproductive age group. The program is voluntary.

It was calculated that, in the late 20th century, a child born in India could expect to live to the age of 53, whereas in the mid-1950s life expectancy was about 32 years. As a result of the drive against preventable disease, some diseases have been successfully brought under control. Plague and smallpox, for instance, have been completely eradicated. Deaths from cholera have fallen radically; though the disease is still endemic in a few scattered areas, epidemics are rare. Malaria cases have also decreased.

The improvement of water supplies and of sanitation also form part of the public-health program. Thousands of children and nursing mothers are given free milk by citizens committees with the help of international organizations. Children also obtain free meals during school hours in a few states.

POLICE

The states control their own police forces through their home ministries. At the centre, the home minister of the union government exercises a coordinating role, controlling various all-India bodies, such as the Central Bureau of Intelligence, the Central Bureau of Investigation, the Central Detective Training School, the Central Forensic Science Laboratory, the Central Fingerprint Laboratory, and the Sardar Vallabhbhai Patel National Police Academy at Mount Abu, Rājasthān, where the Indian police service is trained. The Indian police service recruits university graduates by competitive examination and provides all the senior officers of the state police forces. The cities of Calcutta, Madras, Bombay, Delhi, Bangalore, Ahmedābād, Nāgpar and Hyderābād have separate police organizations.

CULTURAL LIFE

Indian art and culture has developed over many centuries and has many different manifestations, varying from folk art to modern painting and sculpture. Since independence, the union government and some state governments have attempted to promote artistic and cultural activities, as well as a heightened artistic consciousness, through various agencies. These include the Lalit Kala Akademi (National Academy of Fine Arts), the Sangeet Natak Akademi (Academy of Music, Dance, and Drama), and the Sahitya Akademi (National Academy of Letters). Some mass-communications facilities, such as the cinema, radio, and television, have also been utilized to make the Indian people more conscious of their cultural heritage. A number of institutions have collaborated in the task of popularizing traditional arts and culture.

The Lalit Kala Akademi, set up in 1954, promotes the development and understanding of the fine arts. It also coordinates the activities of the regional or state academies. The academy holds a national exhibition of art every year and also organizes exhibitions of arts and crafts of eastern and western countries in India, as well as of Indian arts and crafts abroad. In 1968 it began to hold an international exhibition of contemporary art once every three years. The academy also makes annual awards to outstanding artists who participate in the national exhibition; it is also engaged in making facsimiles of important frescoes in various parts of the country.

Drama, dance, and music are primarily encouraged through the medium of the Sangeet Natak Akademi, inaugurated in 1953; the academy promotes research, encourages the establishment of theatrical centres and training institutions, organizes seminars and festivals, awards prizes, and fosters cultural exchanges. It also maintains libraries of records, books, and films and maintains a museum of musical instruments, masks, and costumes, as well as a sound-recording studio. The academy also runs three training institutions—the National School of Drama and Asian Theatre Institute at New Delhi; Kathak Kendra (a school specializing in Indian dances) at New Delhi; and the Jawaharlal Nehru Manipuri Dance Academy at Imphāl.

The All-India Radio popularizes these activities and broadcasts programs of Indian drama and literature in Hindī and the regional languages. All-India Radio sponsors weekly and monthly national programs of music; it also sponsors an annual music festival to give young musicians and dramatists an opportunity to display their talents.

Indian literature is fostered by the Sahitya Akademi, inaugurated in 1954; it published (the *National Bibliography of Indian Literature*, 4 vol. (1901–53), including books published in the 15 major languages, as well as books in English published in India or written by Indian authors. In addition, it has translated a number of Indian and foreign classics into various Indian languages; it has also published 90 translations of works by the poet Rabindranath Tagore. The academy publishes three journals, a bi-monthly in English called *Indian Literature*, a quarterly in Hindī called *SanKāleen Bhāratiya Sāhitya*, and a semi-annual in Sanskrit called *Sanskrita Pratibha*.

The National Book Trust, established in 1957, is designed mainly to encourage literary production and to make moderately priced literary works available to libraries, educational institutions, and the general public. The trust also publishes standard works on education, science, culture and the humanities, and classical Indian literature, as well as translations of foreign classics in Indian languages. It also organizes an international book fair, held every two years, and national and regional book fairs, exhibitions, seminars, writer's camps, and workshops. The trust has organized exhibitions of Indian books at major international book fairs abroad.

The National Book Trust

To promote cultural relations with foreign countries the Indian Council for Cultural Relations was established in 1950 as an autonomous organization with the object of establishing, reviving, and strengthening cultural relations between India and other countries. The council, which has its headquarters in New Delhi, has regional offices in Bombay, Calcutta, Madras, Bangalore, Chandīgarh, and Vārānasi. It also has cultural centres in Suva (Fiji), Georgetown (Guyana), and Paramaribo (Suriname).

Press and broadcasting. During the era of British rule the press was subjected to fairly rigorous control. Upon independence the right to freedom of speech and expression was guaranteed by Article 19 of the constitution, which is accepted as including the freedom of the press. In general, the Indian press enjoys a remarkable degree of freedom, respected by both the central and state governments.

There are more than 18,000 Indian- or English-language periodicals—published as dailies, tri-weeklies, bi-weeklies, annuals, monthlies, and bi-monthlies. The largest number of newspapers are published in Hindī, followed by English, Bengali, and Urdū. The highly literate state of Kerala has more than 90 dailies in various languages, but a large number of these are no more than two- to four-page news sheets. A number of foreign-language newspapers are also published in India, although their circulation is small; these include papers published in Arabic, Burmese, Chinese, Nepālī, Portuguese, and Swahili.

Radio and television

In a country in which only about a third of the population is literate, the spoken word is more important than the written one. Because of this, audio-visual media, such as radio and television, reach out to a far bigger audience than does the press. Also in recent years a new type of consumer buying, the purchase of video cassette recorders and colour television sets, has been engulfing the country. Hundreds of video libraries and parlours have sprung up in the major cities. All-India Radio, in its medium-wave broadcasts, covers much of the country's population; its short-wave broadcasts cover practically the whole of the country.

Since radio and television are entirely operated under a government monopoly, controlled by the central government, they represent the most powerful media of publicity and mass communication in the country.

The cinema. India ranks with Japan and the United States among the largest producers of films in the world. In addition to the cinemas in the cities, movie theatres are found in all the towns, townships, and major villages.

The cinema, as well as radio and, to a smaller extent, television, has developed into a powerful medium of mass communication, and the government has appointed sev-

eral committees to investigate the workings of the film industry. Films in the Dravidian languages of the south predominate, accounting for well over half of the total output. The national language, Hindī, comes next, followed by Bengali and Marāṭhī.

The National Film Archive of India was established in 1964 in order to trace, acquire, and preserve the heritage of the national film industry. The archive is housed on the premises of the film institute in Pune. In addition, the National Film Development Corporation and the Directorate of Film Festivals promote Indian films that reflect official economic policies and political objectives. A children's film society promotes the production and distribution of children's films. The Film Institute of India, established by the Ministry of Information and Broadcasting, provides technical training in film production and facilities for research into film techniques; training is given in film direction, screen writing, motion-picture photography, sound recording, film editing, and acting.

Indian film theme

Social themes dominate Indian films, though their presentation may often seem crude by Western standards; mythology and legend as well as devotional subjects also play some part in Indian films. Apart from commercial productions, the Films Division of the Ministry of Information and Broadcasting produces many documentaries and newsreels, some of which are of a high quality and have won international awards. Documentaries and newsreels are produced in India's 15 regional languages; every cinema is required, under the terms of its license, to exhibit at each performance not less than 2,000 feet of films approved by the authorities that either are documentary films, have scientific and educational value, or deal with news and current events.

The government has established the Film Finance Corporation that grants loans to producers of films of good quality, giving preference to those projecting national problems and dealing with subjects of social and cultural value. In 1980 the Film Finance Corporation and the Indian Motion Pictures Export Corporation became part of the National Film Development Corporation. Established markets in foreign countries are mainly confined to the United States, East Africa, Mauritius, the West Indies, Sri Lanka, Singapore, Malaysia, Burma, Thailand, Fiji, Indonesia, the Persian Gulf, and West Africa. Soviet and eastern European countries have evinced interest in Indian films. The major demand abroad is for Hindī films, with Tamil and Bengali films next in popularity. The Indian filmmaker with the greatest following in the West, Satyajit Ray, works in Bengali.

Films are subject to censorship, and for this purpose a central board of film censors was constituted in 1951 to certify films for public exhibition. Members of the board, which is headquartered in Bombay, are government appointed. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.*

(F.R.M./Ed.)

HISTORY

The Indian subcontinent, the great landmass of South Asia, is the home of one of the world's oldest and most influential civilizations. Here, the subcontinent (which for historical purposes is usually called simply "India") is understood to comprise the areas of not only the present Republic of India but also of the republics of Pakistan and Bangladesh to the year of independence, 1947. (For the histories of these latter two nations since independence, see PAKISTAN and BANGLADESH.)

The earliest periods (before 1750 BC) of the history of India are known only through historical reconstructions from archaeological evidence. These early phases are conveniently divided into two periods: (1) the prehistoric period (before 2300 BC) and (2) the period of the Indus civilization (2300–1750 BC). The Indus civilization, an early urban culture in the Indus River Basin, is also known as Harappan culture or civilization after the name of one of the most important archaeological sites (Harappā) of the period. Consequently, the prehistoric period is often termed pre-Harappan to indicate that it predates the development of urban civilization in the Indus Valley.

Since early times the Indian subcontinent appears to have provided an attractive habitat for man. Toward the south it is effectively sheltered by wide expanses of ocean, which tended in ancient times to isolate it culturally, while to the north it is protected by the massive Himalayan ranges, which also sheltered it from the arctic winds and the air currents of Central Asia. Only in the northwest and northeast is there an easier access by land, and through these two sectors most of the early contacts with the outside world took place. Within the framework of hills and mountains represented by the Indo-Iranian borderlands on the west, the Indo-Burmese borderlands in the east, and the Himalayas to the north, the subcontinent may in broadest terms be divided into two major divisions: in the north are the basins of the Indus and Ganges rivers, and to the south is the block of Archean rocks that forms the Indian Peninsula. The expansive alluvial plains of the river basins provided the environment and focus for the rise of two great phases of city life: at the end of the 3rd millennium BC the civilization of the Indus (Sindhu) Valley and during the first millennium BC that of the Ganges (Gaṅgā). To the south of this zone, and separating it from the peninsula proper, is a belt of hills and forests, running generally from west to east and to this day largely inhabited by tribal people. This belt has played mainly a negative role throughout Indian history, but it is traversed by various routes linking the more attractive areas north and south of it. Toward the west through this belt flows the Narmada (Narbada) River, which has long been regarded as the symbolic boundary of North and South India.

Moving from west to east, the northern parts of India represent a series of contrasting regions, each with its own distinctive cultural history and with its own distinctive modern population. In the west, the valleys of the Baluchistān uplands are a low-rainfall area, producing mainly wheat and barley and to this day having a low density of population. These mainly tribal people are in many respects closely akin to their Iranian neighbours. The adjacent Indus plains are also an area of extremely low rainfall, but the annual flooding of the river in ancient times, and the exploitation of its waters by canal irrigation in modern times, have given greater agricultural productivity, and the population is correspondingly denser than that of Baluchistan. The Indus Valley may be divided into three parts: In the north are the plains of the five tributary rivers of the Punjab (Panjāb). In the centre the consolidated waters of the Indus and its tributaries flow through the alluvial plains of Sind, and in the south pass naturally into the Indus Delta. East of the latter is the Thar, or Great Indian Desert, which is in turn bounded on the east by the Arāvalli Hills, the northernmost extent of the peninsular block. Beyond them is the hilly region of Rājasthān and the Mālwa Plateau. To the south is the Kāthiāwār Peninsula, forming both geographically and culturally an extension of Rājasthān. All these regions have a relatively denser population than the preceding group, but for topographical reasons they have tended to be somewhat isolated, at least during historical times.

Geography of the Indus Valley

East of the Punjab and Rājasthān, North India develops in a series of belts running broadly east to west and following the line of the foothills of the Himalayan ranges in the north. The southern belt consists of a hilly, forested area broken by the numerous escarpments of the edge of the peninsular block and comprising the Vindhya Range, which includes the Bhāndrer, Rewa, and Kaimur plateaus. Between the hills of central India and the Himalayas lies the Ganges Valley proper, constituting an area of high-density population, moderate rainfall, and high agricultural productivity. Archaeology suggests that, from the beginning of the 1st millennium BC, rice cultivation has played a large part in supporting this population. The

Gangetic Plains divide into three major parts: to the west is the *doab,* or "mesopotamia," of the Yamuna (Jumna) and Ganges rivers; east of the confluence of these rivers lies the Middle Ganges Valley, in which population tends to increase and cultivation of rice predominates; to the southeast lies the extensive delta of the combined Ganges and Brahmaputra rivers. The Brahmaputra flows from the northeast, rising from the Tibetan Himalayas and emerging from the mountains into the Assam Valley, being bounded on the east by the Pātkai Bum and Nāga Hills and on the south by the Mikīr, Khāsi–Jaintia, and Gāro hills. There is plenty of evidence that influences reached India from the northeast in ancient times, even if they are less prominent than those that arrived from the west.

The materials available for a reconstruction of the history of India prior to the 3rd century BC are almost entirely the products of archaeological research. Traditional and textual sources, transmitted orally for many centuries, are available from the closing centuries of the 2nd millennium BC; but their use must depend largely upon the extent to which any passage can be dated or associated with archaeological evidence. For the rise of civilization in the Indus Valley and for contemporary events in other parts of the subcontinent, the evidence of archaeology is still the principal source of information. Even when it becomes possible to read the short inscriptions of the Harappan seals, it is unlikely that they will provide much information to supplement other sources. In these circumstances it is necessary to approach the early history of India largely through the eyes of the archaeologists, and it will be wise to retain a balance between an objective assessment of archaeological data and its synthetic interpretation.

Importance of archaeological evidence

In the following pages certain terms have been used that have not always gained universal currency even among archaeologists writing about India. The Stone Age will be treated in terms of Early, Middle, and Late, because since 1961 there has been fairly general acceptance of these terms to denote the three major stages of the Indian Stone Age. The first and second appear to correspond in general typological terms with the Early and Middle Palaeolithic stages in Europe, while the third might be said to correspond with the Mesolithic. The terms Neolithic and Chalcolithic have been used very sparingly, preferring in general to name cultures broadly in terms of their methods of subsistence (when known). In India-Pakistan the two terms are used in a broadly overlapping sense. The terms pre-Harappan and Harappan are used primarily in a chronological way, but also loosely in a cultural sense, relating respectively to periods or cultures that precede the Indus civilization, and thus the appearance of city life in the Indus Valley, and to the Indus, or Harappan, civilization itself.

India from the Late Stone Age to the decline of the Indus civilization (1750 BC)

THE EARLY PREHISTORIC PERIOD

Late Stone Age hunters. The oldest traces of human activity discovered so far in India–Pakistan belong to the Late Stone Age. Evidence then begins to have a direct bearing upon the later history of the subcontinent, since there are clear indications that groups of Late Stone Age hunters long continued to flourish and had contact with later settled communities of agriculturalists. Late Stone Age sites, identified by assemblages of microliths, have been discovered in large numbers in many parts of the subcontinent, from Baluchistan in the west to Bengal in the east, and from the North-West Frontier Province in the north to Sri Lanka in the south. As the vast majority of the finds were from surface sites, they long remained without precise dates or cultural contexts. Only in recent years have a number of cave and dune sites been excavated from which radiocarbon samples and animal remains have been obtained. These excavations are as yet incompletely published, but they have already added a new dimension to the scholarly view of the origins of Indian civilization. Archaeological sites at Maraiyur in the eastern foothills of Kerala, at Lānghnaj in Gujarāt, at Bāgor near Udaipur in Rājasthān, and at Ādamgarh, near Moshangābād in Madhya Pradesh, are particularly notable. The site at Maraiyur has revealed paintings, drawings, and engravings that cover an interior of a solitary rock wall 39 feet (12 metres) high and 31 feet wide with dates ranging from the 8th to the 2nd millennium BC. Bāgor has produced evidence of long occupation of a sand dune, with radiocarbon dates ranging between the 6th and 3rd millennia BC, while Ādamgarh produced a date from the 6th millennium. From later phases at Bāgor and Lānghnaj copper objects were discovered as well as pottery and, in the latest level, objects of iron. The animal remains, as far as they have been analyzed, suggest that these communities subsisted mainly by hunting and gathering, but remains of domesticated species, including cattle, buffalo, sheep, and goats, are also present, in the later stages. The equipment of these people included the bow and arrow, and their arrows and harpoons were barbed with composite points using stone microliths. Burials have been excavated at a number of Late Stone Age sites, but only those from Lānghnaj have been studied and published. The skeletons were considerably deformed, and it is difficult to draw conclusions regarding their ethnic type, but it has been stated that they indicate traits characteristic not only of Mediterranoids and Veddoids but also of other racial groups.

Important sites

Thus, it is now evident that, for an as yet indefinite period before agricultural and pastoral communities appeared in India–Pakistan, there were communities of stone-using and hunting and gathering—and perhaps fishing—peoples spread widely throughout the country; many of these communities continued to exist as recognizably separate entities for long after the appearance of the first agriculturalists, or users of copper, bronze, and even iron. The relations of these groups of hunter–gatherers and agriculturalists thus present a most interesting field of research.

First settlements in Baluchistan. The Indo-Iranian borderlands form the eastern extension of the Iranian Plateau and, from many points of view, mirror the environment of the Fertile Crescent (an arc of agricultural lands extending from the Tigris-Euphrates to the Nile Basin) in the Middle East. Across the plateau, lines of communication existed from very early times, and throughout the whole area it is to be expected that settled life with agriculture and domestication of animals would have spread, once developed, without undue delay. Yet the eastern portions are still less well-known than their western counterparts, and to date the earliest radiocarbon date for a settlement in Baluchistan is not much older than 3500 BC. It is evident, however, that settlements of some kind had existed for a considerable period prior to that date.

Stages of development before the appearance of urban life

The stages of development leading to the appearance of city life in the Indus Valley may be classified as three. The first stage, at present known from southeastern Afghanistan (Mundigak), from the Quetta Valley (Kili Ghul Mohammad), Kalāt (Anjira), and from northern Baluchistan (Rana Ghundai), appears to have consisted of seminomadic pastoralism with limited cultivation. All of these sites lie at relatively high levels in the valleys of Baluchistan or in the more open valleys of the Helmand River in southern Afghanistan. At some sites there are no traces of permanent habitations, while at others the first houses of mud brick or with pressed-earth walls occur. Stone blades and grinding stones are found, along with bone points and crude pottery, some made by pressing wet clay into baskets. The indications are that metal in the form of copper was present from the first at Mundigak, but at other sites it is not attested in this period. The people kept cattle and apparently sheep and goats (materials for study are still very limited), and cultivated club wheat and perhaps barley. The picture is of semipermanent camp sites at which a greater or lesser part of the year was spent, probably much like those of the seminomadic tribes of northwest Pakistan down to the present day. The first stage appears to have ended *c.* 3500 BC.

After this in Baluchistan there followed the second stage, a period of consolidation, with the development of cultivation and of pastoralism. This second stage corresponds with a time when the settlements grew in size and the houses became larger, probably indicating an increasing density of population. Copper tools became more com-

mon, and the traditions of fine painted pottery for which Baluchistan is renowned came into their own. This period also saw a marked regional diversification in the styles of pottery painting, though what, if any, wider cultural implications this may have is not yet known. One scholar suggests that this stage continued until around 2500 BC when, according to him, it passed into a third stage—one of fully developed sedentary village life, with regionalization but with growing evidence of interregional contacts. This latter third stage coincides with the transition to urbanization in the Indus Valley, or at Mundigak.

The climate and geography of the valleys of Baluchistan make large-scale cultivation impossible, and hence it is probable that pressure was felt to expand to adjacent regions as the density of population increased. It appears that a movement of this sort eastward toward the Indus plains took place around the end of the 4th millennium. This movement is as yet not well documented, but the evidence is sufficient to testify that in the south, on the northern flank of the Indus Delta and in the hills adjoining the plains, a group of sites were settled at this time. Of these Amri situated 300 miles north of Karāchi, has been recently excavated by J.M. Casal, and it has been shown that during the first period a long development took place, with features of culture recalling those of Baluchistan. Several of the sites of this period appear to have been surrounded by massive defensive walls, in some cases perhaps as defense from floods but in others clearly as defense against men or animals. Two radiocarbon dates and other indications suggest that the pre-Harappan occupation at Amri lasted from *c.* 3000 to *c.* 2300 BC.

Early settlements in the Indus Valley. It must have been at some such site as Amri that the enormous agricultural potentials of the Indus floodplains were first appreciated. Rejuvenated by an annual deposit of silt, these plains offer the prospect of food production employing a minimum of skill, labour, and implements. It would seem that the pattern was repeated over a wide area in the succeeding centuries. At Kot Diji to the north of Amri a similar settlement was established around 2600 BC, situated on a small outcrop of rocks near one of the flood channels east of the Indus. It is probable that Mohenjodaro, some 30 miles away across the river, was also the site of a similar settlement; and it is now known that there was another situated beneath the later city of Harappā some 300 miles to the northeast in the Punjab. Another pre-Harappan settlement, dating from around 2450 BC, was established at Kāli Banga on the banks of the ancient Saraswatī (now Ghaggar) River in northern Rājasthān, over 100 miles southeast of Harappā. This site recently has been carefully excavated by the Archaeological Survey of India. Here, too, the mud-brick houses were contained within a broad surrounding wall of brick, but the houses lack the regular alignment of those of the later Indus civilization, and the bricks are of a different size. Perhaps the most exciting find of this period at Kāli Banga is part of the plowed surface of a field, with furrows running in two directions, just as in the modern plowing of the region. Still, it is too soon to piece together a coherent account of the life of these people. Objects of copper are present from the beginning at Amri, but they are reportedly very rare at Kot Diji and rare also at Kāli Banga. A stoneblade industry was a regular feature, and at Kot Diji leaf-shaped arrowheads of stone, recalling those of contemporary Baluchistan, are reported. Other objects include querns (hand mills for grinding grain), steatite beads and shell bangles, and (from Kot Diji) a fragment of a terra-cotta bull anticipating a Harappan type. Some of the painted pottery recalls that of northern Baluchistan; some at Amri has a tendency to geometric patterns, again recalling Baluchistan; while the pottery from Kāli Banga in particular has a strongly individual character that distinguishes it from the rest and that appears to represent the beginnings of a process of Indianization. In spite of these differences there are many common traits—some direct anticipations of the mature Harappan, or urban, culture—at all the sites, and they give the impression of an overriding uniformity throughout the considerable area they cover.

Kāli Banga

Some idea of the extent of this pre-Harappan phase of settlement can be gauged when recently reported discoveries from the vicinity of Dera Ismāīl Khān and Taxila (Sarai Khola) are included. These indicate that there were sites almost from the mouth of the Indus to the foothills of the Himalayas and from west of the Indus to well beyond the modern borders of India. Indeed, by this time such a distribution begins to approximate that of the Harappan civilization itself. Thus, this pre-Harappan phase in the Indus Valley and the Punjab must now be discussed as much because of its extent as because of the suggestions of cultural homogeneity, and because it can be seen to be antecedent to the urban stage that succeeds it in all these areas.

First settlements east of the Indus system. It is still not at all clear at what date settlements of agriculturalists or pastoralists first appeared east of the Indus system, but available evidence suggests that this development took place in pre-Harappan or pre-urban times. Nor is it clear whether the highly individual character of the remains indicates movements from the west, via the Iranian Plateau; movements from the east, from the area of South China or Southeast Asia; or a primarily local and spontaneous development. Probably elements of all three of these possibilities were factors. As indicated above, there were groups of Late Stone Age hunters spread throughout India-Pakistan, and some evidence points to their having made a major contribution to the cultures that developed in peninsular India; other traits strongly suggest the movement of peoples from the direction of Iran; and still others appear to favour the arrival of influences from the east. A further complicating factor is that in not all of these regions is there definite evidence to associate the "Neolithic" ground or polished stone axes found there with actual settlements from which the date and other aspects of the culture (including details of economy and subsistence) may be gauged. Yet when evidence has become available, it has in each case tended to support the hypothesis that this development began in pre-Harappan times, even if it continued in some instances long after the Indus civilization had disappeared. The coexistence of groups of different cultural levels in neighbouring areas has always been a special feature of India.

Stone axes

Four main areas have attracted attention through the discovery of very large numbers of stone axes. The first is the Vale of Kashmir, where at Burzahom, near Srīnagar, a settlement has been excavated revealing deep pits, apparently dug as dwellings for protection from the cold, and a burnished, gray pottery. Radiocarbon samples indicate that this settlement was already in existence by *c.* 2375 BC and was continued down to the iron-using period (that is, the opening of the 1st millennium BC). A series of bone and antler harpoons and needles was also discovered. The ground-stone axes are of an unspecialized type and do not throw much light upon the cultural affinities of the group. There is as yet no clear evidence of what (if any) grains were cultivated or of what animals were domesticated, but evidently the hunt played a large part in the economy. It is strange that there is so far no direct evidence of cultural contact with the Indus civilization, even though the Kashmir Neolithic culture flourished at the same time. There are also a number of curious traits, including some stone tools and the custom of burying dogs or animals of the hunt in ritual fashion, that appear to indicate contacts with regions far to the north of the Himalayas rather than with any other parts of India or Pakistan.

The second main area in which stone axes have been discovered is located in Karnātaka and in Andhra Pradesh. In the northern parts of Karnātaka the nucleus from which stone-ax-using pastoralists spread to many parts of the southern peninsula has been located. The earliest radiocarbon dates obtained in this area are from ash mounds formed by the burning on these sites of great masses of cow dung inside cattle pens. These indicate that the first settlers were semi-nomadic and that they had large herds of zebu cattle. The earliest known settlements date to *c.* 2400 BC. Other important sites are Brahmagiri and Tekkaḷkotā in Karnātaka, and Utnūi and Nāgārajunikonda in Andrhra Pradesh. At Tekkaḷkotā three gold ornaments were excavated, indicating exploitation of local gold deposits,

but no other metal objects have been found, suggesting a relative scarcity of metals. These early sites produced distinctive burnished gray pottery, smaller quantities of black-on-red painted pottery, stone axes, and bone points, and in some instances evidence of a stone-blade industry. The axes have a generally oval section and triangular form with pointed butts. Cattle bones are in the majority; those of sheep or goats are less common but are also certainly present. Other settlements have been excavated in recent years in this region, but so far they have produced only dates from the 2nd millennium, suggesting that the culture continued with little change for many centuries.

Stone axes of a generally similar form have been found widely throughout the southern peninsula and may be taken as indications of the spread of pastoralists throughout the region during the 2nd millennium BC.

A third concentration of ground-stone tools has been found in the hills of Assam. These have as yet barely been located in actual settlements, but in the few excavations made so far they are found to be associated with a crude cord-impressed pottery, suggestive of the fabric of Chinese Neolithic wares. A number of the axes have been ground to a rectangular section, again suggestive of Chinese forms. This culture is still undated, however, and little more can be said of its age or affinities.

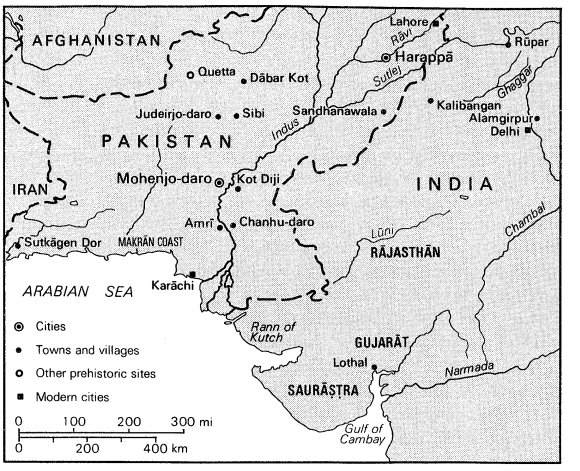
Finally, in a fourth area—the hills and forests of central India—ground-stone axes are found to be widely distributed. In these areas the major finds are from the surface and no early sites have been excavated. These finds must be treated with caution, since stone axes occur throughout the area down to the 1st millennium BC.

These developments have been discussed at some length because, problematic though much of the data may be, they may be inferred to provide (at least in some regions) a major element of the population and culture of succeeding periods. Also, if the relations between the Indus cities and adjacent regions are to be understood, it is necessary to know something of the cultural levels of the peoples they encountered. The technical equipment of these "Neolithic" cultures may have had certain common features, but both the technical equipment and the economies show profound differences. Thus, there is no evidence of cattle or domestication in Kashmir, while cattle seem to have been the mainstay in Karnātaka. A microlithic blade industry developed as one element in the south, while in Kashmir there was none. The burnished gray pottery of these two areas stands in complete contrast to the cord-impressed pottery of Assam and other localities. Much more work needs to be done before the questions raised can be dealt with and the origins, relationships, and later histories of these Neolithic cultures properly understood.

THE INDUS CIVILIZATION

Origins. The origins of the civilization are not easy to define in detail. The idea of civilization, *i.e.*, of a considerable literate population living in towns and cities, had been matured in Mesopotamia before the end of the 4th millenium and there can be little doubt that the Mesopotamian pattern was in principle present in the thinking of the founders of the Indus civilization. But the differences between the two civilizations are sufficient to preclude any direct colonization of the Indus from Mesopotamia. It has been seen above how the oldest settlements of agriculturalists in India-Pakistan appear to have been in the upland valleys of Baluchistan. Throughout this area settlements occur during the 4th millennium, perhaps earlier. By the beginning of the 3rd millennium there is evidence of a fairly dramatic extension of settlements onto the floodplains of the Indus, where there were almost unlimited possibilities for agriculture. This appears to have developed first in the south, in the vicinity of the modern Indus Delta region (which was considerably smaller at that time than it is today), and to have spread north and east during succeeding centuries, reaching Kāli Banga only around the middle of the millennium. The size of Kāli Banga, the only settlement for which there is any reasonable estimate, was considerable—roughly 240 by 180 metres (800 by 600 feet). It may be inferred that by the middle of the 3rd millennium the population was

Agriculture and the rise of cities



Principal sites of the Indus civilization.

Adapted from Sir Mortimer Wheeler, *Civilizations of the Indus Valley and Beyond* (1966), Thames and Hudson, Ltd., London, and McGraw-Hill Book Company, New York; original map by Shalom Schotten

expanding rapidly, and the number of settlements and probably the area settled grew in the same manner.

It is striking that when Sir John Marshall and his colleagues completed the first round of excavations at Mohenjo-daro, Harappā, and Chanhu-daro (before World War II), scarcely any traces of this pre-Harappan stage had been uncovered or recognized, and hence even Stuart Piggott, among the most conservative scholars of his time, could still write in 1948 of the Indus civilization as having "no known beginnings, no tentative early phases." Hence some writers also tried to explain its origins in terms of immigration from outside. These views must be reconsidered in the light of more recent discoveries, which reveal a widespread local culture as the direct antecedent, even if there are also indications that some sort of external influence contributed to the rise of the Indus cities. Suffice it to say that somewhere around or after 2300 BC a combination of causes led to the fairly profound changes that mark the end of the pre-Harappan stage at all the sites so far known and the appearance of the mature Harappan or urban phase. At two sites, there is evidence that the transformation may have been a violent one. At Kot Diji and at Amri thick layers of ash suggestive of a major conflagration were revealed. At other sites, particularly Kāli Banga, the evidence is less clear and the transformation may well have been peaceful. This need not imply that the Harappan culture did not owe an important debt to stimuli, in the form of trade arriving from Mesopotamia, where the empire of Sargon of Akkad (2334–2279 BC) heralded a period of unprecedented foreign contacts; nor that the particular events that initiated the urbanization may not have coincided with some sort of incursions of people, perhaps from southern Baluchistan or beyond, where the Kulli culture had flourished for some time, or even simultaneously from northern Baluchistan. Rather it may indicate that the economic potentials and expanding population on which the civilization were based were already present within the Indus valley itself.

Nature and significance. The Indus civilization marks the culmination of a period of rapid development. Seen in this light it appears as an indigenous response to the challenge of the environment. The direct contact with Mesopotamia must have provided a strong incentive to this development, but the process itself must be seen as indigenous. Nearly all the earlier writers have sensed the Indianness of the civilization, even when they were largely unable to articulate it. Thus, V. Gordon Childe wrote that:

The unique character of the new civilization

> India confronts Egypt and Babylonia by the 3rd millennium with a thoroughly individual and independent civilization of her own, technically the peer of the rest. And plainly it is deeply rooted in Indian soil. The Indus civilization represents a very perfect adjustment of human life to a specific environment. And it has endured; it is already specifically Indian and forms the basis of modern Indian culture.

The force of Childe's words can be appreciated even without an examination of the Indus Valley script found on seals; the attention paid to domestic bathrooms, the drains, and the Great Bath at Mohenjo-daro can all be compared with elements in the later Indian civilization. The bullock carts with a framed canopy, called *ikkās*, and boats are little changed to this day. The absence of pins, the love of bangles and of elaborate nose ornaments are all peculiarly Indian. The religion of the Indus also is replete with suggestions of traits known from later India. The significance of the bull, the tiger, and the elephant; the composite animals; the seated yogi god of the seals; the tree spirits and the objects resembling the Śiva *liṅga* (a device, symbolic of the god Śiva, that has a phallic shape) of later times—all these are suggestive of enduring forms in later Indian civilization.

It is still impossible to do more than guess at the social organization or the political and administrative control implied by this vast area of cultural uniformity. The evidence of widespread trade in many commodities, the apparent uniformity of weights and measures, the common script, the uniformity—almost common currency—of the seals, all indicate some measure of political and economic control and point to the great cities Mohenjo-daro and Harappā as their centres. The presence of the great granaries on the citadel mounds in these cities and of the "citadels" themselves suggests, partly on the analogies of the cities of Mesopotamia, the existence of priest-kings, or at least of a priestly oligarchy, that controlled the economy and civil government. The intellectual mechanism of this government and the striking degree of control implicit in it are still matters of speculation. Nor can scholars yet speak with any certainty regarding relations between the cities and surrounding villages. Much more research needs to be done, on many such topics, before the full character of the Indus civilization can be revealed.

Chronology. Until radiocarbon dating became available, the chronology of these events, and indeed of the whole life of the Indus civilization, was dependent upon often tenuous cross-dating with Mesopotamia for objects of Indian origin discovered in the excavations there. These indicated a central period of 2300 to 2000 BC, with probable extensions at either end, and led the British archaeologist Sir Mortimer Wheeler to propose the generally accepted overall span of 2500 to 1700 BC. Radiocarbon dating has, in general, tended to confirm the correctness of this view. On the basis of some 30 samples, mainly from Kāli Banga, it would appear that at both that site and Kot Diji the beginning of the urban phase—that is to say, the mature Harappan—was as late as 2150 to 2100 BC. Even if, as appears probable, it was somewhat earlier at Mohenjo-daro (which may be regarded as the epicentre of the whole civilization), it is not likely that the crucial transformation took place much before 2300 BC. The end of the civilization is even more unclear, but it appears likely that the mature Harappan culture came to a close around 1750 BC, to be succeeded at Harappā, Chanhu-daro, and other sites by late or post-Harappan phases. Thus, the present indications are that the mature Harappan civilization had a time span of not more than five centuries, rather than the 1,000 years it was formerly allotted.

Extent. All the earlier writers have stressed the remarkable uniformity of the products of the Harappan civilization, and for this reason they provide a definite hallmark for its settlements. The present evidence suggests that, if the outermost sites are joined by lines, the area enclosed will be a little less than 500,000 square miles—considerably larger than modern Pakistan—and if, as is generally inferred, this cultural uniformity coincided with some sort of political and administrative unity, the size of the resulting "empire" is truly vast. Within this area over 70 sites have been identified, the great majority of which are on the plains of the Indus or its tributaries or on the now dry course of the ancient Saraswatī River, which flowed south of the Sutlej and then southward to the Indian Ocean, east of the main course of the Indus itself. Outside the Indus system a few sites occur on the Makrān Coast, the westernmost of which is at Sutkā-gen Dor, near the modern frontier with Iran. These sites were probably ports or trading posts, supporting the sea trade with the Persian Gulf, and were established in what otherwise remained a largely separate cultural region. The uplands of Baluchistan, while showing clear evidence of trade and contact with the Indus civilization, appear to have remained outside the direct Harappan rule.

To the east of the Indus Delta other coastal sites are found beyond the marshes of Kutch (Kacch) and in the interior of the Kāthiāwār Peninsula (Saurāṣṭra). The most important thus far is the estuarine trading post at Lothal on the Gulf of Cambay. Other coastal sites have been claimed on the mainland in Gujarāt, but it is doubtful if any were actually established during Harappan times. West of the Indus River a number of important sites are situated on the alluvial Kacchi plain toward Sibi and Quetta. This area may have been one of the routes by which cultural influence from Baluchistan reached the Indus Valley. East of the Indus system, toward the north, a number of sites occur right up to the edge of the Himalayan foothills, where at Ālamgīrpur, east of Delhi, the easternmost Harappan (or perhaps, more properly, Late Harappan) settlement has been discovered and partly excavated. If the area covered by these sites is compared with that of the pre-Harappan settlements, it will be seen that there is an expansion in several directions, along the coast to both the west and east, and eastward through the Punjab toward the Ganges-Yamuna *doab*.

Planning and architecture. The Harappan sites range from extensive cities to small villages or outposts. The two largest are Mohenjo-daro and Harappā, each perhaps originally about a mile square in overall dimensions. Each shares a characteristic layout, oriented roughly north–south with a great fortified "citadel" mound to the west and a larger "lower city" to the east. A similar layout is also discernible in the somewhat smaller town of Kāli Banga (originally perhaps about 1,300 feet [400 metres] square), and several other major settlements appear to have shared this scheme. Other large sites are Dābar Kot in the Loralai Valley of northern Baluchistan; Amri, Chanhu-daro, and Judeirjo-daro in Sind; Sandhanawala in Bahāwalpur; and Rūpar (Ropar) in the Punjab. Among the smaller sites, special interest attaches to Lothal, where a number of unique and problematic features were discovered in excavations. Of all the sites, Harappā, Mohenjo-daro, Kāli Banga, and Lothal have been most extensively excavated, and more can be said of their original layout and planning. Thus, they are considered in greater detail below.

Layout of cities

The layout of the cities commands respect. At the three major sites, the citadel mound is on a north–south axis and about twice as long as it is broad. The lower city is laid out in a grid pattern of streets; at Kāli Banga these were of regularly controlled widths, the major streets running through, while the minor lanes were sometimes offset, creating different sizes of blocks. At all three sites the citadel was protected by a massive, defensive wall of brick, which at Kāli Banga was strengthened at intervals by square or rectangular bastions. At Kāli Banga traces of a somewhat less substantial wall around the lower town have also been discovered. In all three cases the city was situated close by a river, although in modern times the rivers have deserted their former courses.

The most common building material at every site was brick, but the proportions of burnt brick to unburnt mud brick vary. Mohenjo-daro employs burnt brick, perhaps because timber was more readily available, while mud brick was reserved for fillings and mass work. Kāli Banga, on the other hand, reserved burnt brick for bathrooms, wells, and drains. Most of the domestic architecture at Kāli Banga was in mud brick. Brick was generally bonded in courses of alternate headers and stretchers, the so-called English bond. Stone was rarely, if ever, employed structurally. Timber was occasionally used as a lacing for brickwork, particularly in large-scale work such as the defenses or the granary at Mohenjo-daro. The common bricks were made in an open mold, but for special purposes sawn bricks were also employed. Timber was used for the universal flat roofs, and in some instances the sockets indicate square cut beams with spans of as much as 14 feet.

The houses were invariably entered from the side lanes,

Shops, workshops, and houses

the walls to the main streets presenting a blank brick facade, broken only by the drainage chutes. Apart from domestic structures, a wide range of shops and craft workshops have been encountered, including potters' kilns, dyers' vats, and the shops of metalworkers, shellworkers, and bead makers. There is surprisingly little evidence of public places of worship, although at Mohenjo-daro a number of possible temples were unearthed in the lower city, and other buildings of a ritual character were reported in the citadel. The size of houses varies very considerably. At the one extreme are single-roomed barracks, with cooking and bathing areas formed within by partition walls, and at the other are large houses around a central courtyard or sometimes with a set of intersecting courtyards, each with its own adjoining rooms. Nearly all the larger houses had private wells. In many cases brick stairways led to what must have been upper stories or flat roofs. Almost every house had at least one bathroom, and there is also evidence of public bathrooms or privies. The wastes from these bathrooms were carried by a drainage channel to a chute built in the thickness of the wall, in turn emptying into the main street drains. A number of pottery drainpipes also have been discovered. The bathrooms were usually indicated by the fine quality of the brickwork in the floor and by these drains. Some of the smaller examples were probably used as privies. The flooring of common rooms was of either beaten earth or brick, and for certain special purposes floors were made of sawn brick.

Important sites. *Mohenjo-daro.* The mounds of Mohenjo-daro lie near the right bank of the Indus in the Lārkāna district of Sind. The excavations revealed that occupation continued to a depth of about 30 feet below the modern level of the plain, because of the annual deposition of alluvial silt during the floods. The lowest levels are thus below the modern water table and are still largely unexcavated. The main features of the layout of Mohenjo-daro, with a citadel to the west and a lower city and grid of streets to the east, have already been alluded to above. Enough has been said of the general features of the lower city to make it unnecessary to say more of the considerable areas excavated in that part. The citadel, however, demands further attention. In the citadel Sir John Marshall discovered a massive platform of mud brick and clay approximately 20 feet in depth, above which were six main building levels. Under this platform lay the remains of the early period. The upper strata he designated as the intermediate and late periods. It appears that the mature Indus culture embraces all three of these periods and thus that any pre-Harappan remains must lie still lower. It is probable, but by no means certain, that the platform was raised as protection against floods. Both it and the great brick defensive wall around the perimeter were built at the beginning of the intermediate period.

The Great Bath

The main buildings of the citadel apparently belong to the same period. The most striking of these is the Great Bath, which occupies a central position in the better preserved northern half of the citadel. It is built of fine brickwork, measures 897 square feet (83 square metres), and is eight feet lower than the surrounding pavement. The floor of the bath consists of two skins of sawn brick set on edge in gypsum mortar, with a layer of bitumen sealer sandwiched between the skins. Water was evidently supplied by a large well in an adjacent room, and an outlet in one corner of the bath led to a high corbelled drain disgorging on the west side of the mound. The bath was reached by flights of steps at either end, originally finished with timbered treads set in bitumen. The significance of this extraordinary structure can only be guessed at, but it has generally been thought that it is linked with some sort of ritual bathing. To the north and east of the bath were groups of rooms that evidently were also designed for some special function, probably associated with the group of administrators or priests who controlled not only the city but the great state which it dominated. To the west of the bath a complex of brick platforms about five feet high and separated from each other by narrow passages formed a podium of some 150 by 75 feet, which has been identified by Sir Mortimer Wheeler as the base of a great granary similar to that known at Harappā. Below the granary were brick loading bays. In the southern part of the mound an oblong "assembly hall" was discovered, having four rows of fine brick plinths, presumably to take wooden columns. In a room adjacent to this hall a stone sculpture of a seated male figure was discovered and nearby a number of large worked-stone rings, possibly of some architectural significance. In view of the scarcity of all kinds of stone or sculpture at Mohenjo-daro, it seems certain that this area was invested with some special significance and may well have been a temple or connected with some religious cult.

Harappā. The vast mounds at Harappā stand on the left bank of the now dry course of the Rāvi River in the Punjab. The ruins attracted the attention of travellers early in the 19th century and were visited by General Cunningham in 1853 and 1873. Between these years they provided a source of ballast for local railway construction, and irreparable damage appears to have been done. They were excavated between 1920 and 1934 by the Archaeological Survey of India, in 1946 by Sir Mortimer Wheeler, and more recently by the Pakistan Archaeological Department. When first discovered, the extensive surviving brick ramparts led to the site's being described as a ruined brick castle. The lower city is partly occupied by a modern village, and it has been seriously disturbed by erosion and brick robbers. The citadel, to the west, is roughly a parallelogram on plan, measuring approximately 1,300 by 650 feet (400 by 200 metres). Excavation here revealed a great platform of mud brick about 20 feet in thickness, with a massive brick wall around the perimeter. Below the defenses were discovered traces of the pre-Harappan period, and in the uppermost levels an intrusive pottery named after its presence in the nearby site known as cemetery H was found alongside the normal Harappan pottery. The excavations were not extensive enough to reveal the layout of the interior, but about six building periods were discovered above the platform. The most interesting remains were discovered immediately north of the citadel, close to the bed of the river. Here there were a series of circular platforms evidently intended to hold mortars for pounding grain; a remarkable series of brick plinths, which are inferred to have formed the podium for two rows of six granary buildings, each of 50 by 20 feet and of a different design from those at Mohenjo-daro; a series of pear-shaped furnaces, apparently used for metallurgy; and two rows of single-roomed barracks, which have generally been assigned to a coolie or "slave" community. Two other discoveries at Harappā were made to the south of the citadel. Here two cemeteries were found—"R. 37," belonging to the Harappan period, and "H," dating from the late or even post-Harappan period. These contained different styles of burial and will be discussed below.

Granary buildings

Kāli Banga. Third place in magnitude and importance among excavated Harappan sites must now be accorded Kāli Banga. Excavations by the Archaeological Survey of India have only recently been completed and the results are still largely unpublished. Kāli Banga stands on the left bank of the dry bed of the Saraswatī River in northern Rājasthān. As mentioned above, a pre-Harappan settlement lies beneath the later remains, and the main Harappan township has a layout strikingly similar to that of Mohenjo-daro and Harappā. In the lower town excavation has revealed as many as nine building phases. The citadel mound is a parallelogram on a plan of about 430 feet (130 metres) on the east–west axis and 850 feet on the north–south. The whole site has been drastically reduced by brick robbers, but careful excavation revealed the foundation courses of an accurately laid rhomboid central section with oblong bastions at each corner and smaller bastions on the north and south walls. The principal access was from the south via a flight of steps. Access from the north was via a narrow postern reached by a stairway, beyond which was a further rhomboid section, having an inset gateway in the northwest corner, near the riverbank. Traces of a brick wall around the lower town were also encountered. The central sector of the citadel contained a series of high brick platforms divided by narrow passages. The upper parts of these platforms had been very seriously damaged, and their function is at present mysterious, but they do not appear to have been the foundation for a granary.

The northern sector contained normal domestic housing. A cemetery was discovered a short distance to the west of the town. It may be expected that when the excavation of this site is published, it will add greatly to knowledge of the Indus civilization.

Lothal. One other among the excavated sites deserves special attention; this is Lothal, a small settlement built on low-lying ground near a tributary of the Sābarmatī River on the west side of the Gulf of Cambay. It appears to have served as a port or trading station. Its layout is distinctive: the site is roughly rectangular, measuring about 1,180 feet on the long north–south axis and 690 feet on the east–west. It was surrounded by a massive brick wall, probably used for flood protection. The southeastern quadrant takes the form of a great platform of brick with earth filling, rising to a height of about 13 feet. On this were built a series of further smaller platforms with intersecting air channels, reminiscent of the granary at Mohenjo-daro, with overall dimensions of about 159 by 139 feet. Behind this block were other buildings including a row of 12 bathrooms with connected drains, also strongly reminiscent of those found on the citadel at Mohenjo-daro. The remaining enclosed area was evidently taken up by houses and shops. Among the significant finds were a bead maker's factory and the shops of goldsmiths and coppersmiths. The main street ran from north to south.

The Lothal dock

The most unexpected discovery at Lothal, however, was a great brick basin measuring some 718 by 121 feet with extant brick walls of 15 feet in height. This lay east of the settlement, alongside the platform on which the granary block stood. At one end of the basin was a small sluice or spillway with a locking device.

The excavator has inferred that the basin was a dock to which ships could be brought from the nearby estuary via an artificial channel that would have been kept clear of silt by the controlled flow of water from the spillway. This view has not been universally accepted, but no other explanation of the function of the basin appears to be so satisfactory. A cemetery was found outside the perimeter of the wall, west of the site. The final report on this excavation also is not yet published.

Other important sites. Mention has been made above of the pre-Harappan settlements at Amri and Kot Diji in Sind. Each of these lies beneath a Harappan settlement, but neither compares in size with the major cities. Moreover, the area hitherto excavated is comparatively small, and therefore little can be said regarding the later period except that in each case there are the same indications of the cultural uniformity that dominates all aspects of the civilization. The excavator at Amri divides the period into four phases, the last being equivalent to the post-Harappan phase named after the type-site of Jhukar. Chanhu-daro is another site in Sind, some 20 miles east of Amri, excavated in 1935–36, and showing three phases of Harappan culture followed by the distinctive Jhukar phase and a subsequent phase with coarse gray pottery, probably dating to *c.* 1000 BC and named after the typesite of Jhangar. The Harappan remains were characteristic, and a remarkable bead maker's factory was also discovered. Another excavated site is at Rūpar northwest of Chandīgarh in the Punjab, but the report on this work is not yet published.

Population. There have been two independent estimates of the population of Mohenjo-daro. Both are based on an estimation of the original area covered and the density of the people living there, using traditional settlements in the region in recent times for comparison. H.T. Lambrick proposed a figure of 35,000 for Mohenjo-daro and a roughly similar figure for Harappā; while W.A. Fairservis estimated the former at around 41,250 and the latter around 23,500. These figures are probably conservative. It would be possible to produce estimates of the population for other sites along similar lines—notably for Kāli Banga, of which the lower city has an area about one-fifth that of Mohenjo-daro.

Recently the physical type of skeletons recovered from Harappan sites has been studied in much greater detail than hitherto. The results have shown that the earlier classifications of very small samples that included a variety of racial types—Mongoloid, Proto-Australoid, Mediterranean, and Alpine—are largely meaningless and that the Harappan population was far more homogeneous than has so far been supposed. The people of Mohenjo-daro and Harappā appear to have been tall, long-headed, and high-domed; they were markedly broad-nosed. The population of Lothal seems to have had a more notably brachycephalic (broad-skulled) element. To postulate the presence of several distinct races thus appears to be erroneous on the basis of this evidence.

Agriculture and animal husbandry. It is certain that such great concentrations of population had never been seen in India-Pakistan before that date. Nor could the hunting–fishing and collecting economy of the Late Stone Age groups, nor the small-scale agriculture of the dry valleys of Baluchistan and of the settlements east of the Indus, have sustained such numbers. Clearly the exploitation of the Indus floodplains and the use of the plow attested in pre-Harappan times by finds in Kāli Banga were matters of supreme importance. Lambrick has shown how the traditional exploitation of the floods could provide a simple means of growing the principal crops without even plowing, manuring, or using major irrigation. The main cereals would be sown at the end of the inundation on land that had recently emerged from the floods, and the crop would be harvested in March or April. Other crops might be sown in embanked fields at the beginning of the floods so that they could receive necessary water while growing and be harvested in the autumn. Wheat samples from the Indus cities have been identified as of two subspecies of *Triticum sativum, vulgare* and *compactum,* and of *Triticum sphaerococcum.* Barley is also found, of the species *Hordeum vulgare,* variety *nudum* and variety *hexastichum.* Rice is recorded in Harappan times at Lothal in Gujarāt, but whether wild or cultivated is not yet clear. Other crops include dates, melon, sessamum, and varieties of leguminous plants, such as field peas. From Chanhu-daro seeds of mustard (most probably *Brassica juncea*) were obtained. Finally, there is evidence that cotton was cultivated and used for textiles.

Principal crops and domesticated animal species

A number of domesticated animal species have been found in excavations at the Harappan cities. The Indian humped cattle (*Bos indicus*) were most frequently encountered, though whether along with a humpless variety, such as that shown on the seals, is not clearly established. The buffalo (*Bos bubalis*) is less common and may have been wild. Sheep and goats occur, as does the Indian pig (*Sus cristatus*). The camel is present as well as the ass (*Equus asinus*). Bones of domestic fowl are not uncommon; these fowl were domesticated from the indigenous jungle fowl. Finally, the cat and dog were both evidently domesticated. Present, but not necessarily as a domesticated species, is the elephant. The horse is possibly present but extremely rare and apparently only present in the last stages of the Harappan period.

Communications. It is clear that to achieve the degree of uniformity of material culture evidenced in the excavations, considerable contact must have been maintained between the towns and cities of the Indus state. Such contact may have been by both land and river, just as the foreign trade must have employed both overland and sea routes. For land travel the predominant means was probably the pack bullock, camel, or ass. All of these animals are still, or were until recently, used for pack transport in the more remote country districts of India-Pakistan. For travel on the flat alluvial plains, the bullock cart was probably the main vehicle. Terra-cotta models of such carts, apparently very little different from the modern Indian cart, are frequently encountered. For the transport of persons, smaller carts, with a body raised above the level of the axle and a framed canopy (much like the modern *ikkā*), are known from small bronze models. Several representations of boats also occur. They are mostly of simple design without masts or sails, and would be more suitable for river travel than for sea travel. A terra-cotta model of another type of boat with a socket for mast and eye holes for rigging was discovered at Lothal. This appears to be a somewhat more seaworthy vessel. The dock basin at Lothal could provide berth for ships of the size of the country craft that still ply between India and the Persian Gulf. Heavy pierced stones

discovered in the vicinity of the dock basin at Lothal were assumed by the excavator to be similar to stones still used by the local boatmen as anchors.

Use of copper and bronze

Craft and technology. The Indus civilization reveals a wide range of crafts and technical skills. It may be remarked that while these skills ultimately depended upon the same discoveries as were exploited by the Egyptians or the Sumerians, their employment in India is often distinctive and gives to each craft a character of its own.

Throughout the period copper and bronze were the principal metals used for making tools and implements. These include flat oblong axes, chisels, knives, spears, arrowheads (of a kind that was evidently exported to neighbouring hunting tribes), small saws, and razors. All these could be made by simple casting, chiselling, and hammering. Throughout the sites, bronze is less common than copper, and it is notably rarer in the lower levels. Four main varieties of metal have been found: crude copper lumps in the state in which they left the smelting furnace; refined copper, containing trace elements of arsenic and antimony; an alloy of copper with 2 to 5 percent of arsenic; and bronze with a tin alloy, often of as much as 11 to 13 percent. (A copper arsenic alloy has also been found in Egypt.) The copper and bronze vessels of the Harappans are among their finest products, formed by hammering sheets of metal. Casting of copper and bronze was understood, and figurines of men and animals were made by the cire-perdue (lost-wax) technique. These, too, are technically outstanding, though the overall level of copper–bronze technology is not considered to have reached the level attained in Mesopotamia.

Other metals used were gold, silver, and lead. The latter was employed occasionally for making small vases and such objects as plumb bobs. Silver is relatively more common than gold, and more than a few vessels are known, generally in forms similar to copper and bronze examples. Gold is by no means common and was generally reserved for such small objects as beads, pendants, and brooches. Much of the Indus gold is light in colour, indicating a high silver content.

Other special crafts include the manufacture of faience (earthenware decorated with coloured glazes)—for making beads, amulets, and sealings, and for making small vessels—and the working of stone for bead manufacture and for seals. The seals were generally cut from steatite and were carved in intaglio or incised with a copper burin (cutting tool). Beads were made from a variety of substances, but the carnelians are particularly noteworthy. They include several varieties of etched carnelian and long barrel beads made with extraordinary skill and accuracy. Shell and ivory were also worked and were used for beads, inlays, combs, bracelets, and the like.

After copper and bronze, stone probably played a major role in the life of the Harappans. Blocks of chert were dressed at the quarries at Sukkur on the Indus, some 50 miles northeast of Mohenjo-daro, and cores were prepared from which stone blades of great regularity and size were struck. These blades are a common find at almost all of the Harappan sites. Other special uses for stone were in making bowls and small vessels of alabaster (probably imported); while hammer-dressed blocks of limestone were imported from the nearby hills for occasional use, for example, as covers for brick drains.

Pottery

The pottery of the Indus cities has all the marks of mass production. A substantial proportion is thrown on the wheel (probably the same kind of footwheel that is still found in the Indus region and to the west to this day, as distinguished from the Indian spun wheel common throughout the remaining parts of the subcontinent). The majority of the pottery is competent plain ware, well formed and fired but lacking in aesthetic appeal. A substantial portion of the pottery has a red slip and is painted with black decoration. Larger pots were probably built up on a turntable. Among the painted designs, conventionalized vegetable patterns are common, and the elaborate geometric designs of the painted pottery of Baluchistan give way to simpler motifs, such as intersecting circles or a scale pattern. Birds, animals, fish, and more interesting scenes are comparatively rare. Of the vessel forms, a shallow platter on a tall stand (the so-called offering stand) is noteworthy, as is also a tall cylindrical vessel perforated with small holes over its entire length and often open at top and bottom. The function of this latter vessel remains a mystery.

Although little has survived, very great interest attaches to the fragments of cotton textiles recovered at Mohenjo-daro. These provide the earliest evidence of a crop and industry for which India has long been famous. It is assumed that the raw cotton must have been brought in bales to the cities to be spun, woven, and perhaps dyed, as the presence of dyers' vats would seem to indicate.

Trade and external contacts. It has been seen above that the area covered by the Indus civilization had a remarkably uniform level of material culture. Such uniformity suggests a closely knit and integrated administration and implies internal trade within the state. Evidence of the actual exportation of objects is not always easy to find, but the wide diffusion of chert blades made of the characteristic Sukkur stone and the enormous scale of the factory at the Sukkur site strongly suggest trade. Other items also appear to indicate trade, such as the almost identical bronze carts discovered at Chanhu-daro and Harappā, for which a common origin must be postulated.

The wide range of crafts and special materials employed must also have caused the establishment of economic relations with peoples living outside the Harappan state. Such trade may be considered to be of two kinds: first, the obtaining of raw materials and other goods from the village communities or forest tribes in regions adjoining the Indus culture area; and second, trade with the cities and empires of Mesopotamia. There is ample indication of the former type, even if the regions from which specific materials were derived are not easy to pinpoint. Gold was almost certainly imported from the group of settlements that sprang up in the vicinity of the goldfields of northern Karnātaka; and copper could have come from several sources—some certainly from Rājasthān, some perhaps from East or South India on the one side, and some from Baluchistan or farther west. Lead may have come from Rājasthān or elsewhere in India. Lapis lazuli was probably imported from Iran rather than directly from the mines at Badakhshan, for the recent discovery of a large depot in Sīstān seems to indicate the centre from which this material was traded throughout the ancient East. Turquoise probably came from Iran; fuchsite from Karnātaka; alabaster from Iran; amethyst from Mahārāshtra; and jade from Central Asia. There is unfortunately little evidence of what the Harappans gave in exchange for these materials—possibly nondurable goods such as cotton textiles and probably various types of beads. They may have also bartered tools or weapons of copper.

Trade with Mesopotamia

For the trade with Mesopotamia there are literary evidences as well as archaeological. The Harappan seals were evidently used to seal bundles of merchandise, as clay seal impressions with cord or sack marks on the reverse side testify. The presence of a number of Indus seals at Ur and other Mesopotamian cities and the discovery of a "Persian Gulf" type of seal at Lothal—otherwise known from the Persian Gulf ports of Bahrain (ancient Dilmun, or Tilmun) and Faylahkah, and from Mesopotamia—provide convincing corroboration of the sea trade suggested by the Lothal dock. Timber and precious woods, ivory, lapis lazuli, gold, and luxury goods such as carnelian beads, pearls, shell and bone inlays, including the distinctly Indian kidney shape, were among the goods sent to Mesopotamia in exchange for silver, tin, woollen textiles, and grains and other foods. Copper ingots appear to have been imported to Lothal from Magan (possibly Oman, the Mahran region, or southeastern Iran). Other possible trade items include products originating exclusively in each respective region, such as bitumen, occurring naturally in Mesopotamia; and cotton textiles and chickens, major products of the Indus region not native to Mesopotamia.

Mesopotamian trade documents, lists of goods, and official inscriptions mentioning Meluhha (the ancient Akkadian name for the Indus region), supplement Harappan seals and archaeological finds. Literary references to meluhhan trade date from the Akkadian, Ur III, and

Isin and Larsa periods (*i.e.*, *c.* 2350–1770 BC), but as texts and archaeological data indicate, the trade probably started in the Early Dynastic period, *c.* 2600 BC. During the Akkadian period Meluhhan vessels sailed directly to Mesopotamian ports, but by the Isin and Larsa period, Dilmun (modern Bahrain) was the entrepôt for Meluhhan and Mesopotamian traders. By the subsequent Old Babylonian period, trade between the two cultures evidently had ceased entirely.

Language and scripts, weights and measures. The maintenance of so extensive a set of relations as those implicit in the size and uniformity of the Harappan state and the extent of trade contacts must have called for a well-developed means of communication. The Harappan script has long defied attempts to read it, and therefore the language remains unknown. Recent analyses of the order of the signs on the inscriptions have led several scholars to the view that the language is not of the Indo-European family; nor is it close to Sumerian, Hurrian, or Elamite; nor can it be related to the structure of the Munda languages of modern India. If it is related to any modern language family it appears to be the Dravidian, presently spoken throughout the southern part of the Indian Peninsula. The script, which was written from right to left, is known from the 2,000-odd short inscriptions so far recovered, ranging from single characters to inscriptions of around 20 characters. There are more than 500 signs, many appearing to be compounds of two or more other signs; but it is not yet clear whether these signs are ideographic, logographic, or other. Numerous attempts to read the script have been made during the past decades, most recently by a Russian team under Y. Knorozov and a Scandinavian group led by A. Parpola. The Scandinavian group claims to have read many of the inscriptions, taking the language to be Dravidian, akin to Old Tamil, but few scholars are as yet prepared to accept their results.

The Harappans also employed regular systems of weights and measures. An analysis of a fair number of the well-formed chert weights indicates that they followed a binary system for lower denominations and a decimal system for higher; that is, a series of 1, 2, 4, 8, 16, 32, 64, and then 320, 640, 1,600, 3,200, 8,000. Several scales of measurement were found in the excavations, one a decimal scale of 1.32 inches (3.35 centimetres) rising probably to 13.2 inches, apparently corresponding with the "foot" that was widespread in western Asia; another is a bronze rod marked in lengths of 0.367 inches, apparently half a digit of a "cubit" of 20.7 inches, also widespread in western Asia and Egypt. Measurements from some of the structures show that these units were accurately applied in practice.

It has also been suggested that certain curious objects may have been accurately made optical squares with which surveyors might offset right angles, and in view of the accuracy of so much of the architectural work this appears quite plausible.

Art. The excavations of the Indus cities have produced much evidence of artistic activity. Such finds are important because they provide an insight into the minds, lives, and religious beliefs of their creators. Stone sculpture is extremely rare, and much of it is very crude. The

Sculpture

total repertory cannot compare with the work done in Mesopotamia during the same periods. The figures are apparently all intended as images for worship. Such figures include seated men, recumbent composite animals, or in unique instances (from Harappā) a standing nude male and a dancing figure. The finest pieces are of excellent quality. There is also a small but impressive repertory of cast-bronze figures, including several fragments and complete examples of dancing girls, small chariots, carts, and animals. The technical excellence of the bronzes suggests a highly developed art, but the number of examples is still very small. Nevertheless, they appear to be Indian workmanship rather than imports.

The popular art of the Harappans was in the form of terra-cotta figurines. The majority are of standing females, often heavily laden with jewelry, but standing males—some with beard and horns—are also present. It has been generally agreed that these figures are largely deities (perhaps a Great Mother and a Great God); but some small figures of mothers with children or of domestic activities are probably toys. There are varieties of terra-cotta animals, carts, and toys—such as monkeys pierced to climb a string and cattle that nod their heads. Painted pottery is the only evidence that there was a tradition of painting. Much of the work is executed with boldness and delicacy of feeling, but the restrictions of the art do not leave much scope for creativity. The steatite seals, to whose manufacture reference was made above, form the most extensive series of objects of art in the civilization. The great majority show a humpless "unicorn" or bull in profile, while others show the Indian humped bull, the elephant, bison, rhinoceros, or tiger. The animal frequently stands before a ritual object, variously identified as a standard, manger, or even incense burner. A considerable number of the seals contain scenes of obvious mythological or religious significance. The interpretation of these seals is, however, often highly problematic. The seals were certainly more widely diffused than other artifacts and show a much higher level of workmanship. Probably they functioned as amulets as well as more practical devices to identify merchandise.

Religion and burial customs. In spite of the unread inscriptions, there is a considerable body of evidence that allows for conjecture concerning the religious beliefs of the Harappans. First, there are the buildings identified as temples or as possessing a ritual function, such as the Great Bath. Then there are the stone sculptures found to a large extent associated with these buildings. Next, there are the terra-cotta figures, and finally, the seals and amulets that depict scenes with evident mythological or religious contents. The interpretation of such data necessarily involves a largely subjective element, but most commentators have felt that they indicate a religious system that was already distinctly Indian. It is assumed that there was a Great God—who had many of the attributes later associated with the Hindu god Śiva—and a Great Mother who was the Great God's spouse and shared the attributes of Śiva's wife Durgā-Pārvatī. Evidence also exists of some sort of animal cult, related particularly to the bull, the buffalo, and the tiger. Mythological animals include a composite bull-elephant. Some seals suggest influence from or at least traits held in common with Mesopotamia; among these are the Gilgamesh (Mesopotamian epic) motif of a man grappling with a pair of tigers and the bull-man Enkidu (a human with horns, tail, and rear hooves of a bull). Among the most interesting of the seals are those that depict cult scenes or symbols; a god, seated in a yogic (meditative) posture and surrounded by beasts, with a horned headdress and erect phallus; the tree spirit with a tiger standing before it; the horned tree spirit confronted by a worshipper; a composite beast with a line of seven figures standing before it; the pipal leaf motif; and the swastika.

Many burials have been discovered, giving clear indica-

Form of burial

tion of belief in an afterlife. The cemeteries excavated at Harappā, Lothal, and Kāli Banga are clearly separated from the settlement and show that the predominant rite was extended inhumation, with the body lying on its back and the head generally positioned to the north. Quantities of pottery were placed in the graves, and sometimes personal ornaments adorned the bodies. Some graves took the form of brick chambers within which the body was placed. At Lothal several pairs of skeletons were found in the same grave, and it has been suggested that this is an indication of some form of suttee (*satī*; a later Hindu custom in which the wife ends her life after the death of her husband).

The end of the Indus civilization. Around 1750 BC the uniform culture of this great area broke up, apparently with different results in the various provinces. The cause or causes of the end of the Indus civilization are not easy to determine. At Mohenjo-daro the latest levels show a marked decline in civic standards, and small groups of sprawling skeletons in this period suggest some sort of violence and disorder, perhaps a raid. At Harappā in the top strata of the citadel there was a similar decline, and a new pottery style occurs alongside the Harappan. Such pottery is found also in the cemetery H, where it is associated with two periods of burial, the earlier inhumations and the later fractional burials in urns. At Chanhu-daro and Amri

there is no evidence of a major break in the occupation, but a distinct pottery and material culture seems to have taken over (the so-called Jhukar culture). At Lothal, too, in the late period, a somewhat similar material culture follows directly on the Harappan.

How is this evidence to be interpreted, and where in all this, if at all, are traces of the arrival of early groups of Indo-European-speaking tribes, or of the Aryans to be found? It has recently been suggested that the end of Mohenjo-daro was associated with a vast and calamitous flood brought on by tectonic (structural) change farther down the Indus, but several serious objections to this view have been raised. More plausible is the suggestion that the Indus suffered one of its periodic major alterations of course, cutting off the floodwaters from economically vital agricultural land and resulting either in desiccation around Mohenjo-daro or in an influx of refugees to that city from elsewhere. Another somewhat similar and equally plausible explanation sees the growth of population leading to an over-expolitation of the countryside and hence to the exhaustion of necessary food supplies for men or cattle, and to deforestation. Whatever the causes, they must have led to a growing strain within the system and must have made the city an easy prey for raiders. There is likely to have been more than one wave of Indo-Iranian-speaking peoples into India-Pakistan, and archaeology can supply evidence of several influxes between 1750 and 1000 BC. Such evidence, however, is still extremely difficult to interpret and may largely reflect a period long after the breakdown of the Harappan civilization. The period succeeding this event used to be called a dark age, and in terms of scholarly knowledge this was once justified; but archaeology is steadily throwing more light upon the hiatuses, and in many parts of India the period is becoming quite well known. What emerges is that, during and after the downfall of the Harappan civilization, a whole series of regional cultures appear in India-Pakistan, in some cases as direct descendants of the old Neolithic cultures, in others as new developments. These cultures appear to have developed out of, or shared in some ways the legacy of the Harappans and in their turn provide the framework for the integration of Indian and Aryan elements that contributed to the rise of civilization in the Ganges Valley. Most scholars have concluded that later Indian civilization owes an incalculable debt to the Harappans and that an as yet indefinable part of later Indian tradition and ideology must derive from them. Thus, it is probable that, by analogy with such tangible evidence of survival as the types of carts or ornaments, one may legitimately infer the Harappan origin of less tangible aspects of Indian life and institutions, such as the caste system or certain patterns of thought. (F.R.Al./Ed.)

The development of Indian civilization from c. 1500 BC to c. AD 1200

Though the Indian people have been concerned with both material and spiritual aspects of life, Indian history has in general been thought of as static and concerned only with things spiritual. Indologists, preeminently a 19th-century German, Max Müller, relied heavily on the Sanskritic tradition and saw Indian society as an idyllic village culture emphasizing qualities of passivity, meditation, and otherworldliness. In sharp contrast was the approach of the British historian James Mill and the Utilitarians, who condemned Indian culture as irrational and inimical to human progress. Mill first formulated a periodization of Indian history into Hindu, Muslim, and British periods. Direct contact with Indian institutions through administration in the 19th century, together with the utilization of new evidence from recently deciphered inscriptions, numismatics, and local archives, has provided fresh insights. Nationalist Indian historians of the early 20th century have tended to exaggerate the glory of the past but nevertheless, introduced controversy into historical interpretation, which in turn resulted in more precise studies of Indian institutions of the past.

A major change in the orientation of Indian history has been a questioning of an older notion of Oriental despotism as the key to Indian history. Arising out of a traditional European perspective on Asia, this image of despotism grew to enormous proportions in the 19th century and provided an intellectual justification for both colonialism and imperialism. Its deterministic assumptions clouded the precise understanding of early Indian interrelationships between political forms, economic patterns, and social structure.

The change in the orientation of Indian historical studies

In a study of political forms during this period, a considerable change is noticeable in the role of the institutions. Tribal societies had assemblies, whose political role changed with the transformation of tribe into state and with republican and monarchical governments. Centralized imperialism, which emerged in the Mauryan period, gave way gradually to decentralized administration and to what has been called a type of feudalism in the post-Gupta period. Although the village as an administrative and social unit remained constant, its relationship with the mainstream of history varied. Divinity in kingship was known but rarely taken seriously, the claim to the status of the caste of royalty becoming more important. Because conformity to the social order had precedence over allegiance to the state, the idea of representation found expression not so much in political institutions as in caste and village assemblies. The pendulum of politics swung from large to small kingdoms, with the former attempting to establish empires—the single, totally successful attempt being that of the Mauryan. Thus, true centralization was rare because local forces were often the determinants of historical events. Though imperial or near-imperial periods were marked by attempts at the evolution of uniform cultures, the periods of smaller kingdoms (often referred to as the Dark Ages by earlier historians) were more creative at the local level and witnessed significant changes in society and religion. It was also these small kingdoms that boasted of the most elaborate and impressive monuments.

The major economic patterns were those relating to land and to commerce. The transition from tribal to peasant society was a continuing process, with the gradual clearing of wasteland and the expansion of the village economy based on plow agriculture. Recognition of the importance of land revenue coincided with the emergence of the imperial system in the 4th century BC; and from this period, although the imperial structure did not last long, land revenue became central to the administration and income of the state. Frequent mentions of individual ownership, references to crown lands, numerous land grants to religious and secular grantees in the post-Gupta period, and detailed discussion in legal sources of the rights of purchase, bequest, and sale of land all clearly indicate that there was private property in land. Much emphasis has been laid on the state control of the irrigation system; yet a systematic study of irrigation in India reveals that it was generally privately controlled and that it serviced small areas of land. When the state built canals, they were mainly in the areas of the winter and summer monsoons and where village assemblies played a dominant part in revenue and general administration, as for example in the Cōḷa kingdom of South India.

The urban economy was crucial to the rise of the Indus Valley civilization (*c.* 2300–*c.* 1750 BC). Later, the 1st millennium BC saw an urban civilization in the Ganges Valley and in coastal South India. The emergence of towns was based on the needs and requirements of trade. In the 1st millennium AD, when commerce expanded to include trade with western Asia, the eastern Mediterranean, and Central and Southeast Asia, revenue from trade contributed substantially to the economies of the participating kingdoms, as indeed Indian religion and culture played a significant part in the cultural evolution of Central and Southeast Asia. Gold coins were issued for the first time by the Kuṣāṇas and in large quantity by the Guptas, both of which kingdoms were active in foreign trade. Gold was imported from Central Asia and Rome and later perhaps from eastern Africa because, in spite of the country's recurring association with gold, its sources were limited. Expanding trade encouraged the opening up of new routes, and this coupled with the expanding village economy led to a marked increase of knowledge about the geography of

Emergence of urban civilization

the subcontinent during the post-Maurya period. With increasing trade, guilds became more powerful in the towns. Members of the guilds participated in the administration, were associated with politics, and controlled the development of trade through merchant embassies sent to places as far afield as Rome and China. Not least, guilds and merchant associations held envied and respectable positions as donors of religious institutions.

The structure of Indian society was unique in being characterized by caste, which largely accounts for the survival of that structure to the present day. The constant features of a caste society were endogamous kinship groups (*jāti*) arranged in a hierarchy of ritual ranking, based on notions of pollution and purity, with an intermeshing of service relationships. Although ritual hierarchy was unchanging, there appears to have been considerable mobility within the framework. Migrations of peoples both within the subcontinent and from outside encouraged social mobility and change. The nucleus of the social structure was the family, with the pattern of kinship relations varying from region to region. The family related to the village, and the village in turn to the region. The urban structure was a more complex one in which the guilds occasionally took on *jāti* functions, and there was a continual emergence of new social and professional groups.

Religion in early Indian history did not act as a monolithic force. Even when attempts were made by royalty to encourage certain religions, there was no idea of a state religion. In the main, there were three levels of religious expression. The most widespread was the worship of local cult deities with vague associations of a major deity, as seen in fertility cults, in the worship of mother-goddesses, in the so-called Śākta-Śakti cult, and in Tantrism. Less widespread but popular, particularly in the urban areas, were the more puritanical sects of Buddhism, Jainism, and the *bhakti* tradition of Hinduism. A third level included classical Hinduism and more abstract levels of Buddhism and Jainism, with an emphasis on the major deities in the case of the former and on the teachings of the founders in the case of the latter. It was this level, endorsed by affluent patronage, that provided the base for the institutionalization of religion. But the three levels were not isolated; the shadow of the third fell over the first two, whose more homely rituals and beliefs often crept into the third. This was the case particularly with Hinduism, the very flexibility of which was largely responsible for its survival. Forms of Buddhism, ranging from an emphasis on the constant refinement of doctrine, on the one hand, to an incorporation of magical fertility cults in its beliefs, on the other, faded out toward the end of this period.

Sanskrit literature and the building of Hindu and Buddhist temples and sculpture both reached apogees in this period. Although literary works in Sanskrit continued to be written and temples were built in later periods, the achievement was never again as inspiring.

C. 1500–C. 600 BC

Decline of the Harappā culture

By about 1500 BC an important change began to occur in the northern half of the Indian subcontinent. The Harappā culture in the Indus Valley had declined by about 1750 BC, and the stage was being set for a second and more continuous urbanization in the Ganges Valley. The Harappā decline resulted in the collapse of city civilization and marked the end of the Indus region as a centre of civilization. The new areas of occupation were contiguous with, but seldom identical to, the Harappan area. There was continuity of occupation in the Punjab (Pañjāb) and Gujarāt. Geological and ecological factors, such as changing river courses, the encroachment of the desert, and repeated flooding, may well have made the area physically inhospitable, ultimately resulting in the decline of the cities.

Urbanization in the Ganges Valley. A new thrust toward urbanization came from the migration of peoples from the Punjab into the Ganges Valley.

Early Ganges cultures. Coterminous with the Harappā culture was a series of Neolithic cultures with their nuclei in Kashmir, eastern India, and the Deccan. In the immediate post-Harappan phase, the Ganges Valley sheltered two cultures, which scholars have so far been unable to connect. The Doab was dotted with small settlements using a crude ochre-colour pottery, rolled, and heavily weather-beaten through wind or water effects; its origin is uncertain, and the suggestion that it was a degenerate pottery produced by migrants from the Harappā culture moving eastward generally has not been accepted. The Ganges Valley to the east of the Doab has produced isolated hoards of copper objects—harpoons, spearheads, celts, swords, anthropomorphic figures—often discovered in fields rather than in excavations. Both the Ochre Colour Pottery people and the Copper Hoard people date to approximately 1800–1200 BC.

Substantial remains in the Doab show that, subsequent to the Ochre Colour Pottery culture, there was a so-called Painted Gray Ware culture, which formed the substratum to the urbanization of the region, with its upper limit in about the 5th century BC. Remains suggest communities of pastoralists and agriculturalists who used a distinctive gray pottery decorated generally in black with floral and geometrical patterns.

The Painted Gray Ware culture was strikingly different from the Harappā culture. Apart from the fact that it was not an urban culture, it introduced a new metal technology of iron, which facilitated the transition to urbanization in the Ganges Valley. It also introduced the horse, the domestication of which had been unknown to the country. A gradual change in food habits from wheat and barley in the Doab to rice in the east resulted in changing pottery shapes. To the east of the Doab, the appearance of what has come to be called the Northern Black Polished Ware coincided with urbanization.

The earliest literary evidence in India is the Vedic literature. The Vedas, an orally transmitted tradition during the early period, consist of four major texts—the Rig (Ṛg), Sāma, Yajur, and Atharva Veda. Of these, the Rigveda (Ṛgveda) is believed to be the earliest. The language of the texts is Vedic Sanskrit, and the contents relate to hymns, charms, spells, and ritual observations of the Aryan-speaking peoples in India. ("Aryan" is used here only as a linguistic term, referring to the people who spoke Indo-Aryan, and does not carry any ethnic connotation.)

Origin of the Aryans in India

Theories of the origin of the Aryans in India relate to the question of what has been called the Indo-European homeland. In the 17th and 18th centuries AD, European scholars who first studied Sanskrit were struck by the similarity in its syntax and vocabulary to Greek and Latin. This resulted in the theory that there had been a common ancestry for these and other related languages, which came to be called the Indo-European group of languages. This, in turn, resulted in the notion that the Indo-European speaking people had had a common homeland from which they had migrated to various parts of Asia and Europe. The theory led to unlimited speculation, which continues today, regarding the original homeland and the date of the dispersal from it. The early history of India is still beset by "the Aryan Problem," which often clouds a genuine search for historical insight into this period.

That there was a migration of Indo-European speakers, possibly in waves, which can be dated to the 2nd millennium BC, is clear from archaeological and epigraphic evidence in western Asia. Mesopotamia witnessed the arrival, in about 1760 BC, of the Kassites, who introduced the horse and the chariot and bore such obviously Indo-European names as Surias, Indas, and Maruttas (Sūrya, Indra, and Marutaḥ in Sanskrit). A treaty *c.* 1400 BC between the Hittites, who had arrived in Anatolia at about the beginning of the 2nd millennium BC, and the Mittannis invoked four deities—Indara, Uruvna, Mitira, and the Nāsatyas (names that occur in the Rigveda as Indra, Varuṇa, Mitra, and the Aśvins). An inscription at Boğazköy in Anatolia of about the same date refers to certain Indo-European technical terms in the training of horses. Clay tablets dating to about 1400 BC, written at Tell el-Armarna in Babylonian cuneiform, mention the names of princes, such as Biridashva and Artamanya, which are also Indo-European. The association of the horse with these peoples would point to Central Asia or the southern Russian steppes as their place of origin.

Nearer India, the Iranian Plateau was subject to a similar migration. The literature of the Iranian Aryans, the Avesta, when compared with Vedic literature indicates that there had once been a close relationship between the two. It would appear that a branch of the Iranian Aryans migrated to northern India and settled in the Sapta Sindhu region, the area extending from the Kābul River in the north to the Saraswatī and upper Doab in the south. The Saraswatī, the sacred river at the time, flowed through northern Rājasthān but soon after disappeared into the desert. It was in the Sapta Sindhu that the majority of the hymns of the Rigveda were composed.

The Rigveda is divided into 10 *maṇḍala*s (books), of which the 10th is believed to be somewhat later than the others. Each *maṇḍala* consists of a number of hymns, and most *maṇḍala*s are ascribed to a priestly family, such as Kaṇva, Gautama, Viśvāmitra, Vāmedeva, Atri, Bhāradvāja, and Vasiṣṭha. The ninth *maṇḍala* is a collection of all the hymns dedicated to *soma,* the sacred and intoxicating juice that was drunk on ritual occasions. The hymns include invocations to the gods, ritual hymns, battle hymns, narrative dialogues, and satirical observations. They provide incidental information on geographical location, social and political organization, tribal names and incidents connected with them, and, to a much greater degree, the evolution of Vedic religion.

Few events of political importance are related to the hymns. Perhaps the most impressive of those that are is a description of the battle of the 10 kings in the Dasarājanya hymn: when Sudās, the king of the preeminent Bharata tribe of southern Punjab, replaced his chief priest Viśvāmitra with Vasiṣṭha, Viśvāmitra organized a confederacy of 10 tribes, including the five famous tribes of the Pūru, Yadu, Turvaśas, Anu, and Druhyu, which went to war against Sudās. The Bharata tribe survived and continued to play an important role in historical tradition. Among the tribes there is little distinction between Aryan and non-Aryan, but the hymns refer to a local people, the Dāsas, who are said to have had an alien language and a dark complexion and to worship strange gods. Some Dāsas were rich in cattle and lived in fortified places (*pur*) often attacked by the god Indra. In addition to the Dāsas, there were the wealthy Paṇis, who were hostile and stole cattle.

Settled village communities

This was the period of transition from nomadic pastoralism to settled village communities (*grāma*) with an intermixture of pastoral and agrarian economies. That the notion of cattle remained the dominant factor is indicated by the use of such words as *gotra* ("cowpen") for the endogamous kinship group and *gaviṣṭi* ("searching for cows") for war. That the family structure was patriarchal and the extended family constituted the unit is seen in the practice of *niyoga* ("levirate"), which permitted a widow to marry her husband's brother. A group of families constituted a *grāma.* The tribal unit was the *viś.* Tribal assemblies appear to have been frequent in the early stages, with meetings for both secular and religious purposes. Various categories of assemblies are referred to, such as the *vidatha, pariṣad, sabhā,* and *samiti,* the latter two being more common, although the precise distinction between them is not clear. The king, whose powers were checked by the *sabhā* and *samiti,* was primarily a war leader with the responsibility for protecting the tribe. He received a *bali* (tribute) for performing this function, which included the maintenance of civil order. Punishment was based on the system of wergild, in which a slain man's family was compensated according to the value of his life. The chief priest, or *purohita,* was largely a magic man whose rituals ensured prosperity and valour for the king. The other occasion for the tribal assembly was the *yajña* (religious sacrifice), which was the central ritual of Vedic religion and involved a large hierarchy of priests. The ceremonies lasted many days and entailed the conspicuous consumption of wealth in kind.

Later Vedic period (c. 1000–500 BC). The sources for this period are both archaeological and literary. In the Ganges Valley the Painted Gray Ware culture continues to provide the major archaeological evidence. The literary sources are primarily the Sāmaveda, Yajurveda, and Atharvaveda, mainly ritual texts that include the *Brāhmaṇa* as well as the *Upaniṣad* and the *Āraṇyaka*—collections of philosophical and metaphysical discourses. Associated with these Vedas are the *sūtra* texts, largely explanatory aids to the other sections, being manuals on sacrifices and ceremonies, domestic observances, and social and legal relations. Because the texts were continually revised, they cannot be dated accurately to the early period. The *Dharma-sūtra* texts of this period became the nuclei of the sociolegal *Dharma-śāstra*s of later centuries.

Some historians date the two epics, the *Mahābhārata* and the *Rāmāyaṇa,* to this period, though the date of each is a subject of considerable controversy. Both are mixtures of the historical and the legendary: both were rewritten and edited, and both suffered from frequent interpolations even as late as the early centuries AD. Consequently, important as they are to the literary and religious tradition (both epics were later accepted as sacred literature, and their heroes were deified), it is difficult to identify them as historical source material. The central event of the *Mahābhārata,* whose geographical background is the upper Doab and adjoining areas, was a war between two groups of cousins—the Kauravas and the Pāṇḍavas. Though the traditional date for the war is 3102 BC, most historians would prefer a date between 1400 and 900 BC, and a few attempts to correlate the archaeological evidence with the epic actually point to *c.* 800 BC as more likely. The events of the *Rāmāyaṇa* relate to Ayodhyā, the middle Ganges Valley, and central India, with later interpolations extending the area as far southward as Ceylon.

Another group of texts that are taken as reference material for this period, although they were composed much later, are the *Purāṇa*s. The major *Purāṇa*s contain a section on royal genealogies, which, initially preserved as part of the oral tradition, claim to trace the descent of the two important royal lineages, the Sūryavaṃśa (Solar lineage) and the Candravaṃśa (Lunar lineage), and dynasties from the period of the Flood in the 4th millennium BC to the middle of the 1st millennium AD. Much of the material appears to be mythological, but the dynasties listed for the historical period (*c.* 600 BC–AD 500) do refer to known dynasties.

The geographical focus of the later Vedic period moves from the Sapta Sindhu region into the Ganges-Yamuna Doab and the territories on its fringe, the Ganges by then being the sacred river. The people beyond these areas were the *mlecchas,* the impure barbarians, unfamiliar with the speech and customs of the *ārya,* the noble ones. The land to the south of the Vindhyas, *dakṣināpatha* (literally, the "route to the south"), is referred to, but in a rather vague way.

Tribes of the regions

The literature is replete with the names of tribes occupying these regions. The most powerful tribe, commanding the greatest respect, was the Kuru-Pañcāla, which incorporated the two families of Kuru and Pūru (and the earlier Bharata tribe) and of which the Pañcāla was a confederation of lesser known tribes. This tribe occupied the upper Doab and the Kurukṣetra region. In the north the Kāmboja, Gandhāri, and Madra tribes were the dominant people. In the middle Ganges Valley the neighbours of the Kuru-Pañcālas were the Kāśī, Kosala, and Videha, who worked in close cooperation with each other and were rivals of the Kuru-Pañcālas. The Magadha, Aṅga, and Vaṅga peoples in the lower Ganges Valley and Delta were outside the Aryan pale and regarded as *mlecchas.* Magadha (Patna and Gayā districts of Bihār) is also associated with the *vrātya* people, who occupied an ambiguous position between the *ārya* and *mleccha.* Other tribes frequently mentioned are the Sātvants of the Chambal Valley and such tribes of the Vindhyan and northern Deccan region as the Andhras, Vidarbhas, Niṣadha, Pulinda, and Śabara. These were all *mleccha* tribes, and none was yet Aryanized. Because there was a consciousness of *mleccha* peoples and *mleccha* lands, there was an equally strong awareness of the pure land of the *ārya*s, called the Āryāvarta. This was virtually the Ganges-Yamuna Doab and the encircling areas. The location of all these tribes is of considerable historical interest because they gave their names to the geographical area. By the end of the period, tribal identity had changed gradually to territorial identity,

and the areas of settlement were now regarded as states, each with a recognized political system.

Political institutions

The *sabhā* and *pariṣad* (referring to an assembly, generally of a few persons of expertise) continued in general use and as political institutions into later periods. The larger assemblies declined. Rudimentary notions of taxation were the genesis of administration as also were the *ratnin*s (jewel bearers), consisting of officials and of representatives of various professions advising the king. A major transformation occurred in the notion of kingship, which ceased to be merely an office of a war leader; territorial identity provided it with power and status, symbolized by a series of lengthy and elaborate ceremonies—the *abhiṣekha,* generally followed by three *yajñ*as, *rājasūya, vājapeya,* and *aśvamedha.* The latter ceremony was a famous horse sacrifice, in which a specially selected horse was permitted to wander at will and was followed by a body of soldiers; the area through which the horse wandered unchallenged was claimed by the king conducting the sacrifice. Thus, theoretically at least, only kings of considerable power could perform this sacrifice. The assumption of these sacrifices was a permanent settling of a tribe in a particular area, marking the end of large-scale nomadism. This in turn led to the idea of the king owning the land, because he was permitted to give away portions of it as part of his charitable activities (*dāna*). The new trends emphasized the importance of the priests and the aristocracy (Brahmins and Kṣatriyas), who were the mainstay of kingship. The introduction, through royal sacrifices, of notions of divinity in kingship further strengthened the role of the priests. This was also the period in which kingship became hereditary.

Iron technology, or the use of *kṛṣṇa ayas* (dark metal), as it was apparently referred to in the later Vedic literature, and the migration into the Ganges Valley had resulted in the stabilization of agriculture and the formation of settlements of a permanent nature. Some of these settlements along the rivers evolved into towns, essentially as craft and commercial centres. By the middle of the 1st millennium BC, the second urbanization, this time in the Ganges Valley, had begun.

The development of caste

The factor with the most far-reaching consequences on Indian culture, and frequently referred to in literature, is the structure of society that has come to be called caste. In a late section of the Rigveda, a famous "Puruṣasūkta" hymn, containing a description of the primeval sacrifice, refers to the emergence of four groups from the body of the god Prajāpati—the Brahmins (Brāhmaṇs) from his mouth, the Kṣatriyas from his arms, the Vaiśyas from his thighs, and the Śūdras from his feet. This is clearly a mythologized attempt to describe the origin of the four *varṇa*s, which came to be regarded as the four major castes of Indian society. *Brahma,* meaning magical or divine knowledge, is the base of *Brāhmaṇ*—one who possesses such knowledge; *kṣatra,* he who is endowed with power or sovereignty, became the Kṣatriya; Vaiśya derived from *viś,* a settlement, and was thus a person settled on the land or a member of the *viś* or tribe; the derivation of Śūdra is not clear, which may suggest that it was a non-Aryan word. The texts continually repeat that it is the business of the Śūdra to serve the three higher castes. In addition to *varṇa* there are references to the term *jāti* (birth), which gradually came to acquire a closer connotation to caste and appears to mean the endogamous kinship group. In course of time the Brahmins became the preeminent priestly group, the intermediaries with the gods at the sacrificial rituals, and the recipients of large donations for priestly functions; and in the process they acquired a number of privileges, such as exemption from taxes and inviolability.

Claims to purity of birth added to their prestige and were only occasionally challenged by the theory that learning is the primary criterion of being a Brahmin. The Kṣatriyas, which appear to have been the landowning families, assumed the role of military leaders and of the natural aristocracy having connections with royalty. In the early sections of the texts, *rājanya* is sometimes used as an alternative term for the Kṣatriya. The Vaiśyas were more subservient, and, although their status was not as inferior as that of the Śūdras, they appear to have been crucial to the economy. The traditional view of the Śūdras is that they were non-Aryan cultivators, who came under the domination of the Aryans and in many cases were enslaved and therefore had to serve the upper three castes. But not all references to the Śūdras are to slaves. Sometimes rich Śūdras are mentioned, and in later centuries some of them even became kings. The traditional view that *varṇa* reflected the organization of Indian society has recently been questioned; it has been suggested that perhaps there is a distinction between the model of such a social structure and the actual manner of its functioning and that the concept of *jāti* is more central to caste functioning than *varṇa,* which may be the theoretical rationalization. The ingredients of caste may be listed as conformity to the rules of endogamy, rigid concepts of purity and pollution, the hierarchical ordering of occupations and an emphasis on their being hereditary, and an adherence to a geographical location. Because not all of the *jātis* could have been discussed in the legal codes, *varṇa* may have been used as a kind of model largely to indicate the process of social relationships. This view is strengthened by the fact that the non-Brahmanical literature of later periods does not always conform to the picture of caste society depicted in the *Dharma-śāstras.*

Other cultures of the period. This was broadly the picture in the Ganges Valley. In other parts of the subcontinent, there had evolved a series of cultures of the general Neolithic–Chalcolithic variety, concentrated mainly in the east and the south. Limited excavations in Assam suggest connections with parts of China and Southeast Asia. Sites in West Bengal, such as Pāṇdu-rājārdhibi (in Burdwān district) and Mahisadal (in Midnapore district), cover a range from pre-Iron Age into Iron Age and reveal black and red ware in the early stages, with the arrival of iron dating to about the 7th century BC. Traces of such cultures continued along the eastern coast, as at Patpad and Kesarpalli (Andhra Pradesh) and as far south as Pondicherry. A more concentrated group of southern sites around the Raichūr Doab, such as Sangankalla, Piklihāl, Maski, Tekkalkoṭā, and Hallur, marked the transition to metal technology, and late phases at these sites suggest a cattle-rearing people with gram and millet as their staple diet (unlike the eastern sites, which cultivated rice). Noticeable is the association with the northwestern Deccan and central India—the Jorwe-type sites, which suggest Harappan affinities. Perhaps the most interesting of these Neolithic–Chalcolithic sites are the ones (*c.* 1600–1200 BC) that carry traces of post-Harappan connections evident in the pottery and artifacts of such sites in Kāthiāwār as Rojdi, Rangpur, and Somnāth. The Banās Valley sites of Ahār and Gilund (on the outskirts of Udaipur), although less influenced by Harappan types, show evidence of contact with the western coastal sites as well as with those of Mālwa (Nāgda and Nāvadaṭoli). The Mālwa sites in turn have connections with those of the northern Deccan, such as the Jorwe-culture sites at Jorwe, Prakāsha, Bahāl, Nevāsa, and Daimabād, among others. Mālwa and the northwestern Deccan have traditionally been on the route from the Ganges Valley to the western coast and to the south; thus the presence of spouted vessels of the Jorwe kind in the sites of the Raichūr Doab is not surprising. It has also been suggested that the central Indian sites were in contact with Iran. Most of the sites provide evidence of iron in the period 1000–500 BC, which can therefore be regarded as a period of change. There is no conclusive evidence, however, to prove the hypothesis that the arrival of iron in these areas coincided with the arrival of the Indo-Aryan language.

Neolithic–Chalcolithic and Megalithic cultures

A more mature display of Iron Age culture during and subsequent to this period comes from the southern half of the peninsula—what has frequently been termed the Megalithic culture. The Megalithic culture remains problematical because neither its origin nor its evolution are yet clear. Some scholars have linked it with the megaliths of western India, others with those of central India, and many have suggested maritime connections with the eastern Mediterranean. Carbon-14 dates point to *c.* 1000 BC as the earliest level, and the discovery of Roman pottery in the upper levels at some sites provide a terminal date of

the 1st century AD. Burial methods vary in type from urn burials to menhirs and dolmens (stone monuments) and burial chambers marked off by a circle of stones, some of which are referred to in early Tamil sources. Some of the northern sites of this culture bear traces of contact with the Jorwe type. The Megalithic sites indicate a fairly sophisticated industry in iron artifacts and pottery, and the frequency of horse bits in the graves points to the use of horses. Curiously, however, it is precisely in the regions that used iron and were associated with the horse that the Indo-Aryan languages did not spread. Even today these are the regions of the Dravidian language group.

Indian Dynasties

dynasty	location	dates (in centuries)
Nanda	Magadha	4 BC
Maurya	India, barring the area south of Mysore (Karnātaka)	4–2
Indo-Greeks	north India	2–1
Śuṇga	Ganges valley and parts of central India	2–1
Śātavāhana	north Deccan	1 BC–AD 3
Śaka	western India	1 BC–AD 4
Kuṣāṇa	N India and C Asia	AD 1–3
Gupta	N India	4–6
Harṣa	N India	7
Pallava	Tamil Nādu	4–9
Cālukya	western and central Deccan	6–8
Pāla	Bengal	8–12
Pratihāra	W India and Upper Ganges Valley	9–11
Rāṣṭrakūṭa	western and central Deccan	8–10
Cōḷa	Tamil Nādu	9–13
Chandelā	Bundelkhand	10–12
Cauhān	Rājasthān	10–12
Caulukya	Gujarāt	10–13
Paramāra	western and central India	10–11
Later Cālukyas	western and central Deccan	10–12
Hoysaḷa	central and southern Deccan	12–14
Yādava	N Deccan	12–13
Pāṇḍya	Tamil Nādu	13–14

THE BEGINNING OF THE HISTORICAL PERIOD, C. 600–150 BC

For this phase of Indian history a variety of historical sources are available. The Buddhist canon, pertaining to the period of the Buddha (6th century BC) and later, is invaluable as a cross reference for the Brahmanical sources. This is also true, though to a more limited extent, of Jaina sources. In the 4th century BC there are secular writings on political economy and accounts of foreign travellers. The most important sources, however, are inscriptions of the 3rd century BC.

Pre-Mauryan states. Buddhist writings referring to the 6th and 5th centuries BC with a perspective on politics in eastern India mention 16 major states (*mahājanapada*) dominating the northern part of the subcontinent. A few of these, such as Gandhāra, Kāmboja, Kuru-Pañcāla, Matsya, Kāśī, and Kosala, continued from the earlier period and are mentioned in Vedic literature. The rest were new states, either freshly created from declining older ones or new areas coming into importance, such as Avanti, Aśvaka, Śūrasena, Vatsa, Cedi, Malla, Vṛjji, Magadha, and Aṅga. The mention of so many new states in the area of the eastern Ganges Valley was due in part to the eastern focus of the sources and was partly the antecedent to the increasing pre-eminence of the eastern regions.

Buddhist writings on eastern India

Location. Gandhāra lay astride the Indus and included the districts of Peshāwar and the lower Swāt and Kābul valleys. For a while its independence was terminated by its inclusion as one of the 22 satrapies of the Achaemenid Empire of Persia (*c.* 519 BC). Its major role as the channel of communication with Iran and Central Asia continued, as did its trade in woollen goods. Kāmboja adjoined Gandhāra. Originally regarded as a land of Aryan speakers, Kāmboja soon lost its important status, ostensibly because its people did not follow the sacred Brahmanical rites—a situation that was to prevail extensively in the north as the result of the introduction of new peoples and cultures through the processes of invasion, migration, and trade. Kāmboja was noted for its horses, which came from Central Asia and were traded to other parts of India.

The Kekayas had settled in the region between Gandhāra and the Beās River; they and the Madras and Uśīnaras were described as descendants of the Anu tribe. The Matsyas were located to the southwest of Delhi; but the Kuru-Pañcāla, still dominant in the Doab, were extending their control southward and eastward. The Kuru capital was reportedly moved from Hastināpura to Kauśāmbī when the former was devastated by a great flood, the evidence for which is clear in the excavations at Hastināpura and dates to about the 9th century BC. The Mallas were located in eastern Uttar Pradesh. Avanti arose in the Ujjain-Narmada Valley region, with its capital at Mahiṣmatī; during the reign of King Pradyota, there was a matrimonial alliance with the royal family at Kauśāmbī. Śūrasena had its capital at Mathura, and the tribe claimed descent from the Yadu clan. A reference to the Sourasenoi in later Greek writings is often identified with the Śūrasena, and the city of Methora, with Mathura. The Vatsa state was based on Kauśāmbī. The Cedi state (in Bundelkhand) lay on the route to the northwestern Deccan. South of the Vindhyas, on the Godāvari River, Aśvaka continued to thrive.

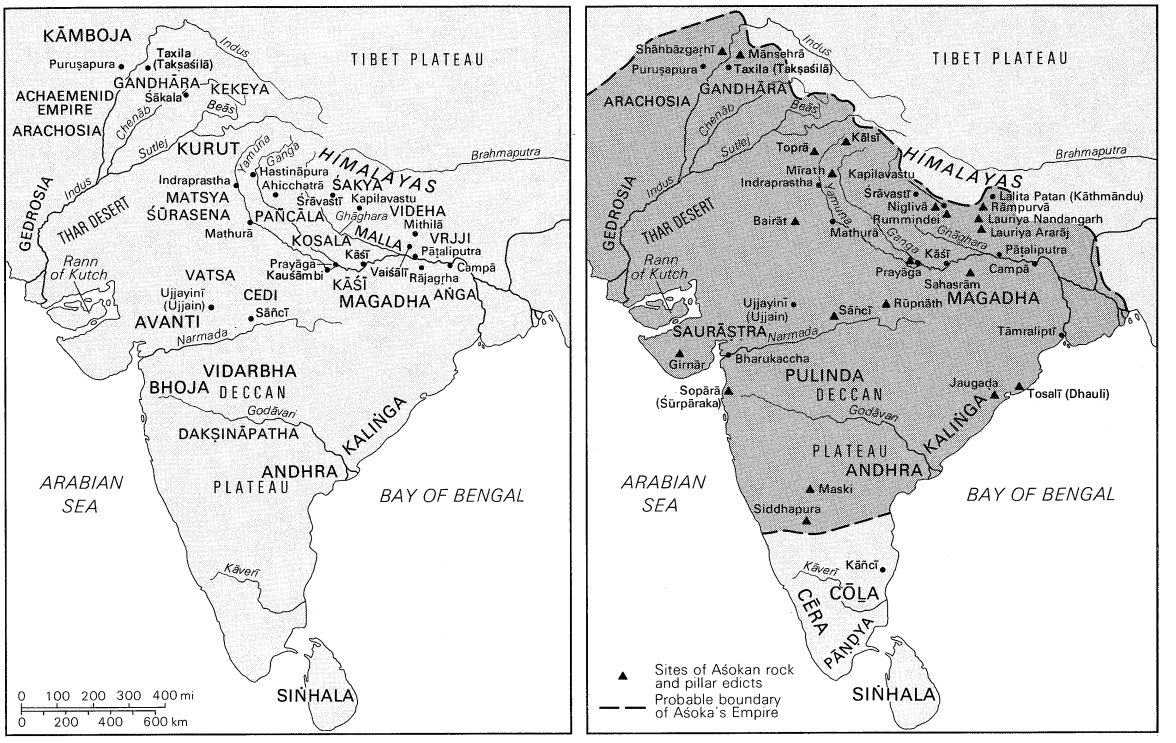
The mid-Ganges Valley was dominated by Kāśī and Kosala. Kāśī maintained close affiliations with its eastern neighbours, and its capital was later to acquire renown as the sacred city of Vārānasi (Benares). Kāśī and Kosala were continually at war over the control of the Ganges; in the course of the conflict, Kosala extended its frontiers to the north and the south, and ultimately it was constituted into Uttar and Dakṣina (northern and southern) Kosala. The new states of Magadha (Patna and Gayā districts) and Aṅga (north of the delta) were also interested in controlling the river and soon made their presence felt. The conflict was joined by the Vṛjji state (Besar and Muzaffarpur district). For a while, Videha (modern Tirhut), with its capital at Mithilā, also remained powerful. References to the states of the northern Deccan appear to repeat statements from sources of the earlier period, suggesting that there was little further knowledge of the region.

Political systems. The political system in these states was either monarchical or a type of representative government that variously has been called republican or oligarchical. Sometimes within the state itself there was a gradual change from one system to the other, as in the case of Vaiśālī, the nucleus of the Vṛjji state. Apart from the major states there also were many smaller republican states, such as those of the Koliyas, Moriyas, Jñātṛkas, Śākyas, and Licchavis. The Jñātṛkas and Śākyas are especially remembered because they were the tribes to which Mahāvīra and Gautama Buddha belonged; the Licchavis were eventually to become extremely powerful.

The republics

The republics were constituted by either a single tribe or a confederacy of tribes. The elected chief or the president (*gaṇapati, gaṇarāja*) functioned with the assistance of a council of elders probably selected from the Kṣatriya families. The most important institution was the general assembly, or *pariṣad,* to the meetings of which members were summoned by kettledrum. There was a prescribed seating arrangement for the members, an agenda, and an order of speaking and debate that terminated in a decision. A distinction was maintained between the families represented and the others. Sovereignty was vested in the *pariṣad,* which included in its powers the election of important functionaries. An occasional lapse into hereditary office on the part of the chief may account for the tendency toward monarchy. The danger of factions destroying the system was realized and was warned against.

The political system of these republics suggests a stabilized agrarian economy. There is frequent mention of the wealthy Kṣatriyas (*khhattiya-gahapati*) employing slaves and hired labourers to work on their lands. That there was an increase in urban settlements and trade is clear not only from references in the literary sources but also from the introduction of two characteristics of urban civilization—a script and coinage. Evidence for the script comes from references to letters of the alphabet in the literary sources and, in the 3rd century, from inscriptions, the latter indicating a considerable period of familiarity with a script. The most widely used script was Brāhmī, which is germane to most Indian scripts used subsequently. A vari-



(Left) India *c.* 500 BC and (right) Aśoka's empire at its greatest extent, *c.* 250 BC.
Adapted from C.C. Davies, *An Historical Atlas of the Indian Peninsula* (1959); Oxford University Press

ant during this period was Kharoṣṭī, used only in northwestern India and indicating an association with Aramaic from Iran. The most commonly spoken languages were Prākrit, which had its local variations in Śaurasenī (from which Pālī evolved), and Māgadhī, in which the Buddha preached. Sanskrit, the more cultured language as compared to Prākrit, was the language of the educated few and of the Vedic religion. Pāṇiṇi's grammar, the *Aṣṭādhyāyī,* and Yāska's etymological work, the *Nirukta,* suggest considerable sophistication in the development of Sanskrit.

Economy. Silver bent bar coins and silver and copper punch-marked coins came into use in the 5th century BC. It is not clear whether the coins were issued by a government authority or were the legal tender of moneyers. The gradual spread in the same period of a characteristic type of luxury ware, the northern black polished ware, is an indicator of expanding trade. The main trade routes were along the Ganges River and across the Indo-Gangetic watershed and the Punjab to Taxila and beyond, and alternatively from the Ganges Valley via Ujjain and the Narmada Valley to the western coast or southward to northwestern Deccan. The route to the Ganges Delta became more popular, and this permitted maritime contact with ports on the eastern coast of India. The expansion of trade and consequently of towns resulted in an increase in the number of artisans and merchants; they were organized in guilds (*śreṇi*), each of which tended to inhabit a particular part of a town. The guild system encouraged specialization of labour and the hereditary principle in professions, which became characteristic of caste functioning. Gradually some of the guilds acquired caste status. The practice of usury encouraged the activity of financiers, some of whom formed their own guilds and found that investment in trade proved increasingly lucrative.

Religion. These changing features of social and economic life were linked to religious and intellectual changes. Orthodox traditions maintained in certain sections of Vedic literature were questioned by teachers referred to in the *Upaniṣads* and *Āraṇyakas* and by others whose speculations and philosophy are recorded in other texts. There was a sizable heterodox tradition current in the 6th century BC, and speculation ranged from idealism to materialism. The Ājīvikas and the Cārvākas, among the smaller sects, were popular for a time as were the materialist theories of the Buddha's contemporary, Ajita Keśakambalin. Even though such sects did not sustain an independent religious tradition, the undercurrent of their teachings cropped up time and again in later religious trends.

Jainism and Buddhism

Of all these sects the only two that acquired the status of major religions were Jainism and Buddhism; the former remained within the Indian subcontinent; the latter spread to Central Asia, China, Japan, and Southeast Asia. Both religions were founded in the 6th century BC; Mahāvīra gave shape to earlier Nirgrantha ideas and formulated Jainism (the teachings of the Jina, or Conqueror, Mahāvīra), and the Buddha (the Enlightened One) preached a new doctrine.

There were many similarities among these two sects. Religious rituals were essentially congregational. Monastic orders (the *saṅgha*) were introduced with monasteries organized on democratic lines and initially accepting persons from all strata of life. Some of the monasteries developed into centres of education. The functioning of monks in society was greater, however, among the Buddhist orders. Wandering monks, preaching and seeking alms, gave the religions a missionary flavour. The recruitment of nuns signified a special concern for the status of women. Both religions questioned Brahmanical orthodoxy and the authority of the Vedas. Both were opposed to the sacrifice of animals, and both preached nonviolence. Both derived support in the main from the wealthy Kṣatriya landowners and the mercantile community; because trade and commerce did not involve killing, the principle of *ahiṃsā* (nonviolence) could be observed in these activities. The Jainas participated widely as the middlemen in financial transactions and in later centuries became the great financiers of western India. Both religions disapproved of the inequality of castes and, although not directly attacking the assumptions of a caste society, generally projected themselves as non-caste movements, thereby securing the support of low-caste groups. Because Buddhism borrowed its rituals and practices from popular local cults, women of the royal families became its supporters.

Struggle over the Ganges Valley

Magadhan ascendancy. The political focus in the 6th century BC centred on the control of the Ganges Valley. The states of Kāśi, Kośala, Magadha, and the Vṛjjis battled for this control for a century until Magadha emerged victorious. Magadha's success was due in part to the political ambition of its king, Bimbisāra, who came to the throne in *c.* 543 BC. According to some sources he was of the

Śiśunāga dynasty, though others maintain that the Śiśunāgas came later. At any rate, Bimbisāra conquered Aṅga, which gave him access to the Ganges Delta—a valuable asset in terms of the nascent maritime trade. Bimbisāra's son Ajātaśatru came to the throne in *c.* 491 BC and fully implemented his father's intentions within about 30 years. Ajātaśatru strengthened the defenses of the Magadhan capital, Rājagṛha, and built a small fort on the Ganges at Pāṭaligrāma, which was to become the famous capital of Pāṭaliputra (Patna). He then attacked and annexed Kāśī and Kośala. He still had to subdue the confederacy of the Vṛjji state, and this turned out to be a protracted affair lasting 16 years. Ultimately the Vṛjjis were overthrown, having been weakened by Ajātaśatru, who was able to sow dissension in the confederacy. The most important tribe in the confederacy was the Licchavi, which became historically quiescent for many centuries and then came to the fore again in the 4th century AD.

The success of Magadha was due not alone to the ambition of Bimbisāra and Ajātaśatru. Magadha had an excellent geographical location controlling the lower Ganges and thus drew revenue both from the fertile plain and the river trade. Access to the delta also brought in lucrative profits from the eastern coastal trade. Neighbouring forests provided timber for building and elephants for the army. Above all, rich deposits of iron ore gave Magadha a lead in technology.

Bimbisāra had been one of the earliest Indian kings to emphasize efficient administration, and the beginnings of an administrative system took root. Rudimentary notions of land revenue developed. Each village had a headman who was responsible for collecting taxes and another set of officials who supervised the collection and conveyed the revenue to the royal treasury. But the full understanding of the utilization of land revenue as a major source of state income was yet to come. The clearing of land continued apace, but the agrarian settlements were probably small because literary references to journeys from one town to another mention long stretches of forest paths.

Ajātaśatru died in *c.* 459 BC, and, after a series of ineffective rulers, a new dynasty was founded by Śiśunāga and lasted for about half a century until ousted by Mahāpadma Nanda. The Nandas are universally described as being of low origin, perhaps Śūdras. Despite these rapid dynastic changes, Magadha retained its position of strength. The Nandas continued the earlier policy of expansion. They are proverbially connected with wealth, probably because they realized the importance of regular collections of land revenue.

Invasions of Alexander the Great. The northwestern part of India had seen the campaign of Alexander of Macedon, who, in 327 BC, in pursuing his campaign to the eastern extremities of the Achaemenid Empire of Darius, entered Gandhāra. He campaigned successfully across the Punjab as far as the Beās River, where his troops refused to continue fighting. The vast army of the Nandas is referred to in Greek sources, and some historians have suggested that Alexander's Greek soldiers may have mutinied out of fear of this army. The campaign of Alexander made no impression, historically or politically, on India, nor are there any references to it in Indian sources. The most significant outcome of his campaign was that some of his Greek companions, such as Onesicritus, Aristobulus, and his admiral Nearchus, recorded their impressions of India. Strabo and Arrian, as well as Pliny and Plutarch, incorporated much of this material into their writings. But some of the accounts are fanciful, and make for better fiction than history. Alexander established a number of Greek settlements, which provided an impetus for the development of trade and communication with western Asia. Most valuable to historians was a reference to Alexander meeting the young prince Sandrocottos, a name identified in the 18th century as Candragupta, thus providing a chronological landmark in early Indian history.

The Mauryan Empire. The accession (dated to *c.* 325–*c.* 321 BC) of Candragupta Maurya is significant in Indian history because it inaugurated the first Indian empire. The Mauryan dynasty was to rule the entire subcontinent—except the area south of present-day Karnātaka and substantial parts of present-day Afghanistan—under a centralized, imperial system.

Candragupta Maurya. Candragupta overthrew the Nanda power in Magadha and then campaigned in central and northern India. Greek sources report that he engaged in a conflict in 305 BC in the trans-Indus region with Seleucus I Nicator, one of Alexander's generals who, on the death of Alexander, had founded the Seleucid dynasty in Iran. The campaign terminated with a treaty by which Seleucus ceded the trans-Indus provinces to the Maurya and the latter presented him with 500 elephants. A marriage alliance is mentioned, but no details are recorded.

Mauryan-Seleucid friendship

The treaty ushered in an era of friendly relations between the Mauryas and the Seleucids, with exchanges of envoys and gifts throughout the 3rd century BC. The most important of the envoys of Seleucus was Megasthenes, who left his observations in the form of a book, the *Indica.* Though the original has been lost, extensive quotations from it survive in the works of such later Greek writers as Strabo, Diodorus, and Arrian. A work of major importance in Sanskrit is the *Artha-śāstra* of Kauṭilya (or Cāṇakya, as he is sometimes called). Kauṭilya was prime minister to Candragupta Maurya, to whom this book, a treatise on political economy, is ascribed. In describing an ideal government, Kauṭilya indicates contemporary assumptions of political and economic theory, and the description of the functioning of government occasionally tallies with present-day knowledge of actual conditions derived from other sources. The date of origin of the *Artha-śāstra* remains problematical, with suggested dates ranging from the 4th century BC to the 4th century AD. Most authorities now agree that the kernel of the book was originally written during the early Mauryan period, but interpolations continued to be compiled perhaps until the Gupta period.

According to Jaina literary sources Candragupta became a Jaina toward the end of his reign. He abdicated in favour of his son Bindusāra, became an ascetic, and travelled with a group of Jaina monks to South India, where he died in the orthodox Jaina manner by deliberate slow starvation.

Bindusāra. The second Mauryan emperor was Bindusāra, who came to the throne in *c.* 297 BC. Greek sources refer to him as Amitrochates, the Greek for the Sanskrit *amitraghāta,* the "destroyer of foes." This name perhaps reflects a successful campaign in the Deccan, Candragupta having already conquered northern India. Bindusāra's campaign stopped at Mysore (modern Karnātaka), probably because the kingdoms of the extreme south, such as the Cōḷas, Pāṇḍyas, and Cēras, were well-disposed in their relations toward the Mauryas.

Aśoka and his successors. Bindusāra was succeeded by his son Aśoka, either directly in 272 BC or, after an interregnum of four years, in 268 BC (some historians say *c.* 265 BC). Aśoka's reign is comparatively well documented, particularly by epigraphical evidence. He issued a large number of edicts, which were inscribed in many parts of the empire and were composed in Prākrit, Greek, and Aramaic, depending on the language current in a particular region. The inscriptions composed in Greek and Aramaic are limited to Afghanistan and the trans-Indus region.

The first major event in Aśoka's reign, which he describes in one of his edicts, was a campaign against Kaliṅga in 260 BC. The suffering that resulted caused him to reevaluate the notion of conquest by violence, and gradually he was drawn to the Buddhist religion. About 12 years after his accession, he began issuing edicts at regular intervals. In one of them he referred to five Greek kings who were his neighbours and contemporaries—these were Antiochus II Theos of Syria, the grandson of Seleucus I Nicator; Ptolemy II Philadelphus of Egypt; Antigonus II Gonatas of Macedonia; Magas of Cyrene; and Alexander (either of Epirus or Corinth). He had sent envoys to all these kingdoms, and local tradition asserts that he had contacts with Khotān and Nepal. Close relations with Tissa, the king of Ceylon (Sri Lanka), were furthered by the fact that Mihinda, Aśoka's son (or his younger brother according to some sources), was the first Buddhist missionary on the island.

Aśoka ruled for 37 years. After his death a political decline set in, and half a century later the empire was

reduced to a part of the Ganges Valley alone. Tradition asserts that his son Kunāla ruled in Gandhāra. Epigraphic evidence indicates that his grandson Daśaratha ruled in Magadha. Some historians have suggested that his empire was bifurcated. In 185 BC the last of the Mauryas, Bṛhadratha, was assassinated by his Brahmin commander in chief Puṣyamitra, who founded the Śuṅga dynasty.

The Mauryan achievement

Financial base for the empire. The Mauryan achievement was its ability to weld the diverse parts of the subcontinent into a single political unit and to maintain an imperial system for almost 100 years. The financial base for an imperial system was provided by the income from land revenue and, to a lesser extent, from trade. The gradual expansion of the agrarian economy and improvements in the administrative machinery for collecting revenue increased the income from land revenue. This is confirmed by both the theories of Kauṭilya and the account of Megasthenes; Kauṭilya maintained that the state should organize the clearing of wasteland and settle it with villages of Śūdra cultivators. It is likely that some 150,000 persons deported from Kaliṅga by Aśoka after the campaign were settled in this manner. Megasthenes writes that there were no slaves in India, yet Indian sources speak of various categories of slaves, the most commonly used term being *dāsa-bhṛtaka* (slaves and hired labourers). It is likely that there was no large-scale slavery for production, although slaves were used on the land, in the mines, and in the guilds, along with the hired labour.

The nature of land revenue has been a subject of controversy—some scholars maintaining that the state was the sole owner of the land, and others contending that there was private and individual ownership as well. References to private ownership would seem to be too frequent to be ignored. There are also references to the crown lands, the personal property of the king. The problem relates to the rest of the land, where a clear distinction was not always maintained between ownership by the king and by the state. It may be argued that, because the distinction between the king as a symbol of the state and the state per se was rather blurred in political theory, the king collected revenue from the land, on behalf of the state rather than in his personal capacity. Two types of taxes were levied—one on the amount of land cultivated and the other on the produce of the land. State-controlled irrigation was known in limited periods and in limited areas. By and large irrigation systems were privately controlled by cultivators and landowners. There is no support for a thesis that control of the hydraulic machinery was crucial to the political control of the country.

Another source of income, which acquired increasing importance, was revenue from taxes levied on both internal and foreign trade. The attempt at uniform political administration helped to break the economic isolation of various regions. Roads built to ensure quick communication with the local administration inevitably became arteries of trade.

Mauryan society. According to Megasthenes, Mauryan society was divided into seven castes: philosophers, farmers, soldiers, herdsmen, artisans, magistrates, and councillors. He defined caste as endogamous and the professions as hereditary. The philosophers included a variety of priests, monks, and religious teachers; they formed the smallest caste in number but were the most respected and were exempt from taxation. The farmers were the largest in number. The soldiers were very well paid, and, if Pliny's figures for the army are correct—9,000 elephants, 30,000 cavalry, and 600,000 infantry—it must have required a considerable financial outlay. The mention of herdsmen as a socio-economic group suggests that, although the agrarian economy was expanding and had become central to the state income, pastoralism still continued to play an important economic role.

The artisans probably represented a major section of the urban population. The listing of magistrates and councillors as distinct and separate groups suggests that the personnel involved in administration was large and recognizable.

Mauryan government. The Mauryan government was a centralized bureaucracy, organized around the king. Aśoka saw his role as essentially paternal: "All men are my children." He was anxious to be in constant touch with public opinion, and to this end he travelled extensively throughout his empire and appointed a special category of officers to gauge public opinion. His edicts indicate frequent consultations with his ministers, the ministerial council being a largely advisory body. The offices of the *sannidhātṛ* (treasurer), who kept the account, and the *samāhartṛ* (chief collector), who was responsible for revenue records, formed the hub of the revenue administration. Each administrative department, with its superintendents and subordinate officials, acted as a link between local administration and the central government. Kauṭilya believed that a quarter of the total income should be reserved for the salaries of the officers. That the higher officials expected to be handsomely paid is clear from the salaries suggested by Kauṭilya and from the clerk's salary of 500 *paṇa*s and the minister's of 48,000 *paṇa*s. Public works and grants absorbed another large percentage of state income.

Divisions of the empire

The empire was divided into four provinces, and at the head of each was generally placed a prince of the royal family. Local officials were probably selected from among the local people because no method of impersonal recruitment to administrative office is mentioned. The emperor sent officers on inspection once every five years for an additional audit and check on provincial administration. Some categories of officers in the rural areas, such as the *rājjūkas,* combined judicial functions with assessment duties. Fines were the most general form of punishment, although capital punishment was imposed in extreme cases. Provinces were subdivided into districts and these again into smaller units. The village was the basic unit of administration and has remained so throughout the centuries. The headman continued to be an important official, as did the accountant and the tax collector (*sthānika* and *gopa*). For the larger units Kauṭilya suggests the maintenance of a census of the population as well. Megasthenes reports that there was a committee of 30 officials, divided into six subcommittees, who catered to the administration of Pāṭaliputra. The most important single official was the city superintendent (*nāgaraka*), who had virtual control of all aspects of city administration.

Aśoka's edicts. It was against this background of imperial administration and a changing socio-economic framework that Aśoka issued edicts that carried his message concerning the idea and practice of *dhamma,* the Prākrit form of the Sanskrit *dharma,* which is virtually impossible to translate. It carries a variety of meanings depending on the context, such as the universal law, the social order, piety, or righteousness; the Buddhists frequently used it with reference to the teachings of the Buddha. This, in part, coloured the earlier interpretation of Aśoka's use of the word to mean that he was propagating Buddhism. Until his inscriptions were deciphered in 1837, Aśoka was practically unknown except in the Buddhist chronicles of Sri Lanka—the *Mahāvaṃsa, Dīpavaṃsa,* and the *Vaṃsattapakāsinī*—and the works of the northern Buddhist tradition—the *Divyāvadana* and the *Aśokavadana*—where he is extolled as a Buddhist emperor *par excellence* whose sole ambition was the expansion of Buddhism. Most of these traditions were preserved outside India in Sri Lanka, Central Asia, and China. Even after the edicts were deciphered, it was believed that they corroborated the theory of the Buddhist sources because in some of the edicts Aśoka avowed his personal support of Buddhism. More recent analyses suggest, however, that, although he was personally a Buddhist, as his edicts addressed to the Buddhist *saṅgha* (order) attest, the majority of the edicts in which he attempted to define *dhamma* do not suggest that he was merely preaching Buddhism. These edicts addressed to the people were inscribed on rock surfaces or on especially erected and finely polished sandstone pillars, in places where people were likely to congregate.

It has been suggested that the idea of issuing edicts was borrowed from the Achaemenid emperors, especially from Darius; but the tone and content of Aśoka's edicts are quite different. The pillars with their animal capitals

have also been described as imitations of Achaemenid pillars, but there are substantial differences in style. The animal capitals are fine examples of imperial Mauryan art and contrast with numerous small, gray terra-cotta figures found at urban sites, which were clearly expressions of popular art (the official emblem of India since 1947 is based on the four-lion capital of the pillar at Sārnāth near Vārānasi).

Aśoka's ethic of behaviour

In his edicts Aśoka defines the main principles of *dhamma* as nonviolence, tolerance of all sects and opinions, obedience to parents, respect for the Brahmins and other religious teachers and priests, liberality toward friends, humane treatment of servants, and generosity toward all. It suggests a general ethic of behaviour to which no religious or social group could object. It could also act as a focus of loyalty to weld together the diverse strands that made up the empire. Interestingly, the Greek versions of these edicts translate *dhamma* as *eusebeia* (piety), and no mention is made anywhere of the teachings of the Buddha, which would be expected if Aśoka had been propagating Buddhism. His own activities under the impact of *dhamma* included attention to the welfare of his subjects, the building of roads and rest houses, the planting of medicinal herbs, the setting up of centres to cater to the sick, a ban on animal sacrifices, and the curtailing of killing animals for food. He also instituted a body of officials known as the *dhamma-mahāmattas*, who served the dual function of propagating the *dhamma* and keeping him in touch with public opinion. Some of these *dhamma-mahāmattas* were sent on missions to the neighbouring kingdoms.

Mauryan decline. Some historians maintain that the disintegration of the Mauryan Empire was an aftermath of Aśoka's policies and actions and that his pro-Buddhist policy caused a revolt among the Brahmins. The edicts do not support such a contention. It has also been said that his insistence on nonviolence resulted in the emasculation of the army, which was consequently unable to meet the threat of invaders from the northwest. There is, however, no indication in any source that Aśoka deliberately ignored the military wing of his administration despite his emphasis on nonviolence.

Other explanations for the decline of the empire appear more plausible. Among these is the idea that there may have been an economic pressure on the empire. It has been thought that the silver currency of the Mauryas was debased as a result of this pressure. The expense required for the army and the bureaucracy must have tied up a substantial part of the income. It is equally possible that the expansion of agriculture did not keep pace with the expansion of the empire, and, because many parts of the empire were nonagricultural, the revenue from the agrarian economy may not have been sufficient for the maintenance of the empire. It is extremely difficult to compute the population of the empire, but, on a purely impressionistic basis, approximately 50,000,000 can be suggested. For such a population of mixed agriculturalists and others to support an empire of this size would have been extremely difficult. Recent excavations at urban sites show economic standards distinctly higher than the Mauryan levels of these sites, probably based on a sudden increase in trade; but the income from trade was unlikely to have been sufficient to supplement fully the land revenue in financing the empire.

It has been argued that the Mauryan bureaucracy at the higher levels tended to be oppressive. This may have been true during the reigns of the first two emperors, from which the evidence is cited, but oppression is unlikely during Aśoka's reign because he was responsible for a considerable decentralization at the upper levels and for continual checks and inspections. A more fundamental weakness lay in the process of recruitment, which was probably arbitrary, with the hierarchy of officials locally recruited.

The concept of the state. Allegiance presupposes the existence of the concept of the state. A number of varying notions had evolved to explain the evolution of the state. Some theorists pursued the thread of the Vedic monarchies, in which the tribal chief became the king and was gradually invested with divinity. An alternative set of theories arising out of Buddhist and Jaina thought ignored the idea of divinity and assumed instead that, in the original state of nature, all needs were effortlessly provided but that slowly a decline set in and man became evil, developing desires, which led to the notions of private property and of family and finally to immoral behaviour. In this state of chaos, the people gathered together and decided to elect one among them (the *mahāsammata,* or "great elect") in whom they would invest authority to maintain law and order. Thus the state came into being. Post-Buddhist theorists retained the notion of a contract between a ruler and the people (most Brahmanical theorists held that the gods appointed the ruler and a contract of dues was concluded between the ruler and the people). Prevalent was the theory of *mātsya-nyāya*—that in periods of chaos, when there is no ruler, the strong devour the weak, just as in periods of drought big fish eat little fish. The legitimacy of the need for a ruler was made absolute. The existence of the state was primarily dependent on two factors: *daṇḍa* (authority) and *dharma* (the social order; *i.e.,* the preservation of the caste structure).

But the importance of the political notion of the state gradually began to fade, partly because of a decline of the political tradition of the republics and the proportional dominance of the monarchical system, in which loyalty was directed to the king. The emergence of the Mauryan Empire strengthened the political notion of monarchy. The second factor was that the *dharma,* in the sense of the social order, demanded a far greater loyalty than did the rather blurred idea of the state. The king's duty was to protect *dharma,* and, as long as the social order remained intact, anarchy would not prevail. Loyalty to the social order, which was a fundamental aspect of Indian civilization, largely accounts for the impressive continuity of the major social institutions over many centuries. But it also deflected loyalty from the political notion of the state, which might otherwise have permitted more frequent empires and a greater political consciousness. The re-emergence of an empire was to take many centuries.

Loyalty toward the social order

150 BC–AD 300

The disintegration of the Mauryan Empire gave rise to a number of small kingdoms, whose regional pattern was often to be repeated in subsequent centuries. The Punjab and Kashmir were drawn into the orbit of Central Asian politics. The lower Indus Valley became a passage for movements from the north to the west. The Ganges Valley played a largely passive role except when faced with campaigns from the northwest. In the northern Deccan there arose the first of many important kingdoms that were to play a key role as the bridge between the north and the south. Kaliṅga was once more independent. In the extreme south, the prestige and influence of the Cēra, Cōḷa, and Pāṇḍya kingdoms continued unabated. Yet in spite of political fragmentation, this was a period of economic prosperity, resulting partly from a new source of income—foreign trade. Indian trading interests were established in Central Asia, China, the eastern Mediterranean and Southeast Asia.

Rise of small kingdoms in the north. The Seleucids, involved in the politics of the eastern Mediterranean, allowed their control over northeastern Iran to relax, and Bactria and Parthia thus broke away. Diodotus, the Greek governor of Bactria, rose in rebellion against the Seleucid king Antiochus and declared his independence, which was recognized by Antiochus *c.* 250 BC.

Indo-Greek rulers. A later Bactrian king, Demetrius (reigned *c.* 190–*c.* 167 BC), took his armies into the Punjab and finally down the Indus Valley and gained control of northwestern India. This introduced what has come to be called Indo-Greek rule. The chronology of the Indo-Greek rulers is based largely on numismatic evidence. Their coins were, at the start, imitations of Greek issues, but by the 2nd century BC they had acquired a style of their own and were particularly noted for excellent portraiture. The legend was generally inscribed in Greek and Brāhmī, the latter being the script used for Prākrit in this case.

The best known of the Indo-Greek kings was Menander, known to Indian sources as Milinda. He is featured

in the Buddhist text, the *Milinda-pañha* ("Questions of Milinda"), written in the form of a dialogue between the King and the Buddhist philosopher Nāgasena, as a result of which the King is converted to Buddhism. Menander is known to have ruled during the years 155–130 BC. He controlled Gandhāra and Punjab, although his coins have been found farther south. According to one theory, he may have attacked the Śuṅgas in the Yamuna region and attempted to extend his control into the Ganges Valley, but, if he did so, he failed to annex it. Incidental references are made to the Yavanas in Madhyadeśa in the grammar *Mahābhāṣya* of Patañjali (*c.* 2nd century BC). Meanwhile, in Bactria itself, the descendants of the line of Eucratides, who had branched off from the original Bactrian line, now began to take an interest in Gandhāra and finally annexed Kābul and the kingdom of Takṣaśilā (Taxila). An important Prākrit inscription at Besnagar (Bhilsā district) of the late 2nd century BC, inscribed at the instance of a Greek, Heliodorus, records his devotion to the Vaiṣṇava Vāsudeva sect; he describes himself as the envoy of King Antialcidas of Taxila.

Asian rulers. The Bactrian control of Taxila was disturbed by an intrusion of the Scythians, known to Indian sources as the Śakas. They had attacked the kingdom of Bactria and subsequently moved into India. The determination of the Han (rulers of China) to keep the Central Asian nomadic tribes (the Hsiung-nu, Wu-sun, and Yüeh-chih) out of China forced these nomadic tribes in their search for fresh pastures to migrate southward and westward; a branch of the Yüeh-chih, the Ta Yüeh-chih, moved farthest west to the Aral Sea, and this displaced the existing Śakas, who poured into Bactria and Parthia. The Parthian king Mithridates II tried to hold them back, but after his death (88 BC) they swept through Parthia and continued into the Indus Valley; among the early Śaka kings was Maues, or Moga (1st century BC), who ruled over Gandhāra. The Śakas moved southward under pressure from the Pahlavas, who ruled briefly in northwestern India toward the end of the 1st century BC. At Mathura the Śaka rulers of note were Rājuvala and Śoḍāsa. Ultimately the Śakas settled in western India and Mālava and came into conflict with the kingdoms of the northern Deccan and the Ganges Valley, particularly during the reigns of Nahapāna, Caṣṭana, and Rudradāman in the first two centuries AD. Rudradāman's fame is recorded in a lengthy Sanskrit inscription at Junāgadh, dating to 150.

Intrusion of the Śakas

Kujūla Kadphises, the Yüeh-chih chief, conquered northern India in the 1st century AD. He was succeeded by his son Vīma, and after him came Kaniṣka, the most powerful among the Kushān kings, as they came to be called. The date of Kaniṣka is controversial, ranging from 78 to 248. The generally accepted view is 78, based largely on literary and epigraphical evidence from Indian sources and archaeological evidence and carbon-14 dating of finds from Soviet Central Asia. More recent numismatic evidence, however (particularly the similarity between Roman and Kushān coins), as well as evidence from Chinese and Iranian sources, points to a date as late as the 2nd century AD. The era based on 78, which is used in addition to the Gregorian calendar by the present Indian government and presumably was started by the Śakas, is sometimes explained by the assumed date of Kaniṣka's accession; the era was widely used in Mālava, Ujjain, Nepal, and Central Asia. The Kushān kingdom was essentially oriented to the north, with its capital at Puruṣapura (near modern Peshāwar), although it extended southward as far as Sāñcī and into the Ganges Valley as far as Vārānasi. Mathura was the most important city in the southern part of the kingdom. Kaniṣka's ambitions included control of Central Asia, which if not directly under the Kushāns did come under their influence. Kaniṣka's successors failed to maintain Kushān power. The southern areas were the first to break away, and, by the middle of the 3rd century, the Kushāns were left virtually with only Gandhara and Kashmir. By the end of the century, they were reduced to vassalage by the Sāsānian king of Persia.

Not surprisingly there were administrative and political terms in northern India at this time commonly found in western and Central Asia. The Persian term for the governor of a province, *khshathrapavan,* as used by the Achaemenids, was Hellenized into "satrap" and widely used by these dynasties. Its Sanskrit form was *kṣatrapa.* The governors of higher status came to be called *mahākṣatrapa;* they frequently issued inscriptions in whatever era they chose to follow, and they minted their own coins, indicating a more independent status than is generally associated with governors. Imperial titles were also taken by the Indo-Greeks, such as *basileus basileōn* ("king of kings"), similar to the Persian *shāhanshāh,* of which the later Sanskrit form was *mahārājādhirāja.* A title of Central Asian derivation was the *daivaputra* of the Kushāns, which is believed to have come originally from the "son of heaven" current in China.

Local tribal republics and kingdoms. The watershed between the Indus and Ganges valleys was the nucleus of a number of tribal republics whose local importance rose and fell in inverse proportion to the rise and fall of larger kingdoms. According to numismatic evidence, the most important politically were the Ārjunāyanas, Mālavas, Yaudheyas, Śibis, and the Ābhīra. The Ārjunāyanas had their base in the present Bharatpur-Alwar region. The Mālavas appear to have migrated from the Punjab to Jaipur (Rājasthān), perhaps after the Indo-Greek invasions; they are associated with the Mālava era, which has been identified with the Vikrama era, also known as the Kṛta era and dating to 58 BC. It is likely that southern Rājasthān as far as the Narmada River and the Ujjain district was named Mālwa after the Mālavas. Yaudheya evidence is scattered over many parts of the Punjab and the adjoining areas of Rājasthān and Uttar Pradesh, but during this period their stronghold appears to have been the Rohtak district, north of Delhi; the frequent use of the term *gaṇa* in Yaudheya coins points to an adherence to the tribal tradition. References to Śaivite deities, especially Kārttikeya or Skanda, the legendary son of Śiva, are striking. The Śibis also migrated from the Punjab to Rājasthān and settled at Mādhyamika (near Chittor). The Ābhīras lived in scattered settlements in various parts of western and central India as far as the Deccan. Most of these tribes claimed descent from the ancient lineages of the *Purāṇas,* and some of them were later connected with the rise of Rājput dynasties.

In addition to the tribal republics, there were small monarchical states, such as Ayodhyā, Kauśāmbī, and the scattered Nāga kingdoms, the most important of which was the one at Padmāvatī (Gwalior). Ahicchatrā (the Bareilly district of Uttar Pradesh) was ruled by kings who bore names ending in the suffix *-mitra.*

The Śuṅga kingdom. Magadha was the nucleus of the Śuṅga kingdom, which succeeded the Mauryan. The kingdom extended westward to include Ujjain and Vidiśā. The Śuṅgas came into conflict with Vidarbha and with the Yavanas, probably Bactrian Greeks attempting to move into the Ganges Valley. (The word *yavana* is derived from the Prākrit *yonā,* suggesting that the Ionians were the first Greeks to come into contact with Persia and India, presumably during the Achaemenid period. In later centuries Yavana was used for all peoples coming from western Asia and the Mediterranean, and it then included the Romans, Persians, and Arabs). The Śuṅga dynasty lasted for about one century and was then overthrown by the Brahmin minister Vāsudeva, who founded the Kāṇva dynasty, which lasted 45 years and was in turn overthrown by the Andhras, the ruling power of the Deccan. But the Andhras are not known to have ruled in Magadha, which area faded out of history until the 4th century AD.

Successors to the Maurya

Kaliṅga. Kaliṅga rose to importance in the 1st century BC under its king Khāravela. His date is controversial, but an earlier identification of the 3rd century BC would conflict with the fact that Kaliṅga was then a part of the Mauryan Empire. Khāravela perhaps exaggeratedly for a pious Jaina, successful campaigns in the western Deccan and against the Yavanas and Magadha and a triumphal victory over the Pāṇḍyas of South India.

The Andhras and their successors. The Andhras are listed among the tribal peoples in the Mauryan Empire. Possibly they rose to being local officials and then, on the

disintegration of the empire, gradually became independent rulers of the northwestern Deccan. It cannot be ascertained for certain whether they arose in the Andhra region (*i.e.,* the Krishna–Godāvari deltas) and moved up to the northwestern Deccan or whether their settling in the delta gave it their name. There is also controversy as to whether the dynasty became independent at the end of the 3rd century BC or at the end of the 1st century BC. Their alternative name, Sātavāhana, is presumed to be the family name, whereas Andhra was probably that of the tribe. It is likely that Sātavāhana power was established during the reign of Śātarkaṛṇi I, with the borders of the kingdom reaching across the northern Deccan; subsequent to this the Sātavāhanas suffered an eclipse when they were forced out of the northwestern Deccan by the Śakas in the 1st century AD, and they settled in Andhra. In the 2nd century AD the Sātavāhanas reestablished their power in the northwestern Deccan; among other things, coins of the Śaka Nahapāna found in this region are often overstruck by the name Gautamīputra Śātakaṛṇi. That he did not control Mālava and Ujjain is clear from the claim of the Śaka king Rudradāman to these regions. The last of the important Andhra kings was Yajñaśrī Śātakaṛṇi, who ruled at the end of the 2nd century AD and asserted his authority over the Śakas. The 3rd century saw the decline of Sātavāhana power as the kingdom broke into small pockets of control under various branches of the family.

The Sātavāhana feudatories then rose to power. The Ābhīras were the successors in the Nāsik area and were associated with the Kalacuri-Cedi era of 248–249; their inscriptions, like those of other successors to the Sātavāhanas, bear the strong imprint of the Sātavāhana style. The Ikṣvākus succeeded in the Krishna–Guntūr region. The Cūtū dynasty in Kuntala (southern Mahārāshtra) had had close connections with the Sātavāhanas. The Bodhis ruled briefly in the northwestern Deccan. The Bṛhatphalāyanas came to power at the end of the 3rd century in the Masulipatnam area. In these regions the Sātavāhana pattern of administration continued; most of the rulers had matronymics, which have been interpreted as a possible derivation from earlier matriarchal societies. Many of the royal inscriptions record donations made to Buddhist monks and monasteries, often by princesses, and also land grants to Brahmins and the performance of Vedic sacrifices by the rulers.

Southern Indian civilizations. According to legend the sage Agastya travelled southward and introduced both the Sanskrit language and Aryan culture to the far south early in the 1st millennium BC. His disciple composed the first Tamil grammar, the *Tolkāppiyam.* Significant historically attested contact between the north and the Tamil regions can be reasonably dated to the Mauryan period. The third Buddhist council, held in about 250 BC, resulted in Buddhist missionary activities in these regions. Expansion in trade brought in the markets of the southern kingdoms. An inscription of Khāravela refers to the powerful confederacy of the Tamil countries; the most important evidence, however, on the early history of the south are the epigraphs of the region, the Tamil *caṅkam* literature, and archaeological evidence.

Inscriptions in Brāhmī (recently read as Tamil Brāhmī) date to between the 2nd century BC and the 4th century AD. Most of the inscriptions record donations made by royalty or by merchants and artisans to Buddhist and Jaina monks. They are useful to check the evidence from the *caṅkam* literature, a collection of a large number of poems in classical Tamil, which, according to tradition, were recited at three assemblies of poets held at Madurai. Included in this literature are the so-called Eight Anthologies (*Ettutogai*) and Ten Idylls (*Pattuppāṭṭu*). It is said that the *Tolkāppiyam* is also of the same period. Though some scholars regard the *Siḷappadikaram (Śilappadigāram)* and the *Maṇimekhaḷai* as of this period, most maintain that they were written some centuries later. The literature probably belongs to the same period as the inscriptions, although some scholars suggest a much earlier date. The historical authenticity of sections of the *caṅkam* literature has recently been confirmed by archaeological evidence.

The *caṅkam* literature

Tamilākam, the abode of the Tamils, was defined in *caṅkam* literature as the area from Kanniyākumāri to the Tirupati Hill and from the east coast to the west. North of the Tirupati Hill the language changed. Tamilākam was divided into 13 *nāḍu*s (districts), of which the region of Madurai was the most important as the core of the Tamil speakers. The three major kingdoms of Tamilākam were those of the Pāṇḍyas (Madurai), the Cēras (Malabār Coast and the hinterland), and the Cōḷas (Thanjāvūr and the Kāverī Valley). The inscriptions of the Pāṇḍyas, recording royal grants and other grants made by local citizens, go back to the 2nd century BC. The king Neduñjeliyaṇ (early 3rd century AD) is celebrated by the poets of the *caṅkam* as the victor in campaigns against the Cēras and the Cōḷas. Cēra inscriptions of the 2nd century AD referring to the Irrumporai dynasty have been found near Karūr (Tiruchchirāpalli district), which has been identified with Korura, the inland capital of the Cēra kingdom according to Ptolemy. *Saṅgam* literature mentions the names of Cēra kings who have been dated to the 1st century AD. Among them Nedunjēral Ādan is said to have attacked the Yavana ships and held the Yavana traders to ransom. His son Śenguṭṭuvan, much eulogized in the poems, is also mentioned in the context of King Gajabāhu of Ceylon (Sri Lanka), who can be dated to either the first or last quarter of the 2nd century AD, depending on whether he was the earlier or the later Gajabāhu. Among the Irrumporai kings was Perunjēral, who is said to have annexed the neighbouring northern area (Salem district). Karikālaṇ (late 2nd century AD) is the best known of the early Cōḷa kings and was to become almost a kind of eponymous ancestor to many families of the south claiming Cōḷa descent. The early capital was at Uraiyūr, the nucleus of the Cōḷa kingdom, which stretched from the Vaigai River in the south to Toṇḍaimaṇḍalam in the north. The three major kingdoms were frequently at war; in addition there were often hostilities with Ceylon. Mention is also made of the ruler of Toṇḍaimaṇḍalam with its capital at Kāñchīpuram, which was to be the heart of Pallava power. There are also references to the minor chieftains, the Vel, who ruled small areas in many parts of the Tamil country. Some of them claimed ancestry from a person who sprang out of the sacrificial fire of a northern sage, a story which was to be repeated in connection with the ancestry of a group of later northern Indian dynasties. Ultimately all the kingdoms suffered at the hands of the Kalvār, or Kalabras, who came from the border to the north of Tamilākam and were described as evil rulers; but they were overthrown in the 5th century AD, with the rise of the Cālukyas (Chalukyas) and Pallavas.

Caṅkam literature reflects the indigenous cultural tradition as well as elements of the intrusion of the northern Aryan tradition, which by now was beginning to come into contact with these areas, many of which changed from tribal chieftainships to kingdoms. The predominant economy remained a pastoral-cum-agrarian economy, with an increasing emphasis on agriculture. In one of the poems in which reference is made to the different castes, the word *kudi,* literally meaning a "group," is used for caste. Each village had its *sabhā,* or council, which conducted local affairs and attended to the problems of the village. This tradition was to continue for centuries and became the base for the village councils of the later Cōḷa period. Religion concentrated in the main on worshipping a number of deities, preeminently Muruganṇ, and the ritual of worship in this case was sacrifice.

Trade between northern and southern India

The main economic momentum for the South Indian kingdoms was provided by trade with the Yavanas and with the northern parts of the subcontinent. Given the terrain of the peninsula and the agricultural technology of the time, large agrarian-based kingdoms like those of northern India were not feasible. Inevitably trade played more than a marginal role, and overseas trade became a major economic activity. Almost as soon as the Roman trade began to decline, the Southeast Asian trade was initiated; in subsequent centuries this became the focus of maritime interest.

Contacts with the West. Trade between western Asia and the western coast of India is often mentioned in

sources referring to the 1st millennium BC. Hebrew texts refer to the port of Ophir—identified with Sopārā, the Śūrpāraka of the west coast of India near Bombay, described as a market for gold, spices, precious stones, ivory, apes, and peacocks. Indian teak and cedar were used in Babylonian buildings in the 7th and 6th centuries BC. The Buddhist *Jātaka* literature mentions trade with Baveru, the Indian name for Babylon. After the decline of Babylon, Arab merchants from southern Arabia apparently continued the trade, probably supplying goods to Egypt and the eastern Mediterranean. The trade was then taken up by Greek and, to a larger extent, Roman traders. The discovery of the regular seasonal monsoon winds, enabling ships to drive a straight course across the Arabian Sea, made a considerable difference to shipping and navigation on the route from western Asia to India. Unification of the Mediterranean and western Asian world at the turn of the Christian Era under the Roman Empire brought Roman trade into close contact with India—overland with northern India and by sea with southern India. The emperor Augustus received two embassies from India in 25–21 BC, and these almost certainly were trade missions.

Trade with the Romans

The *Periplus Maris Erythraei,* written in the 1st century AD, lists a series of ports along the Indian coast including Muziris (Cranganore), Colchi (Koṛkai), Poduca, and Sopatma. Even more impressive is the evidence from archaeology and numismatics. An excavation at Arikamedu (near Pondicherry), believed to be the Poduca of the *Periplus* and Ptolemy, revealed a Roman trading settlement of the 1st and 2nd centuries AD; quantities of Roman pottery, beads, intaglios, lamps, glass, and coins point to a continuous occupation. It would seem that textiles were prepared to Roman specification and exported from such settlements. Pottery and graffiti in Brāhmī provide the local evidence.

Large hoards of Roman coins substantiate other evidence. Coins of the 1st century AD are found mainly around Coimbatore (Tamil Nādu), suggesting that initially the ships berthed along the western coast and that goods were transported overland to the eastern coast. The coins are mainly of the emperors Augustus (ruled 27 BC–AD 14), Tiberius (ruled 14–37), and Nero (ruled 54–68). Their frequency suggests that the Romans paid for the trade in gold coins, and the fact that many are overstruck with a bar may indicate that they were used as bullion in India; certainly, Pliny complained that the Indian luxury trade was depleting the Roman treasury. The coins are found most often in trading centres or near the sources of semiprecious stones, especially quartz and beryl. *Caṅkam* literature attests the prosperity of Yavana merchants trading in towns such as Kāvēripaṭṭinam (in the Kāverī Delta). The *Periplus* lists the major exports of India as pepper, pearls, ivory, silk, spikenard (an aromatic plant), precious stones, malabathrum, tortoise shells, and textiles. For these the Romans traded glass, copper, tin, lead, realgar, orpiment, antimony, and wine, or else they paid in gold coins. The balance of trade appears to have been in favour of India.

Legend has it that St. Thomas travelled from western Asia to Malabār in AD 52. He is believed to have established a number of Syrian churches, which would perhaps account for Syrian Christianity being the major form of Christianity until the arrival of the Portuguese in India in the 15th century. Historical evidence of a Christian community in India so far dates only to the 7th century AD.

Maritime trade routes

The maritime trade routes from the Indian ports were primarily to the Persian Gulf and the Red Sea from where they went overland to the eastern Mediterranean and to Egypt. Indian merchants also ventured out to Southeast Asia seeking spices and semiprecious stones. Initially merchants from western India, Kaliṅga, and South India settled in small trading communities in Burma, Malaya, Cambodia, Sumatra, Java, and Bali; gradually merchants from other coastal areas also began to participate in this trade. The river valleys and the Mauryan roads were the chief routes within India. Greek sources refer to a royal highway built by the Mauryas, connecting Taxila with Pāṭaliputra and terminating at Tāmraliptī, the main port in the Ganges Delta. On the western coast the major port of Bhṛgukaccha (modern Broach) was connected with the Ganges Valley via Rājasthān or, alternatively, Ujjain. From the Narmada Valley there were routes going into the northwestern Deccan and continuing along rivers flowing eastward to various parts of the peninsula. Goods were transported mainly in caravans of oxen and donkeys, but only in the hot and cool seasons, the rains creating an impossible time for travel. Coastal and river shipping, used where possible, was clearly cheaper than land routes. In the north the route from Taxila went to Kābul and Qandahār and from there branched off in various directions, mainly linking up with routes across Persia to the Black Sea ports and the eastern Mediterranean. The route connecting China with Bactria via Central Asia, soon to become famous as the Old Silk Route, linked the oases of Kashgar, Yarkand, Khotan, Miran, Kucha, Karashahr, and Turfan, in all of which Indian merchants were to establish trading stations. The Central Asian route brought Chinese goods in large quantities into the Indian and west Asian markets.

Society and culture. The commercial economy played a central role during this period. Agrarian expansion was not arrested, and land revenue continued to be a major source of income, but profit from trade made a substantial difference to the economy. Not surprisingly there was a marked improvement in the urban standard of living.

Guilds. Related to commercial activity was the development of the social institution most closely connected with it—the *śreṇi* (guild), through which trade was channelled. The guilds were registered with the town authority, and the behaviour of guild members was regulated by the *śreṇi-dharma* (rules of conduct of the guild). The more wealthy guilds employed slaves and hired labourers in addition to their own artisans, though the percentage of such slaves appears to have been small. Guilds had their own seals and insignia. They often made lavish donations to Buddhist and Jaina monasteries, and some of the finest Buddhist monuments of the period were the result of such donations. In some areas, such as the Deccan, members of the royal family invested money with a particular guild, and the accruing interest became a regular donation to the Buddhist *saṅgha.* This must also incidentally have increased the political importance of the guild.

Finance. With the development of the money economy, the role of the financier and banker became extremely important. Sometimes this role was performed by the richer guilds, but the more regular source of money was the merchant financier (*śreṣṭhin*). There was a large variety of coins, minted with a high degree of professional skill in the various kingdoms. The most commonly used were the gold *dīnāras* and *suvarṇas,* based on the Roman denarius (124 grains); a range of silver coins, such as the earlier *kārṣāpaṇa* or *paṇa* (57.8 grains) and the *śatamāna;* an even wider range of copper coins, such as the *kārṣāpaṇa* (144 grains), *māṣa* (nine grains), *kākiṇī* (2.25 grains), and a variety of unspecified standards; and other coins, issued in lead and potin, particularly in western India. Usury was a regular part of the banker's trade, with 15 percent being the generally accepted interest rate, although this varied according to the enterprise for which the money was borrowed. Increasing trade also led to the use of a great variety of weights and measures.

Impact of trade. The economic impact of foreign trade was probably greatest in the south, but the interchange of ideas appears to have been more substantial in the north. This may have been due to a longer association with western Asia and the colonial Hellenic culture exported from these lands. Greek together with Aramaic was widely spoken in Afghanistan and was doubtless understood in Taxila. The spurt of geographical studies in the Mediterranean resulted in works with extensive descriptions of the trade with India; these included Strabo's *Geography,* Ptolemy's *Geography,* the *Periplus Maris Erythraei,* and Pliny's *Natural History.* The most obvious and visible impact was in Gandhāra art, which depicted Indian themes in the Greco-Roman style of Alexandrian art—an attractive hybrid that accelerated the frequency of the icon in Buddhist art.

Religious patronage

Religion. If art remains are an index to patronage, then Buddhism seems to have been the most favoured religion. Buddhist centres generally had a complex of three structures—the monastery (*vihāra*), the hall of worship (*caitya*), and the sacred tumulus (*stūpa*)—all of which were freestanding structures in the north but were initially rock-cut monuments in the Deccan. The Jainas found more patrons in the Deccan. Hindu temples are referred to in literature, but none has been actually found dating to this period. Apart from the Gandhāra style of sculpture, a number of indigenous centres in other parts of India, such as Mathura, Kārlī, Nāgārjunakoṇḍa, and Amarāvati, portrayed Buddhist legends in a variety of local stone. The more popular medium was terra-cotta, by then changed from gray to red, depicting not only men and women in the ordinary walks of life and animal figures but also large numbers of mother-goddesses, indicating the continued and popular worship of these deities.

The practice of Buddhism was itself undergoing change. Affluent patronage provided the big monasteries with endowments of land and slaves to work the land. Association with royalty gave Buddhism access to power. The proselytizing consciousness that had gradually evolved now saw Buddhist monks as missionaries travelling to Central Asia and China, western Asia, and Southeast Asia. New situations inevitably led to the need for new ideas, as is most clearly seen in the contact of Buddhism with Christianity and Zoroastrianism in Central Asia. Certain aspects of the legend of the lives of the Buddha and Christ are strikingly similar (*e.g.,* a miraculous birth, the fasting in the wilderness, the temptation by the devil). The practices of the Essenes in Roman Judaea also reflect some knowledge of Indian belief and life. Arguments over the original teaching of the Buddha had already resulted in a series of councils called to clarify the doctrine. The two main sects were the Theravāda, with its centre at Kauśāmbī, which compiled the Pāli canon on Buddhist teachings, and the Sarvāstivāda, which arose at Mathura, spread northward, and finally established itself in Central Asia, using Sanskrit as the language for preserving the Buddhist tradition. A fourth council, held in Kashmir during the reign of Kaniṣka, ratified the separation of the two main schools of Buddhism—the Mahāyāna (Greater Vehicle) and the Hīnayāna (Lesser Vehicle).

Jainism had by now also split into two groups—the Digambara (Sky-Clad; *i.e.,* naked), the more orthodox, and the Śvetāmbara (White-Clad), the more liberal. The Jainas were not as widespread as the Buddhists, their main centres being in western India, Kaliṅga for a brief period, and the Mysore (modern Karnātaka) and Tamil country.

Brahmanism also underwent changes with the gradual fading out of some of the Vedic deities. The two major gods were Vishnu (Viṣṇu) and Śiva, around whom there emerged a monotheistic trend perhaps best expressed in the *Bhagavadgītā,* which most authorities would date to the 1st century BC. The doctrine of *karma* and rebirth, emphasizing the influence of actions performed either in this life or in former births on the present and future lives, became central to Hindu belief and influenced both religious and social notions. Vedic sacrifices were not discontinued but gradually became symbols of such ceremonial occasions as royal consecrations. Sacrificial ritual was beginning to be replaced by the practice of *bhakti* (personal devotion), positing a personal relationship between the individual and the deity.

Sacred literature

Literature. Popular epics, such as the *Mahābhārata* and the *Rāmāyaṇa,* were injected with didactic sections on religion and morality and given the status of sacred literature. Their heroes Krishna (Kṛṣṇa) and Rāma were incorporated into Vaiṣṇavism as avatars (*avatāras*; incarnations) of Vishnu. The concept of incarnations was useful in incorporating local deities and cults.

The epics were also used as a treasury of stories, which were converted into poems and plays. The dramas of Bhāsa were among these. His two best known plays, *Svapnavāsavadatta* and *Pratijñāyaugandharāyaṇa,* formed parts of the foundation of Sanskrit drama. Aśvaghoṣa, another major Sanskrit dramatist, based his works on Buddhist themes. The popularity of drama necessitated the writing of a work on dramaturgy, the *Nāṭya-śāstra* ("Treatise on Dramatic Art") of Bharata. The composition of *Dharma-śāstras* (sacred law books), among which the most often quoted is ascribed to Manu, became important in a period of social flux in which traditional social law and usage were important as precedent. A new grammar of Sanskrit had been provided by the *Mahābhāṣya* of Patañjali, timely because even the non-Indian dynasties of the north and west made extensive use of Sanskrit. The two sciences that received attention were astronomy and medicine, both of which indicate an interchange of ideas with western Asia. Doubtless, the new navigation along midocean routes was a considerable incentive to improved knowledge in astronomy. Two basic medical treatises, composed by Caraka and Suśruta, date to this period.

Assimilation of foreigners. The presence of foreigners, most of whom settled in Indian cities and took on Indian habits and behaviour in addition to religion, became a problem for social theorists because the newcomers had to be fitted into caste society. It was easier to accommodate a group rather than an individual into the social hierarchy, because the group could be given a *jāti* status. Technically, conversion to Hinduism was difficult because one had to be born into a particular caste, and it was *karma* that determined one's caste. The theoretical definition of caste society continued as before, and the four *varṇa*s were referred to as the units of society. The problem of including new groups was more acute in the two upper castes. The assimilation of local cults, for instance, demanded the assimilation of cult priests, who had to be accommodated within the Brahmanical hierarchy. The Greeks and the Śakas, clearly of non-Indian origin, who were initially the ruling group were referred to as "fallen Kṣatriyas." The Vaiśya and Śūdra groups did not pose such a serious problem because they were vaguely defined and consequently socially more mobile. It is likely that in such periods of social change some lower caste groups may have moved up the ladder of social hierarchy.

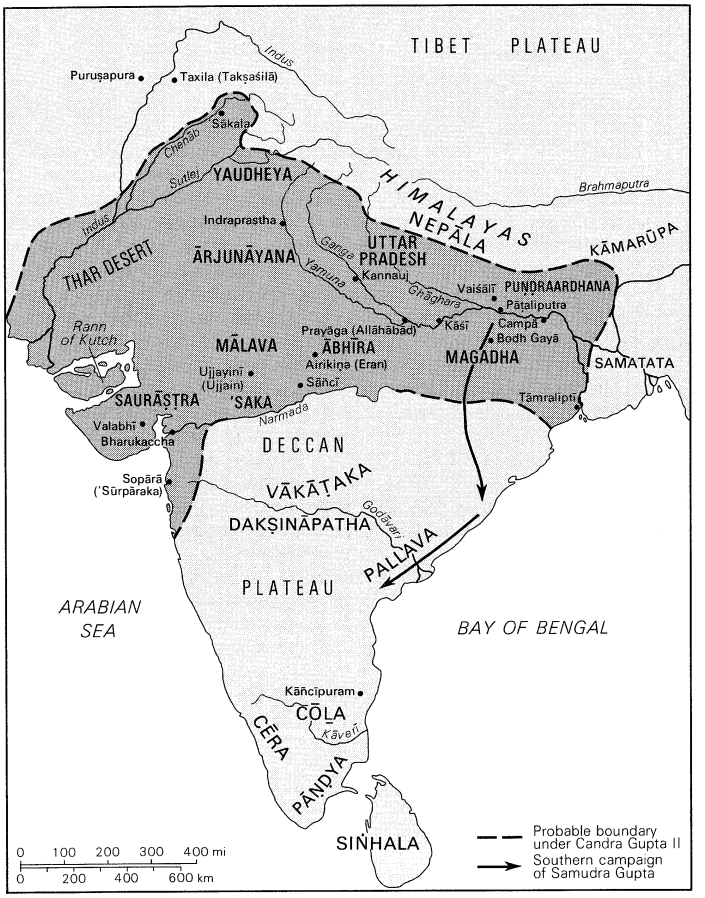
AD 300–750

Northern India. *The Guptas.* The Gupta period (*c.* 320–540) has often been called the Classical Age of India. It was a period during which the norms of Indian literature, art, architecture, and philosophy were established, at least in some parts of India, in the form that came to be regarded as standard. It was an age of material prosperity in most areas, particularly among the urban elite, and it was a period of renascent Hinduism. The Gupta period has also been called an imperial age, but with less justification, since administrative centralization so characteristic of an imperial system is less apparent than during the Mauryan period.

The Guptas, a comparatively unknown family, came from either Magadha or eastern Uttar Pradesh. The third king, Candra Gupta I, took the title of *mahārājādhirāja.* He married a Licchavi princess—an event celebrated in a series of gold coins. It has been suggested that, if the Guptas ruled in Prayāga (modern Allāhābād in eastern Uttar Pradesh), the marriage alliance may have added Magadha to their domain, assuming that the Licchavis had no heir. The Gupta Era began in 320, but it is not clear whether this date commemorated the accession of Candra Gupta or the assumption of the title of independence.

Campaigns of Samudra Gupta

Candra Gupta appointed his son Samudra Gupta to succeed him *c.* 330, according to a long eulogy (*praśasti*) to Samudra Gupta inscribed on a pillar at Prayāga (modern Allhābād). The coins of an obscure prince, Kācha, suggest that there may have been contenders for the throne but that it was Samudra Gupta who was successful. Samudra Gupta's campaigns took him in various directions and resulted in many conquests. Not all the conquered regions were annexed, but the range of operations established the military prowess of the Guptas. According to the Prayāga inscription, eight rulers of Āryāvarta in northern India were "violently uprooted." Samudra Gupta also conquered the states of Dakṣinapatha but reinstated the conquered rulers. The route of this campaign was along the east coast as far south as Kāncipuram (Tamil Nādu state).



The Gupta Empire at the end of the 4th century.

Adapted from C.C. Davies, *An Historical Atlas of the Indian Peninsula* (1959); Oxford University Press

Mention is also made of forest chiefs who were reduced to servitude. Five border states in the east, in addition to Nepal and the north, acknowledged the Gupta king. Nine tribal republics including the Mālavas, Ārjunāyanas, Yaudheyas, and Ābhīras are listed as having become subservient. Another category included more distant rulers who were regarded as subordinate: the Daivaputra-shāhishāhanushāhi of the northwest, Śaka, Muruṇḍa, and the inhabitants of "all the other islands," including Siṁhala (Sri Lanka). It would seem that the campaign extended Gupta power in northern and eastern India and virtually eliminated the republics and the minor kings of central India and the Ganges Valley. The more distant regions may not have been subordinate, but they were included largely because the composition was a eulogy. The islands remain problematical, since they could have been the ones close to India or else those of Southeast Asia with which communication was now close. It would appear that the conquest of the Ganges Valley and central India was the extension of the kingdom and remained the area under direct administrative control. The campaign in the eastern coastal areas may have been prompted by the desire to acquire the trading wealth of these regions. The grim image of Samudra Gupta as a military conqueror is ameliorated, however, by references to his love of poetry and music and by coins on which he is depicted playing the lyre.

Samudra Gupta was succeeded *c.* 380 by his son Candra Gupta II, though there is some evidence that there may have been an intermediate ruler. Candra Gupta II's major campaign was against the Śaka rulers of Ujjain, the success of which was celebrated in a series of silver coins. Gupta interest lay not merely in the political control of the west but in the wealth the area derived from trade with western and southeastern Asia. Gupta territory adjoining the northern Deccan was secured through a marriage alliance with the Vākāṭaka dynasty, the successors of the Sātavāhanas in the area. Though Candra Gupta II took the title of Vikramāditya (Sun of Valour), his reign is associated more with cultural and intellectual achievements than it is with campaigns. It was also during his reign that Fahsien, a Chinese Buddhist, travelled to India and left an account of his impressions.

There is considerable epigraphic evidence relating to the administration of the Gupta kingdom. The kingdom was divided into provinces called *deśa* or *bhukti,* and these in turn into smaller units, the *pradeśa* or *viṣaya.* The provinces were governed by *kumārāmātyas,* high imperial officers or members of the royal family. That there was decentralization is clear from the composition of the municipal board (*adhiṣṭhāna-adhikaraṇa*), which consisted of the guild president (*nagara-śreṣṭhin*), the chief merchant (*sārthavāha*), and representatives of the artisans and of the scribes. During this period the term *sāmanta,* which originally meant neighbour, was beginning to be applied to landed intermediaries who had been given grants of land or to conquered feudatory rulers. There was also a noticeable tendency for some of the higher administrative offices to become hereditary.

The Hūṇa invasion

The first hint of a fresh invasion from the northwest comes in the reign of Candra Gupta's son and successor, Kumāra Gupta (*c.* 415–455). The threat was that of the Huns, or Hūṇas as they are called in Indian sources. Skanda Gupta (*c.* 455–467), who succeeded Kumāra Gupta, and his successors all had to face the full-fledged invasion of the Hūṇas. Skanda Gupta managed to rally Gupta strength for a while, but after his death the situation deteriorated. Dissensions within the royal family added to the problem. Gupta genealogies of this period show considerable variance in their succession lists. By the mid-6th century, when the dynasty apparently came to an end, the kingdom had dwindled to a small size. Northern India and parts of central India were in the hands of the Hūṇas.

The Hūṇas were a branch of the Ephthalites, or White Huns, one section of which migrated to the Volga Valley and another to the Oxus. The latter invaded Persia and Afghanistan, and from there a group swept into northern India. The first Hūṇa king in India was Toramāṇa (early 6th century), whose inscriptions have been found as far south as Eran (Madhya Pradesh). His son Mihirakula, a patron of Śaivism, is recorded in Buddhist tradition as uncouth and extremely cruel. The Gupta rulers, together with Yaśodharman of Mālava, seem to have confronted Mihirakula and forced him back to the north. Ultimately his kingdom was limited to Kashmir and Punjab with the capital at Śākala (Siālkot). Hūṇa power declined after his reign.

The coming of the Hūṇas brought northern India once more into close contact with Central Asia, and a number of Central Asian tribes migrated into India. It has been suggested that the Gurjaras, who gradually spread to Kashmir, Punjab, Rājasthān, Gujarāt, and northern Mahārāshtra, may be identified with the Khazars of Central Asia. The Hūṇa invasion challenged the stability of the Gupta kingdom, even though the ultimate decline may have been caused by internal factors. A severe blow to the Gupta kingdom was the disruption of the Central Asian trade caused by the Hun movement and the consequent disruption of the income that northern India had derived from it. Some of the tribes of northern India migrated to other regions, and this movement of peoples affected changes in the social structure of the post-Gupta period. The rise of Rājput families and "Kṣatriya" dynasties (see below *The Rājputs*) are associated by some scholars with tribal chiefs in these new areas.

Successor states. Of the kingdoms that arose as inheritors of the Gupta territory, the most important were those of Valabhī (Saurāṣṭra and Kāthiāwār); Gujarāta (originally the area near Jodhpur), believed to be the nucleus of the later Pratihāra kingdom; Nandipurī (near Broach); Maukhari (Magadha); the kingdom of the later Guptas (in the area between Mālava and Magadha) and those of Bengal, Nepal, and Kāmarūpa. In Nepal, effective political power passed into the hands of Aṃśuvarman, an erstwhile feudatory chief—a pattern by no means peculiar to Nepal. In the Assam Valley the royal family of Kāmarūpa became independent of Gupta sovereignty. Orissa (Koṇgoda) was

under the Māṇa and Śailodbhava dynasties before being conquered by Śaśāṅka, king of Gauḍa (lower Bengal), who in the early 7th century annexed a substantial part of the Ganges Valley, where he came into conflict with the Maukharis and the rising Puṣpabhūti dynasty of Thānesar (north of Delhi).

The reign of Harṣa

The Puṣpabhūti dynasty aspired to imperial status during the reign of Harṣa (Harṣavardhana). Sthāṇvīśvara (Thānesar) appears to have been a small principality, probably under the suzerainty of the Guptas. Harṣa came to the throne, in Thānesar, in 606 and ruled there for 41 years. The first of the major historical biographies in Sanskrit, the *Harṣacarita* ("Deeds of Harṣa"), was written by Bāṇa, a celebrated author attached to his court, and contains information on Harṣa's early life. A full account of the period also survives, written by the Chinese Buddhist pilgrim Hsüan-tsang, who travelled through India and stayed for some time at a monastery at Nālandā. Harṣa acquired Kannauj (Farrukhābād district) acquired Kannauj and thereafter a large kingdom. A major but unsuccessful campaign was fought against Pulakeśin II, the Cālukya king of the northern Deccan, and Harṣa was confined to the northern half of the subcontinent. Nor was his success spectacular in western India against Valabhī, Nandipurī, and Sind (lower Indus Valley). In his eastern campaign, Harṣa met with little resistance (Śaśāṅka having died in 636), however, and acquired Magadha, Vaṅga and Koṅgoda (Orissa). Although Harṣa failed to build an empire, his kingdom was of no mean size, and he earned the reputation of being the preeminent ruler of the north. Harṣa is remembered as the author of three Sanskrit plays—the *Ratnāvalī* and *Priyadarśikā* and the *Nāgānanda*—the theme of the last indicating his interest in Buddhist thought. The T'ang emperor of China, T'ai Tsung, sent a series of embassies to Harṣa. One of these arrived after Harṣa's death and found that the throne had been usurped. Finally the allies of Harṣa defeated the usurper, who, it is claimed, was taken to China as a prisoner. The kingdom of Kannauj entered a period of decline until in the early 8th century, when it revived with the rise of Yaśovarman, who is eulogized in the Prākrit poem *Gauḍa-vadha* ("The Slaying of the King of Gauḍa") of Vākpati. Yaśovarman came into conflict with Lalitāditya, the king of Kashmir of the Karkoṭa dynasty, and appears to have been defeated.

In the 8th century the rising power in western India was that of the Gurjara-Pratihāra. The Rājput dynasty of the Guhilla had its centre in Mēwar (with Chitor as its base) and arose on the defeat of the Mori Rājputs, who probably then moved to the Kotah area. The Cāpa family were associated with the city of Aṇahilapāṭaka (modern Pātan) and are involved in early Rājput history.

The Arab invasion

The late 7th and early 8th centuries saw a new threat—this time in Sind—the invasion of the Arabs. Inscriptions of the western Indian dynasties speak of controlling the tide of the *mleccha,* which has been interpreted to mean the Arabs. The conquest of Sind was the easternmost extremity of the Arab Empire. A local chronicle from Sind, the *Chach-nāma,* gives an account of these events. The Arabs initially sent a small naval expedition, which was a failure. An overland expedition attacked Kābul, Zābul (the Helmand Valley), and Sind and was finally successful. The Arab hold on Sind was loose at first, and the local chiefs remained virtually independent. By 724, however, the local chiefs were deposed, and direct rule was established with a governor ruling on behalf of the caliph. At this point Arab expansion into western India, as well as their advance into Punjab and Kashmir, was checked. The Arabs are referred to in Indian records either as *mleccha* or as *yavana.* The Indians had only a limited interest in political events in western Asia, and therefore did not fully comprehend the seriousness of the Arab threat. Along the west coast, the Arabs were seen as yet another group of traders from western Asia. The possible competition with Indian trade was not realized.

The Deccan. In the Deccan, the Vākāṭaka dynasty was closely tied to the Guptas. With a nucleus in Vidarbha, the founder of the dynasty, Vindhyaśakti, extended his power northward as far as Vidiśā (near Ujjain). At the end of the 4th century, a collateral line of the Vākāṭakas was established by Sarvasena in Vatsagulma (Bāsim, in Akola district), and the northern line helped the southern to conquer Kuntala (southern Mahārāshtra). That the main Vākāṭaka line dominated the northern Deccan during this period is clearly established by the matrimonial alliances not only with the Guptas but also with other peninsular dynasties such as the Viṣṇukuṇḍins and the Kadambas. The Vākāṭakas were weakened by attacks from Mālava and Kosala in the 5th century. Ultimately, the Cālukyas of Vātāpi (Bādāmi) ended their rule.

Inscriptions issued by the Nala kings show that part of the territory in Vidarbha was ruled by them in the 6th century. Nala territory extended through Bastar as far as Koraput (in Utkala [Orissa]). The Bhojas were also powerful in Vidarbha, probably in the Amraoti region. The Traikūṭakas were ruling in the northern coastal areas from approximately Bombay to Surat. In the second half of the 6th century, the Kalacuris settled in the Narmada Valley and occupied northern Mahārāshtra, Gujarāt, and Mālava. They also adopted the Cedi era dating from 248–249. Harassed by the rising power of the Cālukyas and Gurjaras, the Kalacuris moved east to Jabalpur, where they were to come into prominence again in the 9th century.

In the eastern Deccan, the Guntūr and Tenāli districts were ruled by the Ānanda family. In the 4th century Veṅgi came under the control of the Śālaṅkāyanas and, in the next century, passed into the hands of the Viṣṇukuṇḍin family, who ruled there until the early 7th century. The Viṣṇukuṇḍins, like many of the other newly risen dynasties of the peninsula, took great pride in the performance of Vedic sacrifices. They married into the Vākāṭaka family, probably to secure themselves against their neighbours—the eastern Gaṅgas and the Māṭharas ruling in Kaliṅga. The Piṭṛbhaktas ruled the intervening territory of Vishākhapatnam. The Māṭharas occupied the Srīkākulam area and appear to have been overthrown by the Vasiṣṭha. An eastern branch of the Gaṅga dynasty established itself in southern Karnātaka and soon controlled not only the Ganjam district from where they began but Srīkākulam as well. Dakṣiṇa Kosala or southern Kosala (Raipur, Bilāspur, and Sambalpur areas), once part of the Sātavāhana empire, came under the rule of the Śarabhapurīya family. After the mid-6th century, the kingdom passed into the hands of the Pāṇḍuvaṁśis. A branch line of the dynasty appears to have ruled in Mekala (the Maikala Range near Amarakaṇṭaka).

The Cālukyas

The most significant kingdom of the Deccan was that of the Cālukyas, who are associated with Vāṭāpi (Bījāpur district) in the 6th century AD. The Cālukyas controlled large parts of the Deccan for two centuries. There were many branches of the family, the most important of which were the Eastern Cālukyas, ruling at Piṣṭapura (Pithāpuram in the Godāvari Delta) in the early 7th century; the Cālukyas of Vemulavada (near Karīmnagar, Andhra Pradesh); and the renascent later Cālukyas of Kalyāṇī (between the Bhīma and Godāvari rivers), who rose to power in the 10th century. Cālukya power reached its zenith during the reign of Pulakeśin II (610–642), a contemporary of Harṣa (see above *Successor states*). The early years of his reign were taken up with a civil war, after which Pulakeśin had to reconquer lost territories and reestablish his control over recalcitrant feudatories. Pulakeśin then campaigned successfully in the south against the Kadambas, the Alūpas, and the Gaṅgas. Finally, a matrimonial alliance with the Gaṅgas terminated the campaign. An attack on Konkan resulted in the acquisition of the island of Ghārāpuri (Elephanta Island) near Bombay. Leading his armies north, he defeated the Lāṭas, Mālavas, and Gurjaras. Pulakeśin's final triumph in the north was the victory over Harṣa of Kannauj. Pulakeśin then turned his attention to the eastern Deccan and conquered southern Kosala, Kaliṅga, Piṣṭapuram, and the Viṣṇukuṇḍin kingdom. He started the collateral branch of the Eastern Cālukyas based at Piṣṭapuram with his younger brother Viṣṇuvardhana as the first king. Pulakeśin then launched another major campaign against the powerful South Indian kingdom of the Pallavas, in which he defeated their king Mahendravarman I, beginning a

Cālukya–Pallava conflict that was to continue for many centuries. Pulakeśin II sent an embassy to the court of the Persian king Khusru II. Good relations between the Persians and the Indians of the Deccan were of great advantage to the Zoroastrians of Persia, who, fleeing from the Islāmic persecution in subsequent centuries, sought asylum in India and settled along the west coast of the Deccan. Their descendants today are the Parsi community.

The advantage to the Cālukya king of this conquest was that he was "the lord of the eastern and western waters"; that is, he controlled both coasts of the Deccan. The major river valleys of the northern Deccan—the Narmada, the Tāpti, the Godāvari with its tributaries, and the Krishna were in Cālukya hands, as were the valuable routes in the valleys. This in turn meant the control of the west coast trade with west Asia and the Kalinga and Andhra trade on the east coast with Southeast Asia. The centuries-long conflict between the northern and the southern Deccan, of which the Cālukya–Pallava conflict was but a facet, also had geographical, political, and economic reasons. Any South Indian power wishing to expand would inevitably try to move up the east coast, which was not only the most fertile area of the peninsula but was also rich from the income of trade with Southeast Asia. Therefore, any power wishing to control the northern Deccan had to control the east coast as well, since rivers and routes ran west to east. With the major maritime activity gradually concentrating on southeast Asian trade, in which even the west coast had a large share, the control of both coasts was of considerable economic advantage. It was along the east coast, therefore, that the conflict between the two regions was often expressed. The next 100 years of Cālukya power in the main witnessed the continuation of this conflict, weakening both contenders. Ultimately, in the mid-8th century, a feudatory of the Cālukyas, Dantidurga of the Rāṣṭrakūṭa family, rose to importance and established himself in place of the declining Cālukya dynasty. The Eastern Cālukyas, who had managed to avoid involvement in the conflict, survived longer and came into conflict with the Rāṣṭrakūṭas. Another branch of the Cālukyas established itself at Lāṭa in the mid-7th century and played a prominent role in obstructing the Arab advance, but it too declined soon after.

Southern India. The southern part of the peninsula split into a number of kingdoms each fighting for supremacy. Cēra power was based mainly on a flourishing trade with western Asia. The Cōḷas retired into insignificance in the Uraiyūr (Tiruchchirāppalli) area. The Pāṇḍyas were involved in fighting the rising power of the Pallavas, and this occasionally led them into alliances with the Deccan kingdoms. In Karnātaka the Gaṅgas were dominant, but, placed as they were between the Pallavas and the Kadambas, they could hardly expand. Nor did their distant collaterals, the Eastern Gaṅgas, come to their aid in times of distress. The Kadambas came to power in Kuntala after the decline of the Cūtūs in the 4th century. Frequently under attack from the Vākāṭakas and later the Pallavas, they succumbed finally to the Cālukyas in the 6th century. The Bāṇas occupied the Kolār region adjoining the Pallava kingdom and more often than not were feudatories of the latter.

The rise of the Pallavas

The origin of the Pallava dynasty is obscure. It is not even clear whether the early Pallavas of the 3rd century AD were the ancestors of the later Pallavas of the 6th century, who are sometimes distinguished by the title "imperial." It would seem, however, that their place of origin was Toṇḍaimaṇḍalam, with its centre at Kānchipuram (ancient Kāñcī). Prākrit copperplate charters issued by the early kings from Kānchipuram often mention places just to the north in Andhra Pradesh, suggesting that the dynasty may have migrated to the Kānchipuram area. The Sanskrit and Tamil epigraphic records of the later kings of the dynasty indicate that the later Pallavas became dominant in the 6th century after a successful attack against the Kalabhras, which extended their territory as far south as the Kāverī. The Pallavas reached the height of their power during the reign of Mahendravarman I (*c.* 600–630), a contemporary of Harṣa and Pulakeśin II. Among the sources of the period, Hsüan-tsang's account acts as a connecting link since he travelled through the domains of all three kings. The struggle for Veṅgi between the Pallavas and the Cālukyas became the immediate pretext for a long drawn-out war, which began with the defeat of the Pallavas. Apart from his campaigns, Mahendravarman was a writer and artist of some distinction. His play, the *Maṭṭavilāsaprahasana,* treats in a farcical manner the idiosyncrasies of Buddhist and Śaiva ascetics.

His successor, Narasiṃhavarman I Mahāmalla (*c.* 630–668), avenged the Pallava defeat by capturing Vāṭāpi. He sent two naval expeditions to Ceylon (Sri Lanka) to assist the king Mānavamma to regain his throne. The expeditions set out from Mahābalipuram, where port facilities had been developed. Pallava naval interests laid the foundation for the extensive use that their successors, the Cōḷas, made of the navy. Toward the end of the 8th century, the Gaṅgas and the Pāṇḍyas joined coalitions against the Pallavas. As the Cālukyas declined under pressure from the Rāṣṭrakūṭas, the Pāṇḍyas gradually took on the Pallavas and, by the mid-9th century, advanced as far as Kumbakonam. This defeat was avenged, but, by the end of the 9th century, Pallava power had ceased to matter very much.

Society and culture. Some of the Pallava kings took an interest in the Āḻvārs and Nāyaṇārs, the religious teachers who preached a new form of Vaiṣṇavism and Śaivism based on the *bhakti* cult. Among the Śaivas were Appar (who is said to have converted Mahendravarman from Jainism) and Māṇikkavācakar. Among the Vaiṣṇavas were Nammāḻvār and a woman teacher, Āṇḍāḷ. The movement aimed at preaching a popular Hinduism, in which Tamil was preferred to Sanskrit. Emphasis was placed on the participation of women in the congregations and the concept of the peripatetic teacher. The movement gradually spread northward. The Tamil devotional cult was in a sense competitive with Buddhism and Jainism, both of which suffered a gradual decline. After a long period of quiescence, however, Jainism found a foothold in Rājasthān and Gujarāt, while Buddhism flourished in eastern India in places such as Nālandā, a monastery that attracted vast numbers of students from India and abroad. Buddhism had come to be influenced by various esoteric Tibetan and east Indian cults, particularly the Tantric cults, and a new school called Vajrayāna (Thunderbolt Vehicle) Buddhism developed. The widespread Śakti cult associated with Hindu practice was based on the notion that the male can be activated only by union with the female. Thus the gods were given consorts—Lakṣmī and Śrī for Vishnu and Pārvatī, Kālī and Durgā for Śiva—and ritual was directed toward the worship of the mother-goddess. Much of it was derived from the earlier fertility cults and local rites and beliefs that were assimilated into Hinduism. It may also have been due to the incorporation of non-Aryan cultures into Hindu society and the religion that the peoples of these cultures would have brought with them (see also HINDUISM).

At the same time orthodox Brahmanism received encouragement, especially from the royal families. Learned Brahmins were given endowments of land. The new Brahmanism now acquired a locality and an institution in the form of the temple. The earliest remains of a Hindu temple discovered at Sānchi date to the Gupta period. The structure of these temples was extremely simple and consisted of a shrine room, called a *garbhagṛha* ("womb house"), where the image of the deity was placed and which opened onto a porch. Over the centuries additional structures were added until the temple complexes covered many acres. In the peninsula, the early temples were rock-cut shrines in imitation of Buddhist models. Although the Cālukyas did introduce freestanding temples, most of their patronage extended to rock-cut monuments. The Pallavas also began with rock-cut temples, as at Mahābalipuram, but, when they took to freestanding temples, they produced the most impressive examples of their time.

Temple building

With the growth of temples and monasteries, there also grew up traditions of mural painting and sculpture. Earlier examples of mural painting occur at Bāgh (Dhār district) and Sittanavāsal (Pudukkottai district), and the tradition reached its apogee in the murals at Ajantā (Aurangābād

district), the best of which date to Vākāṭaka and Cālukya periods. The style of murals on the walls of Buddhist monasteries travelled from India to Afghanistan and Central Asia and ultimately to China. Equally impressive was the Buddhist sculpture at Sārnāth, in Uttar Pradesh. It is possible that the Buddhist images led to the depiction of Hindu deities in iconic form. The symbolism of the Hindu deity, however, took precedence over mere representation, and the deities were portrayed with many arms each carrying a symbolic attribute.

Temples were richly endowed with wealth and land. The large endowments could accommodate the colleges of higher learning (*ghaṭikā* and *maṭha*), primarily for priests. These became responsible for much of the formal education, and inevitably the use of Sanskrit became widespread. There was an appreciable interest in philosophy, with heated debates between Buddhist and Hindu philosophers. Hindu philosophy had by now been classified into six major schools: Nyāya, Vaiśeṣika, Sāṃkhya, Yoga, Mīmāṃsā, and Vedānta. An indication of the growing domination of Brahmanical intellectual life is that the ancient *Purāṇas* were now rewritten substantially in their present form under Brahmanical influence.

Sanskrit literature

The flowering of classical Sanskrit literature also belongs to this period. The plays and poems of Kālidāsa (*Abhijñānaśākuntala, Mālavikāgnimitra, Vikramorvaṣīya, Raghuvaṃśa, Meghadūta*) are believed to have been written during the time of Candra Gupta II, although Kālidāsa's date is still placed anywhere between the 2nd and 7th centuries AD. In the south, the propagation of Sanskrit resulted in the *Kirātārjunīya,* an epic written by Bhāravi (7th century), and in Daṇḍin's *Daśakumāracarita,* a collection of popular stories (6th century). Tamil literature was equally productive, as is expressed in the composition of two didactic works, the *Tirukkuṛaḷ* and *Nāladiyār,* and in the more lyrical strain of the *Siḷappadikaram* and the *Maṇimekhaḷai,* two Tamil epics. Representing a less common genre of literature in the Gupta period was the *Kāma-sūtra* of Vātsyāyana, a manual on the art of love. This was a collation and revision of earlier texts and displays a remarkable sophistication and urbanity.

The monasteries and temples were centres of formal learning, and the guilds were centres of technical knowledge. The mixture of the theoretical and practical, however, sometimes occurred, as in the case of medicine, particularly veterinary science. Advances in metallurgy affected sculpture, as in the Sultānganj Buddha and a famous iron pillar now at Meharaulī (Delhi). Gold and silver coins of the Gupta period show a quality of minting technology that was not to be surpassed for many centuries.

Science and mathematics

Mathematics was particularly advanced, probably more so than anywhere in the world at the time. Indian numerals were later borrowed by the Arabs and introduced to Europe as Arabic numerals. The use of the cipher and the decimal system is confirmed by inscriptions. With advances in mathematics there was a comparable advance in astronomy. Āryabhaṭa, writing in AD 499, calculated π to 3.1416, the solar year to 365.3586 . . . days and stated that the earth was spherical and rotated on its axis. That European astronomy was also known is suggested by the 6th-century astronomer Varāhamihira, who mentions the Romaka Siddhānta (School of Rome) among the five major schools of astronomy.

Legal texts and commentaries were abundant—the better known being those of Yājñavalkya, Nārada, Bṛhaspati, and Kātyāyana. The judicial process had become more formalized with a hierarchy of procedures and judges. Also there was a great increase in the amount of legal commentary on social problems and property rights. The post-Gupta period saw considerable social change not only with migrations from outside India but also from the Aryanizing of non-Aryan tribes within India. The expanding village economy opened up new areas geographically, and the increasing importance of guilds in the towns indicated fresh perspectives on social life.

750–C. 1200

Northern India. *The tripartite struggle.* The century saw the start of the tripartite struggle for Kannauj between the Pratihāra, the Rāṣṭrakūṭa, and the Pāla dynasties. The Pratihāras, indirectly associated by some historians with the Jodhpur Pratihāras, rose to strength in the Avanti-Jalaor region in the second quarter of the 8th century. They appear to have used western India as a base from which to extend their power through central India into the Ganges Valley. The capture of Kannauj was not in itself of great political importance, but it symbolized the control of the central Ganges Valley, which gave access to the entire Ganges Plain. The Rāṣṭrakūṭa dynasty, a feudatory of the Cālukyas, rose to power during the reign of Dantidurga, *c.* 753, who not only overthrew the Cālukyas but established the new kingdom. The Rāṣṭrakūṭas were the inheritors of the Sātavāhana and Vākāṭaka kingdoms, and their interest in Kannauj was probably to try to capture the trade routes from the Ganges Valley. This was the first occasion on which a power based in the Deccan made a serious bid for a pivotal position in northern India. According to Tāranātha, a 16th-century Buddhist historian of Tibet, the Pālas came from Puṇḍravardhana (Bogra district), and their first ruler, Gopāla (*c.* 750–770), included Vaṅga in his kingdom and gradually extended his control to the whole of Bengal.

Vatsarāja, a Pratihāra ruler who came to the throne *c.* 778, controlled eastern Rājasthān and Mālava. His ambition to take Kannauj brought him into conflict with the Pāla king, Dharmapāla (*c.* 770–810), who had by this time advanced up the Ganges Valley. The Rāṣṭrakūṭa king Dhruva (*c.* 780–793) attacked each in turn and claimed to have defeated them. This initiated the so-called tripartite struggle, at the beginning of which, the Pratihāras were in an inferior position, as the Pālas were able to consolidate their position in eastern India. Dharmapāla soon retook Kannauj and put his nominee on the throne. The Rāṣṭrakūṭas were preoccupied with problems in the south. Vatsarāja's successor, Nāgabhaṭa II (*c.* 793–833), reorganized Pratihāra power, attacked Kannauj, and for a short while reversed the situation. But soon after he was defeated by the Rāṣṭrakūṭa king Govinda III (793–814), who in turn had to face a confederacy of southern powers that kept him involved in Deccan politics, leaving northern India to the Pratihāras and Pālas. Bhoja (*c.* 836–885) revived the power of the Pratihāras by bringing Kālañjara (Bānda district), and possibly Kannauj as well, under Pratihāra control. Bhoja's plans to extend the kingdom, however, were thwarted by the Pālas and the Rāṣṭrakūṭas. More serious conflict with the latter ensued during the reign of Krishna II (*c.* 878–914). Initially Bhoja met with success, but later some of the territory he gained had to be returned to the Rāṣṭrakūṭas.

An Arab visitor to western India, the merchant Sulaymān, referred to the kingdom of Juzr (which is generally identified as Gurjara) and its strong and able ruler, who may have been Bhoja. Of the successors of Bhoja, the only one of significance was Mahīpāla (*c.* 908–942), whose relationship with the earlier king remains controversial. Rājaśekhara, a renowned poet at his court, implies that Mahīpāla restored the kingdom to its original power, but this may be an exaggeration. By the end of the 10th century, the Pratihāra feudatories—Cāhamānas, Chandelās (Candellas), Guhilas, Kalacuris, Paramāras, and Caulukyas—were asserting their independence, although the last of the Pratihāras survived until 1027. Meanwhile Devapāla (*c.* 810–850) was reasserting Pāla authority in the east and, he claimed, in the northern Deccan. The end of the 9th century, however, saw the decline of the Pāla kingdom with feudatories in Kāmaūpa (modern Assam) and Utkala (Orissa) taking independent titles. Pāla power revived during the reign of Mahīpāla (*c.* 988–1038), although their stronghold now was Bihār rather than Bengal. Further attempts to recover the old Pāla territories were made by Rāmapāla, but, by 1120, even he had failed. After the collapse of the last Pāla king in 1162, the Varman dynasty ruled in Bengal. There was a brief revival of power in Bengal under the Sena dynasty (*c.* 1070–1289). Finally Mithilā (northern Bihār) came under the control of the Karṇāṭa dynasty, probably founded by an officer from Karnātaka.

The Pratihāra feudatories

In the Rāṣṭrakūṭa kingdom, Amoghavarṣa (*c.* 814–878)

faced a revolt of officers and feudatories but managed to survive and reassert Rāṣṭrakūṭa power despite intermittent rebellions. Campaigns in the south against Veṅgi and the Gaṅgas kept him preoccupied and prevented him from participating in northern politics. The Rāṣṭrakūṭa capital was moved to Mānyakheṭa (Malkhed, in Andhra Pradesh), doubtless to facilitate southern involvements, which clearly took on more important dimensions at this time. Sporadic campaigns against the Pratihāras, the Eastern Cālukyas, and the Cōḷas, the new power of the south, continued (see below *The Cōḷas*). Indra III (914–927) captured Kannauj, but, with mounting political pressures from the south, his control over the north was inevitably short-lived. The reign of Krishna III (*c.* 939–968) saw a successful campaign against the Cōḷas, a matrimonial alliance with the Gaṅgas, and the subjugation of Veṅgi. Rāṣṭrakūṭa power declined suddenly, however, after the reign of Indra, and this was fully exploited by the feudatory Taila.

Taila II (973–997) was of the Cālukya family and ruled a small part of Bijāpur. He traced his ancestry to the earlier Cālukyas of Vāṭāpi. On the weakening of the Rāṣṭrakūṭa power, he defeated the King, declared his independence, and founded what has come to be called the later Cālukya dynasty. The kingdom included southern Karnātaka, Konkan, and the territory as far north as the Godāvari. By the end of the 10th century, the later Cālukyas clashed with the expanding power of the Cōḷas. The Cālukyas' capital was subsequently moved north to Kalyāṇī (in Bidār). Campaigns against the Cōḷas took a more serious turn during the reign of Someśvara I (1043–68), with alternating defeat and victory. The later Cālukyas, however, by and large retained control over the western Deccan despite the hostility of the Cōḷas and of their own feudatories. In the middle of the 12th century, however, a feudatory, Bijjala (1156–67) of the Kalacuri dynasty usurped the throne at Kalyāṇī. The last of the Cālukya rulers, Someśvara IV (1181–*c.* 1189), regained the throne for a short period, after which he was overthrown by a feudatory of the Yādava dynasty.

Minor kingdoms

On the periphery of the large kingdoms were the smaller states such as Nepal, Kāmarūpa, Kashmir, and Utkala (Orissa). Nepal had freed itself from Tibetan suzerainty in the 8th century but still remained a major trade route to Tibet. Gradually the *rāna*s, the feudal chiefs, gained power and threatened the position of the king. Kāmarūpa with its capital at Prāgjyotiṣapura (near Gauhāti) was one of the centres of the Tantric cult. In 1253 a major part of Kāmarūpa was conquered by the Ahoms, a Shan (Burmese) people from whom it derives its present name, Assam. Politics in Kashmir were dominated by turbulent feudatories seeking power. By the 11th century Kashmir was torn between rival court factions, and royal oppression of King Harṣa accentuated the suffering of the people. Smaller states along the Himalayan foothills managed to survive without becoming too embroiled in the politics of the plains.

The Rājputs. In Rājasthān and central India there arose a number of small kingdoms ruled by dynasties that came to be called the Rājputs (literally, sons of kings). The term was used by royal families that claimed Kṣatriya status and linked their lineage with either the Sūryavaṃśi or the Candravaṃśi, the royal lineages of the *itihāsa-purāṇa* tradition; or else with the Agnikula (Fire lineage) based on a lesser myth in which the eponymous ancestor arises out of the sacrificial fire. The four major Rājput dynasties—Pratihāra, Paramāra, Cauhān (Cāhamāna), and Caulukya—claimed Agnikula lineage. The claims made in Rājput genealogies to supernatural ancestry suggest that either they were families of obscure origin—perhaps semi-Hinduized local tribes who gradually acquired political and economic status—or else that they were of non-Indian (probably Central Asian) origin.

The Caulukyas of Gujarāt (also known as the Solaṅkis) had three branches: one ruling Mattamayūra (the Mālava–Cedi region), one established on the erstwhile kingdom of the Cāpas at Aṇahilapāṭaka (Patan), and the third at Bhṛgukaccha (Broach) and Lāṭa in the coastal area. By the 11th century they were using Gujarāt as a base and attempting to annex neighbouring portions of Rājasthān and Avanti. Kumārapāla (*c.* 1143–72) was responsible for consolidating the kingdom. He is also believed to have become a Jaina and to have encouraged Jainism in western India. Hemacandra, an outstanding Jaina scholar noted for his commentaries on political treatises, was a well-known figure at the Caulukya court. Many of the Rājput kingdoms had Jaina statesmen, ministers, and even generals, as well as Jaina traders and merchants. By the 14th century, however, the Caulukya kingdom had declined.

Adjoining the kingdom of the Caulukyas was that of the Paramāras in Mālava, with minor branches in the territories just to the north (Mount Ābū, Bānswāra, Dūngarpur, and Bhīnmāl). The Paramāras had started as feudatories of the Rāṣṭrakūṭas and rose to eminence during the reign of Bhoja. An attack by the Caulukyas weakened the Paramāras in 1143, and, although the dynasty was later re-established, it remained weak. In the 13th century the Paramāras were threatened by both rising Yādava power in the Deccan and the Turkish kingdom at Delhi; the latter conquered the Paramāras in 1305.

The Kalacuris of Tripurī (near Jabalpur) also began as feudatories of the Rāṣṭrakūṭas. The Kalacuris became a power in central India in the 11th century during the reigns of Gāṅgeyadeva and his son Lakṣmīkarṇa, when attempts were made to conquer territories as far afield as Utkala (Orissa), Bihār, and the Doab. Here they came into conflict with the Turkish governor of the Punjab, who had extended his territory as far as Vārānasi. To the west, there were conflicts with Bhoja Paramāra, and the Kalacuris declined at the end of the 12th century.

The Chandelās, whose kingdom comprised mainly Bundelkhand (Jejākābhukti), were feudatories of the Pratihāras. Among the important rulers was Dhaṅga (*c.* 950–1008), who issued a large number of inscriptions and was generous in donations to Jaina and Hindu temples. Dhaṅga's grandson Vidyādhara (1017–29), often described as the most powerful of the Chandelā kings, extended the kingdom as far as the Chambal and Narmada rivers. This brought him into direct conflict with the Turkish leader Maḥmūd of Ghazna, though the battles between them were indecisive. The Chandelās had also to face the attacks of the Cauhāns, who were in turn being harassed by the Turks. The Turkish kingdom at Delhi encroached into Bundelkhand, but the Chandelās survived until the 16th century, as minor chieftains.

The Kachchhapaghātas ruled at Gopādri (Gwalior), with branches in the neighbourhood at Dubkund and Narwar. They were ultimately conquered by the Turks. The Gāhaḍavālas rose to importance in Vārānasi after that city was sacked by the Turks in 1034. They extended their kingdom up the Doab including Kannauj. The king Jayacandra (12th century) is mentioned in the poem *Pṛthvīrāja-rāso* of Candbardāi, in which his daughter, the princess Sanyogita, elopes with the Cauhān king Pṛthvīrāja. Jayacandra was killed in battle against the Turkish leader, Muḥammad Ghūri, and his kingdom was annexed. The Guhilas became prominent after their capture of Chitor in the 8th century, their original centres having been Nāgahrada (Nāgda) and Āghaṭa (Ahar), which became the nuclei of the region called Medapāṭa (Mēwar). Despite hostilities with the Caulukyas and Cauhāns, the Guhilas maintained their position. A severe attack from the Turks in the early 13th century was parried under the leadership of Jaitrasiṃha (*c.* 1213–52). In the Śripatha region (Bharatpur) at Bayana ruled the Yaduvaṃśa family, claiming, as the name implies, descent from the ancient Yadu tribe. Their activities were focussed on the area between Āgra and Delhi. The Haryāna region north of Delhi was ruled by the Tomara Rājputs who founded the city of Dhillikā (Delhi) in 736. They were originally feudatories of the Pratihāras. In the 12th century they were overthrown by the Cauhāns.

The Cauhāns

Inscriptional records associate the Cauhāns with the lake Śākambharī and its environs (Sāmbhar Salt Lake in Jaipur district). Cauhān politics were largely campaigns against the Caulukyas and the Turks. In the 11th century they founded the city of Ajayameru (Ajmer) along the southern part of their kingdom, and in the 12th they captured Dhillikā from the Tomaras and annexed some of their

territory along the Yamuna River. The name of Pṛthvīrāja III has come down both in folk and historical literature as the Cauhān king who resisted the Turkish attacks in the first battle at Tarāori in 1191. Pṛthvīrāja, however, was defeated at a second battle in the same place in 1192, which defeat ushered in Turkish rule in northern India.

The coming of the Turks. The establishment of Turkish power in India is initially tied up with politics in the Punjab. The Punjab was ruled by Jayapāla of the Hindu Shāhiya dynasty, which dynasty had in the 9th century wrested the Kābul Valley and Gandhāra from a Turkish Shāhiya. Relations both political and economic were extremely close between the Punjab and Afghanistan. Afghanistan in turn was closely involved with Central Asian politics. Sebüktigin, a Turk, was appointed governor of Ghazna in 977. He attacked the Hindu Shāhiyas and advanced as far as Peshāwar. His son Maḥmūd succeeded to the Ghazna principality in 998. He went to war with the Shāhiya dynasty and almost every year until his death in 1030 led raids against the rich temple centres in northern and western India, using the wealth obtained from the raids to finance successful campaigns in Central Asia and to build a Central Asian empire. He emerged as a patron of culture and was responsible for sending to India the scholar al-Bīrūnī, whose study, *Ta'rīkh al-Hind* is a source of valuable information. A by-product of temple raiding was that it also credited Maḥmūd with religious merit, since he claimed that he was destroying the idols of the infidels. He left his governors in the Punjab with a rather loose control over the region.

In the 12th century the Ghūrid Turks were driven out of Khorāsān and later out of Ghazna by the Khwārezm-Shāhs. Inevitably the Ghūrids sought their fortune in northern India, where the conflict between the Ghaznavids and the local rulers provided an excellent opportunity. Muḥammad Ghūri advanced into the Punjab and captured Lahore in 1185. Victory in the second battle of Tarāin consolidated his success, and he left his general Quṭb-ud-Dīn Aybak in charge of his Indian possessions. Muḥammad was assassinated in 1206 on his way back to Afghanistan. Quṭb-ud-Dīn remained in India and declared himself sultan of Delhi, the first of the Mamlūk (Slave) dynasty.

The Deccan and the south. In the northern Deccan the decline of the later Cālukyas brought about the rise of their feudatories, among them the Yādava dynasty (also claiming descent from the Yadu tribe) based at Devagiri (Daulatābād), whose kingdom (Seunadeśa) included the Khāndesh (now divided into Dhule and Jalgaon districts), the Nāsik, and the Ahmadnagar districts. The kingdom expanded during the reign of Siṃhana (*c.* 1210–47), who campaigned against the Hoysaḷa in northern Karnātaka against the lesser chiefs of the western coast, and against the Kākatīya kingdom in the eastern Deccan. Turning northward, he attacked the Paramāras, the Caulukyas, and the Lāṭa. The Yādavas, however, facing the Turks to the north and the powerful Hoysaḷas to the south, declined in the early 14th century.

In the eastern Deccan, the Kākatīya dynasty was based in the Nalgonda and Warangal areas (Andhra Pradesh) and survived until the Turkish attack in the 14th century. The Eastern Cālukyas ruled in the Godavāri Delta, and, in the 13th century, their fortunes were tied to those of the Cōḷas. The Eastern Gaṅgas, ruling in Kaliṅga, came into conflict with the Turks advancing down the Ganges Valley to the delta in the 13th century.

The Cōḷas. The Cōḷas were by far the most important dynasty in the subcontinent at this time, although their activities mainly affected the peninsula and Southeast Asia. The nucleus of Cōḷa power during the reign of Vijayālaya in the late 9th century was Thanjāvūr, from which they spread northward, annexing in the 10th century what remained of Pallava territory. To the south they came up against the Pāṇḍyas. Cōḷa history can be reconstructed in considerable detail because of the vast number of lengthy inscriptions issued not only by the royal family but also by temple authorities, village councils, and trade guilds. Parāntaka I (907–953) laid the foundation of the kingdom. He took the northern boundary up to Nellore (Andhra Pradesh), where his advance was stopped by a defeat at the hands of the Rāṣṭrakūṭa king Krishna III. Parāntaka was more successful in the south, where he defeated both the Pāṇḍyas and the Gaṅgas. He also launched an abortive attack on Ceylon (Sri Lanka). For 30 years after his death, there was a series of overlapping reigns that did not strengthen the Cōḷa position. There then followed two outstanding rulers who rapidly reinstated Cōḷa power and ensured the kingdom its supremacy. These were Rājarāja I and Rājendra.

Rājarāja I's conquests

Rājarāja (985–1014) began with the south and attacked the Pāṇḍyas and Īllamaṇḍalam (Ceylon). In the case of the latter, the old capital at Anurādhapura was destroyed, and a new capital was established by the Cōḷas at Polonnaruva. Northern Ceylon became a province of the Cōḷa kingdom. A campaign against the Gaṅgas and Cālukyas extended the Cōḷa boundary north to the Tungabhadra River. On the eastern coast the Cōḷas battled with the Cālukyas for the possession of Veṅgi. A marriage alliance gave the Cōḷas an authoritative position, but Veṅgi remained a bone of contention. A naval campaign led to the conquest of the Maldive Islands. The conquest of these islands as also of the Malabār Coast and northern Ceylon can be explained as necessary to the Cōḷa control over trade with Southeast Asia and with Arabia and East Africa. These were the transit areas, ports of call for the Arab traders and ships to Southeast Asia and China, the source of the valuable spice sold at a high profit to Europe.

Rājarāja I's son Rājendra was associated with the government from 1012 and ruled until 1044. To the north he annexed the Raichūr Doab and moved into Mānyakheṭa in the heart of Cālukya territory. A revolt against Mahinda V of Ceylon gave Rājendra the excuse to conquer southern Ceylon as well. In 1021–22 the now-famous northern campaign was launched. The Cōḷa army campaigned along the east coast as far as Bengal and then north to the Ganges River—almost the exact reverse of Samudra Gupta's campaign to Kānchipuram in the 4th century AD. But the most spectacular campaign was a naval campaign against the Śrivijaya kingdom in Southeast Asia in 1025. The reason for the assault on Śrivijaya and neighbouring areas appears to have been the interference with Indian shipping and mercantile interests seeking direct trading connections with South China. The Cōḷa victory reinstated these connections, and throughout the 11th century Cōḷa trading missions visited China.

The Hoysaḷas and Pāṇḍyas. The succession after Rājendra is confused until the emergence of Kulottuṅga I (1070–1122), who was of mixed Cōḷa and Eastern Cālukya descent. But his reign was the last of any significance. The 12th and 13th centuries saw a gradual decline in Cōḷa power, accelerated by the rise of the Hoysaḷas to the west and the Pāṇḍyas to the south.

The Hoysaḷas also claimed Yādava origin and began as hill chieftains, feudatory to the Cālukyas. The core of the kingdom was Dōrasamudra (near Karnātaka), and the kingdom was consolidated by Viṣṇuvardhana (12th century). The Hoysaḷas were involved in conflict with the Yādava kingdom, which was seeking to expand southward, particularly during the reign of Ballāla II (1173–1220). Hostilities also developed with the Cōḷas to the east. The armies of the Turks ultimately diminished the Hoysaḷa kingdom to the point of nonexistence, until, in the 14th century, it gave way to the Vijayanagara kingdom. In the 13th century the Pāṇḍyas became the dominant power in the south, but their supremacy was brief since they were attacked in the 14th century by Turkish armies. Information on the dynasty is supplemented by the account of Marco Polo who visited the region in 1288 and 1293.

Society and culture. *Indian feudalism.* Apart from the political events of the time, a common development in the subcontinent was the recognizable decentralization of administration and revenue collection. From the Cōḷa kingdom, there are long inscriptions on temple walls referring to the organization and functioning of village councils. Village councils

Villages that had been donated to Brahmins had councils called the *sabhā*; in the non-Brahmin villages the council was called the *ur*. Eligibility qualifications generally relating to age and ownership of property were indicated, along

with procedural rules. The council was divided into various committees in charge of the different aspects of village life and administrations. Among the responsibilities of the council was the collection of revenue and the supervision of irrigation.

In the Deccan, the rise and fall of dynasties was largely the result of the feudatory pattern of political relationships. The same held true of northern India and is seen both in the rise of various Rājput dynasties as in their inability to withstand the Turkish invasions. There is considerable controversy among historians as to whether the feudatory pattern can be accurately described as feudalism. Some argue that, although it was not identical to the classical example of feudalism of western Europe, there are sufficient similarities to allow the use of the term. The counter argument is that the emphasis on the economic contract, essential to western European feudalism was absent in India or was, at least, not as clearly stated. In any event, the patterns of land relations, politics, and culture changed considerably, and the major characteristic of the change is decentralization.

The commonly used term for a feudatory was *sāmanta,* a term that designated either a conquered ruler or, more often, a secular official connected with the administration who had been given a grant of land in lieu of a salary and who had asserted ownership over the land and gradually appropriated rights of ruling the area. There were various categories of *sāmanta*s. As long as a ruler was in a feudatory status, he called himself *sāmanta* and acknowledged his overlord in official documents and charters. Independent status was indicated by the elimination of the title of *sāmanta* and the inclusion instead of royal titles such as maharaja (*mahārājādhirāja*). The feudatory had certain obligations to the ruler. Although virtually in sole control administratively and fiscally over the land granted to him, he nevertheless had to pay a small percentage of the revenue to the ruler and maintain a specified body of troops for him. He was permitted the use of certain symbols of authority on formal occasions and was required, if called upon, to give his daughter in marriage to his suzerain. These major administrative and economic changes, although primarily concerning fiscal arrangements and revenue organization, also had their impact on politics and culture. The grantees or intermediaries in a hierarchy of grants were not merely secular officials but were often Brahmin beneficiaries who had been given grants of land in return for religious services rendered to the state. Such grants were, in most cases, free of taxes and dues. The grants were frequently so lucrative that the Brahmins could marry into the families of local chiefs, and this explains the presence of Brahmin ancestors in the genealogies of the period.

The economy. Cultivation was still carried out by the peasants, generally Śūdras, who remained tied to the land. Since the revenue was now to be paid not to the king but to the *sāmanta,* the peasants naturally began to give more attention to his requirements. The *sāmanta*s copied the life-style of the royal court and tried to set up miniature courts in imitation of the royal model. The culture of the royal court was thus disseminated, but at the same time, the system encouraged parochial loyalties and local cultural interests. One of the manifestations of this local involvement was a sudden spurt of historical literature such as Bilhaṇa's *Vikramāṅkadevacarita,* the life of the Cālukya king Vikramāditya VI, and Kalhaṇa's *Rājataraṅgiṇī,* a history of Kashmir.

This trend toward decentralization was coupled with an apparent tendency toward a decline in trade in all but the fringe areas of the subcontinent. The west coast was active, but Indian traders did not venture out so much, since the Arabs had established trading settlements in India. The Cōḷa and Pāla kingdoms were exceptions. Powerful trading guilds had the political and military support of the Cōḷa monarchy. Even the rich Hindu temples of South India invested their money in trade. Pāla contacts were mainly with Śrivijaya and trade combined with Buddhist interests. The monasteries at Nālandā and Vikramaśīla (Bihār) maintained close relations. By now eastern India was the only region with a sizable Buddhist interest. The hinterland was devoid of trade, and there seems to have been a decline in commercial centres. The traditional routes were still used, and some kingdoms were based essentially on the revenue from such routes as those along the Arāvalli Range, Mālava, and the Chambal and Narmada valleys. The paucity of coins for this region may be the result of a decline in the volume of trade. Significantly, the major technological innovation, the introduction of the sakia (Persian) wheel, or (*araghaṭṭa*) as an aid to irrigation in northern India, pertains to agrarian life and not to urban technology. It has been suggested that these economic and administrative changes introduced the self-sufficient economy of the Indian village.

Social mobility. Historians once believed that the post-Gupta period brought greater rigidity in the caste structure and that this rigidity was partially responsible for the inability of Indians to face the challenge of the Turks. This view is now being modified. The distinctions, particularly between the Brahmins and the other castes, were in theory sharper, but in practice it now appears that social restrictions were not as rigid as believed. Brahmins often lived off the land and founded dynasties. Most of the groups claiming Kṣatriya status were those who had recently acquired it. The conscious reference to being Kṣatriya, a characteristic among Rājputs, is a noticeable feature in post-Gupta politics. The fact that many of these dynasties were of obscure origin suggests considerable social mobility: a person of any caste, having once acquired political power, could also acquire a genealogy connecting him with the traditional lineages and conferring Kṣatriya status on him. A number of new castes such as the Kāyasthas (scribes) and Khatris (traders) are mentioned in the sources of this period. According to the legal sources, they originated from intercaste marriages, but this is clearly an attempt at rationalizing a situation. Many of these new castes played a major role in society. The hierarchy of castes did not have a uniform distribution throughout the country. Thus the Kṣatriya and Vaiśya groups were little emphasized in South India, where the majority were Śūdras and participated in almost all professions. The preeminent position of the Brahmin was endorsed not merely by the fact that many had lands and investments but by the fact that they controlled education. Formal learning was almost restricted to the institutions attached to the temples. Technical knowledge was available in the various artisan guilds. Not all Brahmins, however, were prosperous and powerful. Even among the Brahmins there was a hierarchy; some Brahmin castes, who had perhaps been tribal priests before being assimilated into the Hindu tradition, remained ordinary village priests catering to rites and ceremonies and the day to day religious functions.

Religion. The local nucleus of the new culture led to a large range of religious expression, from the powerfultemple religion of Brahmanism to a widespread popular *bhakti* religion and even more widespread fertility cult. The distinctions between the three were not clearly demarcated. Rites and concepts from the one flowed into the other. The traditional worship of Vishnu and Śiva had the support of the elite and the urban *bhakti* groups. Temples dedicated to Vaiṣṇava and Śaiva deities were the most numerous. But also included were some of the chief deities connected with the fertility cult, and the mother-goddesses played an important role. The *Purāṇa*s had been rewritten to incorporate popular religion. Now the *Upapurāṇa*s were written to record rites and worship of more localized deities. Among the incarnations of Vishnu the most popular was Krishna. To the earlier austere hero of the *Bhagavadgītā* was now added the pastoral and the erotic aspect. The love of Krishna and Rādhā was expressed in sensitive and passionate poetry and sculpture.

The erotic theme in Hinduism

The introduction of the erotic theme in Hinduism was closely connected with the fertility cult and Tantrism. The latter, named after its scriptures, the *Tantra*s, influenced both Hindu and Buddhist ritual. Tantrism originated in eastern India but was associated with the folk religion throughout the subcontinent and, to that degree, represented the conversion of a folk religion into a sophisticated one. The emphasis on the mother-goddess strengthened

the status of the female deities and also linked it with the *śākta-śakti* cult. The erotic aspect was also related to the ritual of coition being necessary to full Tantric rites. The depiction of erotic scenes on temple walls therefore had a magicoreligious context.

Vajrayāna Buddhism current in eastern India, Nepal, and Tibet shows evidence of the impact of Tantrism. The goddess Tārā emerges as the saviour and is in many ways the Buddhist counterpart of *śakti.* Buddhism was on the way out—the Buddha had been incorporated as an avatar of Vishnu, and it had lost much of its popular appeal, which had been maintained by the simple habits of the monks. Its traditional source of patronage had declined with declining trade. Jainism, however, managed to maintain some hold in Rājasthān, Gujarāt, and Karnātaka. The protest aspect of these religions, especially the opposition to Brahmanical orthodoxy, had now been taken over by the Tantrics and the *bhakti* cult. The Tantrics expressed their protest through their rather extreme rites: as also did some of the heretical sects such as the Kalamukhas and Kāpālikas. The *bhakti* cult expressed the more puritanical protest of the urban groups, and it gradually spread to the rural areas. Preeminent among the *bhakti* groups during this period were the Liṅgāyats, or Vīraśaivas, traditionally founded by Basava, who were to become a powerful force in Karnātaka, and the Pandharpur cult in Mahārāshtra, which attracted preachers such as Nāmadeva and Jñāneśvara.

Literature and the arts. It was also in the *maṭha* and the *ghaṭikā,* attached to the temples, that the influential philosophical debates were carried out in Sanskrit. Foremost among the philosophers were Śaṅkarācārya (9th century), Rāmānuja (1017–1137) and Madhva (13th century). The discussions centred on religious problems, such as whether knowledge or devotion was the most effective means of salvation, and problems of metaphysics, including that of the nature of reality.

Court literature, irrespective of the region, continued to be composed in Sanskrit, with the many courts competing for the patronage of the poets and the dramatists. There was a revival of interest in earlier literature, with copious commentaries on prosody, grammar, and technical literature. The number of lexicons increased, perhaps necessitated by the growing use of Sanskrit by non-Sanskrit speakers. Literary style tended to be pedantic and imitative, although there were some exceptions, such as Jayadeva's lyrical poem on the love of Rādhā and Krishna, the *Gītagovinda.* The *bhakti* teachers preached not in Sanskrit but in the local languages, giving a tremendous fillip to literature in these languages. Adaptations of the *Rāmāyaṇa, Mahābhārata,* and *Bhagavadgītā* were used regularly by the *bhakti* teachers. In addition, hymns were composed and congregations addressed in the popular language. There was thus a gradual breaking away from Sanskrit and Prākrit via Apabrahmśa (the "crooked language") and the eventual emergence and evolution of Kannaḍa, Telugu, Marāthī, Gujarātī, Bengali, Oṛiyā, and the dialects of Bhojpuri, Maithili, and Māgadhī.

Differentiations of languages

The period was rich in sculpture both in stone and metal, each region registering a variant style. Western India and Rājasthān emphasized ornateness, with the Jaina temples at Mount Ābū attaining a perfection of rococo. Nālandā was the centre of striking but less ornate images in black stone and of Buddhist bronze icons. Central Indian craftsmen used the softer sandstone. In the peninsula the profusely sculptured, rock-cut temples such as the Kailāsa at Ellora, with Cālukya and Rāṣṭrakūṭa patronage, created a style of their own. The dominant style was that of Cōḷa sculpture, particularly in bronze as typified in the famous Naṭarāja image. The severe beauty and elegance of these bronze images mainly of Śaiva and Vaiṣṇava deities and saints remains unsurpassed. A new genre of painting was the illustration of Buddhist and Jaina manuscripts with miniature paintings popular in Nepal, eastern India and Gujarāt.

Temple architecture was divided into three main styles distinguished by the ground plan of the temple and by the shape of the *śikhara* (tower) that rose over the *garbhagṛha* (cubical structure) and that became the commanding feature of temple architecture. The North Indian temples conformed to the *nāgara* style, as is seen at Osiān (Rājasthān state); Khajurāho (Madhya Pradesh state); and Konārak, Bhubaneswar, and Puri (Orissa state). The Orissa temples, however, remain nearest to the original archetype. The South Indian, or *drāviḍa,* style with its commanding *gopuram* (gateways) can be seen in the Rājarājeśvara and the Gaṅgaikoṇḍacōḷapuram temples. The Deccan style, *vasara,* tended to be an inter-mixture of the northern and the southern, with early examples at Vāṭāpi, Aihoḷe, and Pattadakal and, later, at Halebīd, Belūr, and Somnāthpur in the vicinity of Mysore. Artistic achievement was handsomely rewarded from temple funds, which made the temples the focus of attack from invaders.

The question that is frequently posed as to why the Turks so easily conquered northern India and the Deccan has in part to do with what might be called the medieval ethos. A contemporary observed that the Indians had become self-centred and unaware of the world around them. This was substantially true. There was little interest in the politics of neighbouring countries or in their technological achievement, as is evident, for example, from the slow impact of the use of the stirrup in the early cavalry. The medieval ethos expressed itself not only in the "feudatory" attitude to politics and the parochial concerns that became dominant and prevented any effective opposition to the Turks but also in the trappings of chivalry and romanticism that became central to elite activity.

It has been generally held that the medieval period of Indian history began with the arrival of the Turks (dated to either AD 1000 or 1206) because the Turks brought with them a new religion, Islām, which changed Indian society at all levels. Yet the fundamental changes that took place around the 8th century, when the medieval ethos was introduced, would seem far more significant as a criterion of change than the coming of Islām. (R.Th.)

The early Muslim period

NORTH INDIA UNDER MUSLIM HEGEMONY, C. 1200–1526

The first Muslim conquests in the subcontinent were made by Arabs in Sind during the 8th century, and there had been Muslim trading communities in India at least since that time; but the significant and permanent military movement of Muslims into North India dates from the late 12th century and was carried out by a Turkish dynasty that arose indirectly from the ruins of the ʿAbbāsid caliphate. The road to conquest was prepared by Sultan Maḥmūd of Ghazna (modern Ghaznī in Afghanistan), who conducted more than 20 raids into North India between 1001 and 1027 and established in the Punjab the easternmost province of his large but short-lived empire. Maḥmūd's raids, though militarily successful, were mainly for plunder rather than for conquest. They did reveal the wealth of India, however, and indicated that North India was vulnerable to military attacks.

Permanent Muslim conquest

The Delhi sultanate. The decline of the Ghaznavids after 1100 was accentuated by the sack of Ghazna by the rival Shansabānīs of Ghūr in 1150–51. The Ghūrids, who inhabited the region between Ghazna and Herāt, rose rapidly in power during the last half of the 12th century, partly because of the changing balance of power that resulted from the westward movement of the non-Muslim Karakitai Turks into the area dominated by the Seljuq Turks, who had been the principal power in Iran and parts of Afghanistan during the previous 50 years. The Seljuq defeat in 1141 led to a struggle for power among the Karakitai, the Khwārezm-Shāhs, and the Ghūrids for control of parts of Central Asia and Iran. The Ghūrid invasions of North India are part of this Central Asian struggle, for it was during the last quarter of the 12th century that Ghūrid troops, led by Sultan Ghiyās-ud-Dīn Muḥammad and his brother Muʿizz-ud-Dīn Muḥammad ibn Sām, commonly called Muḥammad of Ghūr, conducted campaigns of expansion both into Khorāsān in the west and into North India in the east.

The Ghūrid conquest. The early Ghūrid invasions of India were led by Muḥammad of Ghūr and his lieutenant Quṭb-ud-Dīn Aybak. After some initial successes, their

attempts to follow the southern route into India through the Gumal Pass met obstacles. In 1178 Ghūrid troops, exhausted after crossing the desert of Mt. Ābū, suffered a severe defeat at the hands of the army of Mūlarāja II, the Caulukya ruler of Gujarāt. The Ghūrids then turned to the northern route, through the Khyber Pass, and destroyed the remnants of Ghaznavid power in India with the capture of Peshāwar (1179), Siālkot (1185), and Lahore (1186).

Destruction of the Ghaznavids

By 1185 the Ghūrids had conquered the local Muslim rulers of the northwest and were in a favourable military position to move against the North Indian Rājput powers. The conquest of the Rājputs was not easy, however, and the Ghūrids were by no means always victorious during the next few years. Muḥammad of Ghūr was defeated by the Cāhamānas (Cauhān) under Pṛthvīrāja in 1191 at Tarāorī (Tarāin), northwest of Delhi, but his forces returned the following year to defeat and kill the Rājput king on the same battlefield. The victory opened the road to Delhi, which was conquered in 1193 but left in the hands of a tributary Hindu king. Muḥammad of Ghūr completed his conquests with the occupation of the military outposts of Hānsi, Kuhram, Sursuti, and Sirhind and then returned to Ghazna with a large hoard of treasure, leaving Quṭb-ud-Dīn Aybak in charge of consolidation and further expansion.

Quṭb-ud-Dīn displaced the Cāhamāna chief and made his headquarters at Delhi in 1193, when he began a campaign of expansion. The Gāhāḍavāla chief Jayacandra was defeated and slain, and Vārānasi (Banāras) was taken in 1194. During the next three years Quṭb-ud-Dīn was occupied mainly with battles against Rājput forces that were trying to reconquer Ghūrid-held lands between Delhi and Gujarāt. In 1198–99, however, he subdued Badaun and Kannauj in the upper Ganges region; and in 1202 he conquered part of the Chandelā (Candella) kingdom of Bundelkhand, on the southeastern border of Ghūrid possessions, with the capture of Kālinjar.

One of the most interesting and instructive episodes of the Ghūrid conquest was the rapid rise of an obscure military adventurer, Ikhtiyār-ud-Dīn Muḥammad Bakhtiyār Khaljī. After he was refused service in both Ghazna and Delhi, he took a low position under the commander of the Vārānasi and Oudh (Awadh) division of Quṭb-ud-Dīn's army. He used his small salary to collect a band of soldiers (mostly Khaljī tribesmen, like himself), and *c.* 1200 he began to raid into the Magadha territory in the eastern Gangetic Plain. Bakhtiyār was a brilliant soldier and soon amassed enough plunder to recruit a large force. He received permission from Quṭb-ud-Dīn to continue his campaigns as long as he could finance them through his own resources. By 1202 Bakhtiyār had conquered Nadia, the capital of the Sena kings of Bengal. Within two years, however, the ambitious general had embarked upon what proved to be a disastrous attempt to conquer Tibet. Defeated by the terrain, the weather, and the harassments of the hill tribes, Bakhtiyār managed to return to Bengal with a few hundred men, and there he died in 1206.

Reasons for Ghūrid success

Had it not been for Bakhtiyār Khaljī, the Ghūrids probably would not have conquered Bihār and parts of Bengal during their first thrust into the Gangetic Plain. It is the element of opportunism plus the means by which he gained his success that make Bakhtiyār significant. He combined military genius with the use of extremely mobile cavalry to defeat his foes. He was able to recruit troops, partly because he had money but also because he was successful in his exploits. The availability of a large number of military adventurers from Central Asia who would follow commanders with reputations for success was one of the important elements in the rapid Ghūrid conquest of the major cities and forces of the North Indian plain. Other factors were important as well: better horses contributed to the success of mobile tactics, and the Ghūrids also made better use of metal for weapons, armour, and stirrups than did most of their adversaries. Perhaps most important was the tradition of centralized organization and planning, which was conducive to largescale military campaigns and to the effective organization of post-campaign occupation forces. While the Rājputs probably saw the Ghūrids as an equal force competing for paramount power in North India, the Ghūrids had in mind the model of the successor states to the ʿAbbāsid caliphate, the old Iranian Sāsānid Empire, and particularly the vast centralized empire of Maḥmūd of Ghazna.

In 1205 Sultan Muḥammad of Ghūr suffered a severe defeat at Andkhvoy (Andkhui) at the hands of the Khwārezm-Shāh. News of the defeat precipitated a rebellion by some of the Sultan's followers in the Punjab, and, although the rebellion was put down, Muḥammad of Ghūr was assassinated at Lahore in 1206. Thus, the Ghūrid possessions were insecure everywhere.

The defeat at Andkhvoy proved to be the beginning of the end of the Ghūrid Central Asian empire, and it was still far from certain that the Ghūrids would be able to consolidate their position in India. They held the major towns of the Punjab, Sind, and of much of the Gangetic Plain, but there were military threats from many quarters, and almost all the land outside the cities still was subject to some form of control by Hindu landholders and chiefs. Even in the Doab (the land between the Ganges and the Yamuna rivers, near Delhi) the Gāhāḍavālas held out against the Turks. Most significantly, the Rājput chiefs of Rājasthān had not been permanently subdued but had retreated to less accessible positions in the hills and forests.

The Mamlūk (Slave) dynasty. It is not certain that the Turks' major objective at the time was the consolidation of their Indian conquests, for it appears that the Ghūrids were not ready to give up their struggle against the Khwārezm-Shāhs and that rivalries among the slave successors to Muḥammad of Ghūr were of paramount importance. When Quṭb-ud-Dīn Aybak, Muḥammad of Ghūr's slave and principal commander in India, assumed authority over the Ghūrid possessions in India, he moved from the neighbourhood of Delhi to Lahore. There he set up guard against another of Muḥammad of Ghūr's slaves, Tāj-ud-Dīn Yildiz of Ghazna, who also claimed his former master's Indian possessions. In 1208 Quṭb-ud-Dīn defeated his rival and captured Ghazna but soon was driven out again. He died in 1210 in a polo accident, having made no effort to extend his Indian conquests, but he had managed to establish the foundation of an Indian Muslim state.

Quṭb-ud-Dīn was the first ruler in what has become known, perhaps unreasonably, as the Slave dynasty, even though three families were represented among its nine members, and only Quṭb-ud-Dīn attained a freed status after he had become ruler. Slavery was, however, an integral and important part of the political system. As practiced in eastern Muslim polities of this period, the institution of slavery provided a nucleus of well-trained and loyal military followers (the *mamlūks*) for important political figures; indeed, one of the principal objects of this form of slavery was to train specialists in warfare and government, usually Turks, whose first loyalty would be to their masters. Slave status was honourable and was a principal avenue to wealth and high position for talented individuals whose origins were outside the ruling group. It has been observed that a slave was a better investment than a son, whose claim was not based upon proved efficiency. Yet, slaves with high qualifications could get out of control, and often slaves or former slaves controlled their masters as much as they were controlled by them. The beneficial results for the sultanate of this type of political interaction were that some men of talent had room to rise within the system and thus were less tempted to tear it down and that the responsibilities of government tended to rest in the hands of capable men, whether or not they were the actual rulers.

Importance of slavery

Iltutmish (reigned 1211–36), son-in-law and successor to Quṭb-ud-Dīn Aybak, was able to consolidate many of the military gains of his predecessors, although the sultanate continued to be hard pressed militarily throughout the 13th century. During his reign Iltutmish was faced with three problems: defense of his western frontier, control over the Muslim nobles within India, and subjugation of the many Hindu chiefs who still exercised a large measure of independent rule. His relative success in all three areas gives him claim to the title of founder of the independent sultanate of Delhi. His reign opened with a factional dis-

The reign of Iltutmish

pute in which he and his Delhi-based supporters defeated and killed the rival claimant to the throne, Quṭb-ud-Dīn's son, and put down a revolt by a portion of the Delhi guards. In the west Iltutmish was passive at first and even accepted investiture from his old rival, Yildiz; but when Yildiz was driven from Ghazna into the Punjab by the Khwārezm-Shāh in 1215, Iltutmish was able to defeat and capture him at Tarāorī. Iltutmish might have faced a threat himself from the Khwārezm-Shāh had it not been for Genghis Khan's attack upon the latter. Again Iltutmish waited while refugees, including the heir to the Khwārezm-Shāhī throne, poured into the Punjab and while Nāṣir-ud-Dīn Qabācha, another of Muḥammad of Ghūr's former slaves, maintained a perilous hold on Lahore and Multān. Iltutmish's political talents were pushed to the maximum as he tried desperately to avoid a direct confrontation with the armies of Genghis Khan. He refused aid to the Muslim Khwārezm heir against the pagan Mongols and yet would not attempt to capture him. Fortunately the Mongols were content to send raiding parties into the Salt Range, which Iltutmish wisely ignored, and eventually the Khwārezm-Shāh prince fled from India after causing enormous destruction within Qabācha's domains. Thus, Iltutmish's cause was advanced, and in 1228 he was able to drive Qabācha from Multān and Uch and, by establishing his frontier east of the Beās River, to avoid a direct confrontation with the Mongols. He was not able to gain effective control of the western Punjab, however, largely because the area was subject to raids by hill tribes.

Iltutmish fared better in the east. In 1225 he launched a successful campaign against Ghiyās-ud-Dīn ʿIwāz Khaljī, one of Bhaktiyār Khaljī's lieutenants, who had assumed sovereign authority in Lakhnāwatī (Bengal) and was encroaching upon the province of Bihār. ʿIwāz Khaljī was defeated and slain in 1226, and in 1229 Iltutmish invaded Bengal and slew Balka, the last of the Khaljī chiefs to claim independent power.

Iltutmish's campaigns against the Rājputs were ultimately less successful, although he temporarily captured Ranthambhor (1226), Mandor (Mandāwar; 1227), and Gwalior (1231) and plundered Bhīlsa and Ujjain in Mālwa (1234–35). His generals suffered defeats, however, at the hands of the Cāhamānas of Būndi, the Caulukyas of Gujarāt, and the Chandelās (Candellas) of Narwar, and his victories were not sufficient to suppress the revival of several Rājput polities.

By 1236 the sultanate of Delhi was established as clearly the largest and most powerful of a number of competing states in North India. The new state had internal momentum as well, surviving severe factional disputes during the next 10 years while four of Iltutmish's children or grandchildren were in turn raised to the throne and deposed. This momentum was maintained largely through the efforts of Iltutmish's personal slaves, who came to be known as the Forty (*chihilgān*). A political faction whose membership was characterized by talent and by loyalty to the family of Iltutmish, the Forty managed to fend off or absorb claimants to the throne from other lines. Meanwhile they divided the kingdom among themselves. None was willing to grant supremacy to another of their group, and each sought to build his personal fortune and position; nevertheless, they all appear to have realized that the future of the political system in which they participated depended upon a policy of cooperation in the face of a continued Mongol threat and the danger of revived Rājput power.

Rise of the Forty

The political situation had changed by 1246, when Ghiyās-ud-Dīn Balban, a junior member of the Forty, had gained enough power to attain a controlling position within the administration of the newest sultan, Nāṣir-ud-Dīn Maḥmūd (reigned 1246–66). Balban, acting first as *nāʾib* ("deputy") to the Sultan and later as sultan (reigned 1266–87), was the most important political figure of his time. The period was characterized by almost continuous struggles to maintain Delhi's position against the revived power of the Hindu chiefs (principally Rājputs) and by vigilance against the strife-ridden but still dangerous Mongols in the west. During the first 10 years of Maḥmūd's reign, Balban conducted major campaigns against five Rājput chiefs with only partial success. Ranthambhor could not be taken from the Cāhamānas, and, beginning in 1256, Mewāī raids first threatened the environs of Delhi and later actually penetrated the city. Thus, at least until 1266, the power of the sultanate was stagnant if not in decline. The Sultan's authority was challenged in several frontier regions by Hindu chiefs or discontented Muslim nobles, and even in central regions of the state sultanate rule sometimes was challenged by the nobility.

Balban's accession (1266) brought some improvement. He sought to raise the prestige of the institution of kingship through the use of ceremony, the strict administration of justice, and the formulation of a despotic view of the relationship between king and subject. The success of his policy owed much to the death or incapacity of most of the Forty and to the lack of rival claimants to the throne. Probably the most significant aspect of his reign was this elevation of the position of the sultan, which made possible the reorganization and strengthening of the army and the imposition of a tighter administrative apparatus, particularly in the area around Delhi.

Balban's military strategy

Balban's military strategy was to work outward from the capital. First, he cleared the forests of Mewātīs; then he restored order in the Doab and Oudh and suppressed a revolt in the Badaun and Amroha districts with particular viciousness. Having established the security of his home territory, Balban then chose to consolidate his rule over the provincial governors rather than to embark upon expeditions against Hindu territories. He argued that Turkish forces in India were insufficient to guard against the Mongol threat, man the existing administrative posts, and also conquer Rājput territories; but probably he saw also that it was in his own interests to strengthen his hold on the outlying territories already held by Muslims. Thus, he reacted vigorously and effectively against an attempt to establish an independent state in Bengal in the 1280s. Balban's policy of consolidation was a wise one; it strengthened sultanate rule and enhanced the power of the ruler so that his successors could undertake a number of successful expansionist campaigns after 1290.

Balban's concern over a shortage of manpower arose in part because the Mongols had severed Delhi's contact with Turkistan. Because few pureblood Turks could be recruited, Afghans and Indian-born Muslims had to be added to the administration and army. Although Balban was opposed to this admixture, necessity forced him to begin a practice that was to become much more significant in later years.

The Khaljīs. Balban's successors were unable to manage either the administration or the factional conflicts between the old Turkish nobility and the new forces, led by the Khaljīs; and, after a struggle between the two factions, Jalāl-ud-Dīn Fīrūz Khaljī assumed the sultanate in 1290. He immediately rewarded his supporters with offices and money and attempted to conciliate members of the old nobility. A number of them, as well as a large part of the populace of Delhi, however, continued to view him as a non-Turkish usurper. During his short reign (1290–96), Jalāl-ud-Dīn suppressed a revolt by some of Balban's officers, led an unsuccessful expedition against Ranthambhor, and defeated a substantial Mongol force on the banks of the river Sind. In 1296 he was assassinated by his ambitious nephew and successor, ʿAlāʾ-ud-Dīn Khaljī.

The transition to Khaljī rule was more than a dynastic change, because it represented the triumph of new blood and resulted in significant administrative and economic innovations. The view of medieval Muslim historians that the end of Balban's dynasty represented the end of the Turkish sultanate of Delhi is exaggerated, for the administration was neither so racially pure before 1290 nor so racially mixed thereafter as is sometimes stated. Nevertheless, the Khaljīs were not recognized by the older nobility as coming from pure Turkish stock (although they were Turks), and their rise to power was aided by impatient outsiders, some of them Indian-born Muslims, who might expect to enhance their positions if the hold of the followers of Balban and the Forty were broken. To some extent, then, the Khaljī usurpation was a move toward the recognition of a changing balance of power and thus

functioned as a cohesive force: those Muslims of talent who were outside the old ruling group were able to alter the system enough so that their efforts would be rewarded. Rebellion against the sultanate was not necessary.

The reign of ʿAlāʾ-ud-Dīn

By far the ablest of the Khaljīs, if not of all the sultans of Delhi, was ʿAlāʾ-ud-Dīn Khaljī (reigned 1296–1316). His career demonstrates many features common to ambitious and talented rulers. During his reign the sultanate briefly assumed the status of an empire. In order to achieve his goals of centralization and expansion, ʿAlāʾ-ud-Dīn needed money, a loyal and reasonably subservient nobility, and an efficient army under his personal control. He had earlier, in 1292, partly solved the problem of money when he conducted a lucrative raid into Bhīlsa in central India. Using that success to build his position and a fresh army, he led a brilliant and unauthorized raid on the fabulously wealthy Devagiri (modern Daulatābād), the capital of the Yādavas, in the Deccan early in 1296, with the idea that he could gain enough wealth to finance a coup. The wealth of Devagiri not only financed his usurpation but provided a good foundation for his state-building plans. ʿAlāʾ-ud-Dīn already had the support of many of the disaffected Turkish nobles, and now he was able to purchase the support of more with both money and promotion. He also was able to capture and subsequently to imprison, blind, or kill the sons of Jalāl-ud-Dīn and their supporters.

As sultan, ʿAlāʾ-ud-Dīn soon initiated a number of military, revenue, and economic reforms that were designed to raise more money, put a greater distance between the Sultan and his nobles, and to sink the roots of the sultanate more deeply into Indian soil. Many land grants and revenue assignments were resumed, and nobles were paid in cash whenever possible. The rate of taxation was increased officially to 50 percent of the produce, and officials were sent out to force local Hindu landholders to pay the revenue demand in full. For the first time in the history of the sultanate, taxation appeared as revenue demand based upon actual measurement of the land in many areas of North India rather than as a kind of tribute collected from local landholders. ʿAlāʾ-ud-Dīn also introduced a partly successful system of price and wage controls in the Delhi area so that the inflation caused by the sudden influx of both men and money following the organization of a standing army would devalue his treasury as little as possible. These measures were much more oppressive than any attempted by earlier sultans, and it is not entirely clear why ʿAlāʾ-ud-Dīn was able to enforce them as well as he did. His success must have been due in part to his military and administrative ability, but it also depended upon the creation of a substantial standing army—the first that the sultanate had known—which was directly under his command and which he kept at Delhi for extended periods.

Expansionist activities against Rājput chiefs

The result of ʿAlāʾ-ud-Dīn's reforms and his energetic rule was that the sultanate expanded rapidly and was subject to a more unified and efficient direction than during any other period. ʿAlāʾ-ud-Dīn began his expansionist activities with the subjugation of Gujarāt in 1299. Next he moved against Rājasthān and then captured Ranthambhor (1301), Chitor (1303), and Māndū (1305) and later added Siwāna (1308) and Jālor (1312). These conquests greatly reduced the power of the Rājput chiefs and made them, finally, subject to Delhi. In addition, the Rājput campaigns opened the road for further raids into South India.

These raids were not intended to result in occupation of the land but rather in the formal recognition by Hindu kings of ʿAlāʾ-ud-Dīn's supremacy and in the collection of huge amounts of tribute and booty, which were used to finance his centralizing activities in the north. In 1307 his lieutenant Malik Kāfūr again subdued the Yādava kingdom of Devagiri and in 1309 added the Kākatīya kingdom of Telingāna. In 1310–11 Malik Kāfūr plundered the Pāṇḍya kingdom in the far south, and in 1313 Devagiri was again defeated and finally annexed to the sultanate.

ʿAlāʾ-ud-Dīn also managed to fend off a series of Mongol attacks, which posed a much more serious threat than those of Balban's reign. Mongol invasions into the Punjab (1297–98) and into Sīwistān (1299) were repulsed, and, while another invasion almost reached Delhi in the same year and the Mongols actually invested the city in 1303, ʿAlāʾ-ud-Dīn finally was victorious on both occasions. Other invasions were repulsed in 1305 and 1306, after which the Mongol attacks subsided, probably as much because of an intensification of internal rivalries as of the lack of Mongol success in India.

Ambition, a talent for ruling, and the gold of South India carried ʿAlāʾ-ud-Dīn a long way, but it is also significant that he was one of the first rulers to deliberately expand political participation within the sultanate government. Not only did he partly open the gates to power for the non-Turkish Muslim nobility—some of whom were even converted Hindus—but he also at least made gestures toward the inclusion of Hindus within the political world he viewed as legitimate. Both ʿAlāʾ-ud-Dīn and his son married into the families of important Hindu rulers, and several such rulers were received at court and treated with respect. Like the Mughal ruler Akbar, 250 years later, ʿAlāʾ-ud-Dīn appears to have had the vision of an empire in which both Hindus and Muslims would participate within the ruling group, although both would be subject to the strict discipline of the ruler. ʿAlāʾ-ud-Dīn's centralizing activities also led, however, to a good deal of oppression of both Hindus and Muslims.

Succession dispute

ʿAlāʾ-ud-Dīn's success was not lasting, nor did the sultanate remain in his family for long. The succession dispute following his death in 1316 resulted in the murder of Malik Kāfūr by the palace guards and in the blinding of ʿAlāʾ-ud-Dīn's six-year-old son by Quṭb-ud-Dīn Mubārak Shāh, the Sultan's third son, who assumed the sultanate (1316–20). During Quṭb-ud-Dīn's reign, revolts in Gujarāt and Devagiri were suppressed, and Telingāna was raided again. Quṭb-ud-Dīn was murdered by his favourite general, a Hindu convert named Khusraw Khān, who had built substantial support among a group of Hindus who were not part of the traditional nobility. Opposition to Khusraw's rule arose immediately, led by Ghāzī Malik, the warden of the western marches at Deopalpur; and Khusraw was defeated and slain after four months. Ghāzī Malik was probably able to gain enough support to overthrow Khusraw because of his own military reputation and because of Khusraw's position as an outsider. It is unlikely, however, that Khusraw's Hindu origin was of major significance (for many Muslims supported him). Historians of the succeeding Tughluq period, however—perhaps in justification of the Tughluq usurpation—made much of the claim that Ghāzī Malik saved the sultanate from domination by infidels.

The Tughluqs. Ghāzī Malik, who ascended the throne as Ghiyāṡ-ud-Dīn Tughluq (reigned 1320–25), had first risen to prominence during the reign of Jalāl-ud-Dīn Khaljī and had later distinguished himself by his successful defense of the frontier against the Mongols. His reign was short but successful. He sent an expedition against Telingāna in 1321–22, captured it in 1323, and then raided into Jājnagar. In addition, another Mongol raid was repulsed, and Bengal, which had been independent under Muslim kings since the death of Balban, was again brought under sultanate rule. While returning from the Bengal campaign, the Sultan was killed when a wooden shelter collapsed on him at Afghānpur, near Delhi. Although some historians have argued that Muḥammad ibn Tughluq plotted his father's death, the case never has been proved.

The reign (1325–51) of Muḥammad ibn Tughluq marked both the high point and the beginning of the decline of the sultanate. The period from 1296 to 1335 can be seen as one of nearly continuous centralization and expansion. A contemporary historian counted 23 provinces at the beginning of Muḥammad ibn Tughluq's reign, and there were few places in the subcontinent where the Sultan's authority could be seriously challenged. Even before 1335, however, things began to go wrong, and, while the sultanate enjoyed its greatest geographical extension during his reign, Muḥammad ibn Tughluq was unable to maintain the momentum of consolidation. By 1351 South India had been lost and much of the North was in rebellion.

Muḥammad ibn Tughluq was an innovative and ambitious ruler; yet his innovations and ambitions appear to have worked to his disadvantage. Like the Mughal ruler

Aurangzeb, more than three centuries later, Muḥammad ibn Tughluq was confronted by problems resulting from expansion into South India, which put him in an untenable military and political position. And, like Aurangzeb, he tried to hold his territories in the Deccan and thus paid for his ambition with a weakened hold on the North and the ultimate independence of the South. Eschewing the Khaljī policy of maintaining Hindu tributary states in the South, Muḥammad ibn Tughluq, while still a prince, had brought two of the three remaining southern Hindu powers under the direct control of the sultanate, and in 1326–27 he subdued the third (Kampili). Direct Muslim rule in the South, however, did not necessarily signify control from Delhi. In an effort both to settle additional Muslim nobles in the South and to maintain his control over them, the Sultan made Daulatābād (Devagiri) his second capital in 1327 and forcibly moved a significant proportion of the unhappy inhabitants of Delhi there. The extent to which Delhi was depopulated by the move is questionable, however, and perhaps the severe famine of 1335–36 was as significant a cause of the city's temporary decline.

No sooner was the Sultan established at Daulatābād than trouble broke out in the North. Rebellions by Kishlū Khān on the western border and by Nāṣir-ud-Dīn in Bengal had to be put down in 1327–28. The Sultan was less successful against an invasion by the Mongols, however, and was obliged to pay a large tribute after they had come almost to the gates of Delhi. Perhaps in reaction to this invasion, Muḥammad ibn Tughluq raised an army reported to number 370,000 men and said to have been destined for a campaign into Khorāsān, from whence the Mongols had come. It is more likely that the army's object was Ghazna or Peshāwar, but in any case the troops were never used and were disbanded after being paid for a year, thus depleting the treasury. Another attempt to extend the boundaries of the sultanate some years later—this time in the Kumaon region of the Himalayan foothills—was actually carried out and resulted in the loss of almost an entire army through the ravages of disease and the guerrilla tactics of the hill people.

Loss of the South

By 1335 it was evident that Muḥammad ibn Tughluq's attempt to hold the South was not likely to succeed. The Muslim governor of Ma'bar, the southernmost province of the sultanate, declared his independence in 1334–35 and founded the sultanate of Madura while Muḥammad ibn Tughluq was busy quelling a rebellion in Lahore. Soon rebellions by Hindu chiefs had resulted in the formation of several new states, the most important of which was Vijayanagar. During the next few years, while the Sultan was kept moving from north to south in an attempt to put down rebellions in practically every province, he lost control of the rest of his South Indian possessions after successful rebellions in Gulbarga (1339), Warangal (1345–46), and Daulatābād, which led to the founding of the Bahmanī kingdom (1347). Muḥammad ibn Tughluq spent the last five years of his life trying to suppress yet another rebellion in Gujarāt and thus could not make an attempt to regain Daulatābād.

Muḥammad ibn Tughluq's successor was his cousin Fīrūz ibn Rajab, thereafter known as Fīrūz Shāh (reigned 1351–88). Although this son of a Rājput princess has been depicted by medieval chroniclers as a pious ruler whose reign saw peace and prosperity, he still campaigned in Bengal (1353–54 and 1359), Orissa (1360), Nagarkot (1361), Sind (1362 and 1366–67), Etāwah (1377), and Katehr (1380). Despite these efforts, Fīrūz was unable to recover Bengal for the sultanate, and Sind was no more than a tribute-paying vassal during his reign. Fīrūz's decision (*c.* 1365) not to accept an invitation from Bahrām, a Bahmanī prince, to intervene in Bahmanī politics and reconquer the Deccan was probably a wise one. Fīrūz had witnessed the troubles of his cousin and must have seen how difficult if not impossible it would be for even the most talented and energetic of rulers (which he was not) to hold the South and North simultaneously.

The principal reason for Fīrūz's good reputation among contemporary chroniclers was probably his conciliatory attitude toward the two main influential Muslim groups of the period: the Muslim religious leaders and the nobility. While 'Alā'-ud-Dīn Khaljī had kept religion and religious leaders apart from his political plans and Muḥammad ibn Tughluq had incurred the enmity of at least some Ṣūfīs because of his refusal to give them what they regarded as proper support, Fīrūz rewarded Ṣūfīs and other religious leaders generously and listened to their counsel. He was indeed a pious ruler from the Muslim viewpoint: he created charities to aid poor Muslims, built many colleges and mosques, abolished taxes not recognized by Muslim law, made largely unsuccessful attempts to convert his Hindu subjects, and sometimes persecuted them.

His weakness as a ruler was politically more significant than was his piety, however, and his weak policy toward his nobility was an important reason for the decline of central authority during his reign.

Decline of central authority

Balban, 'Alā'-ud-Dīn, and Muḥammad ibn Tughluq all had made attempts to check the power of the nobility and to centralize authority; the latter two had also realized the necessity for allowing a certain amount of mobility both into and within the army and civil administration for groups that had come to represent significant and articulated interests. The Khaljī revolution had brought Muslim outsiders into positions formerly enjoyed almost exclusively by Balban and his followers, and during the first 30 years of Tughluq rule the gradual integration of new blood into the nobility had continued. Such a policy also enhanced the power of the sultans over all of the nobility because it unsettled old nobles and provided grateful new ones. Judging by the revolts during his reign, however, Muḥammad ibn Tughluq's policy toward his nobility was too autocratic to succeed. Partly for that reason and partly because of his own proclivities, Fīrūz adopted policies that gave his nobles much more autonomy. He agreed that fewer new nobles of questionable heritage (*i.e.*, Indian Muslims and Hindu converts) should be allowed to have important positions. The result was that the Sultan lost both an important means of leverage and a means of adjusting to new political circumstances. Fīrūz also made little or no attempt to pay officers in cash (rather than in assignments of land revenue), granted hereditary appointments, and extended the system of revenue farming. All of these measures, which reversed policies adopted by one or more of the strong rulers of the previous 50 years, tended to decrease Fīrūz's control over his nobility and over the revenue system. Finally, contemporary chronicles suggest that Fīrūz's generally light hand on the reins of administration led to increasing inefficiency and corruption, especially after the death of his very competent vizier, Khān Jahān I, in 1370.

Fīrūz was justly famous as a builder, however. In addition to a large number of mosques, schools, hospitals, caravansaries, and even towns, he constructed five canals for irrigation, the most important of which ran for 150 miles (240 kilometres), and a number of reservoirs and wells for the same purpose.

Decline of the sultanate. Succession disputes were the rule rather than the exception in the Muslim kingdoms of India, but the death of an old ruler after a long reign often led to more than normally destructive disputes because there could be two generations of grown men with claims to the throne. Such a situation arose after the death of Fīrūz, in 1388, and it was further complicated by the relatively great power of the nobility and by the absence of any claimant who exhibited a superior degree of talent. The sons and grandsons of Fīrūz, supported by various groups of nobles, began a struggle for the throne that rapidly diminished the authority of the sultanate and provided opportunities for Muslim nobles and Hindu chiefs to enhance their autonomy. By 1390 the Muslim governor of Gujarāt had declared his independence, and between 1391 and 1394 the important Rājput chiefs of Etāwah rebelled and were defeated four times. Clearly, the authority of Delhi was in doubt. By 1394 there were two sultans, both residing in or near Delhi. The result was bitter civil war for three years; meanwhile the disastrous invasion of Timur (the Tamerlane of Western literature) drew nearer.

Timur was a Barlas Turk who rose from an uncertain position as ruler of a small principality in Transoxania

Timur's invasion of India

to become one of the greatest conquerors the world has known. Timur invaded India in 1398, when he was 62 years old and in possession of a vast empire in the Middle East and Central Asia, and dealt the final blow to the effective power of the Tughluq dynasty and of the sultanate itself. In a well-executed campaign of four months—during which many of the disunited Muslim and Hindu forces of North India either were bypassed or submitted peacefully while Rājputs and Muslims fighting together were slaughtered at Bhatnagar—Timur reached Delhi and, in mid-December, defeated the army of Sultan Maḥmūd, and sacked the city. It is said that Timur ordered the execution of at least 50,000 captives before the battle for Delhi and that the sack of the city was so devastating that practically everything of value was removed—including those inhabitants who were not killed.

After this catastrophe, Delhi became merely one of the regional principalities of North India, competing with the emerging Rājput and Muslim states. After Timur's invasion, Gujarāt, Mālwa, and Jaunpur soon became powerful independent states, old and new Rājput states rapidly emerged, and Lahore, Dīpālpur, Multān, and parts of Sind were held by Khiẓr Khān Sayyid for Timur (and later for himself). In this unstable political situation, many smaller states also began to emerge.

Rise of regional states

During the 15th and early 16th centuries, no paramount power enjoyed effective control over most of North India and Bengal, although the Lodī sultans of Delhi gradually asserted authority over much of the Gangetic Plain and the Punjab. The lack of unified rule has led some historians to describe the period as one of political anarchy and confusion, in which the inhabitants suffered because there was no strong guiding hand. Such a conclusion is far from certain, however, even for the central areas of the Gangetic Plain where many battles were fought. In areas where effective regional rule either was restored or developed—as in Rājasthān, Orissa, Bengal, Gujarāt, Mālwa, Jaunpur, and various smaller states in the North, as well as in the large and small states of the Deccan—the quality of life may well have been as good or better for cultivators, townspeople, landholders, and nobles than it had been in earlier centuries. Although contemporary sources are scarce, the information available does not indicate a significant decline in total cultivation or trade (despite some alteration of trade routes), and various regional histories suggest that the vigour of local political organization was maintained if not enhanced.

The Sayyids. The last of the Tughluqs died in 1413, and Delhi and a small area surrounding it fell into the hands of the Turk Khiẓr Khān in 1414. The Sayyid dynasty ruled the territory until 1451, trying to obtain tribute and recognition of suzerainty from the nearby Rājput rulers and fighting almost continuously against neighbouring states to preserve their kingdom intact. The last Sayyid ruler, ʿAlāʾ-ud-Dīn ʿĀlam Shāh (reigned 1445–51), peacefully surrendered Delhi to his nominal vassal, the Afghan Bahlūl Lodī, and retired to Badaun district, which he retained until his death in 1478. Meanwhile, the neighbouring kingdom of Jaunpur developed into a power equal to Delhi during the reign (1402–40) of Ibrāhīm Sharqī. Ibrāhīm's successor, Maḥmūd, conducted expansionist campaigns against Bengal and Orissa and, in 1452, initiated a conflict with the Lodī sultans of Delhi that lasted at least until the defeat and partial annexation of Jaunpur by Bahlūl Lodī in 1479.

The Lodīs. Before he moved to Delhi, Bahlūl Lodī had already carved out a kingdom in the Punjab that was larger than that of the Sayyid sultans. During the next 75 years Bahlūl and his two successors, Sikandar (reigned 1489–1517) and Ibrāhīm (reigned 1517–26), continued to hold Delhi and intermittently to expand their control over the surrounding territory. Bahlūl pacified the Doab and subdued Etāwah, Chandwar, and Rewāri. Sikandar completed the pacification of Jaunpur (1493), campaigned into Bihār, and founded the city of Āgra in 1504 as a base from which to launch his ultimately unsuccessful attempt to control nearby Rājput states. By the time of Sikandar's death, the Afghans could claim a somewhat uneven control over the Punjab and most of the Gangetic Plain down to Bihār. Still, the question of Lodī hegemony in North India was far from settled, for it can be argued that the strongest military force in India on the eve of the Mughal conquest was that of the Rājput confederacy, under Rānā Saṅa (Sanga), rather than that of the Lodīs. A significant result of the Lodī state's development was the immigration into India of a considerable number of Afghans, who settled in and helped to expand the sultanate and who were the principal rivals to the new Mughal power through most of the 16th century.

Immigration of Afghans

The last Lodī sultan, Ibrāhīm, was more autocratic than his predecessor, and he was ultimately less able to control his skittish nobility, who tended to see the Lodī sultans as first among equals only. He soon faced an Afghan rebellion in the east under the leadership of his brother Jalāl Khān, and, while Ibrāhīm put down this and some other Afghan revolts in the east, the groundwork for the final disaster was laid in the west. Dawlat Khān Lodī, governor of the Punjab, and ʿĀlam Khān Lodī, Ibrāhīm's uncle, appealed to Bābur, the Mughal ruler of Kābul, to aid them in their attempt to overthrow the Sultan. The adventurous Bābur was at that time probably thinking only of annexing the Punjab, but, as his previous history had demonstrated, he was quick to take advantage of political opportunities. In 1524 he led an expedition to Lahore and defeated Ibrāhīm's army. Bābur then passed over his Afghan allies and appointed his own officials in the Punjab. After his allies had indignantly left him, he went on to defeat and kill Ibrāhīm at Pānīpat, near Delhi, in 1526 (see below, *The Mughal Empire*). The Afghan sultanate underwent a short revival under the Sūrs in 1540–55, only to be replaced by the Mughals again under Humāyūn and then Akbar.

Appraisal of the Delhi sultanate. In many ways the sultanate of Delhi can be seen as a continuation of the Hindu political tradition of the preceding period. Except for a brief time under ʿAlāʾ-ud-Dīn Khaljī and the early Tughluqs, the sultanate appeared as one of a number of powers—though usually the greatest—that competed for the paramount position in North India. Although most of the upper level of Hindu chiefs in the plains areas were removed, the Rājput chiefs in the hills and forests were never entirely subdued, and they returned repeatedly to challenge the sultanate. Even in the plains areas, however, it is likely that only a small proportion of the land was directly administered by Muslims, while the rest was left in the hands of Hindu petty chiefs and landholders, who paid what amounted to tribute rather than a tax based upon measurement of the land, except during ʿAlāʾ-ud-Dīn's reign. Even under ʿAlāʾ-ud-Dīn, the area under centralized revenue control was not very large, and Muḥammad ibn Tughluq's attempts to increase the rate of taxation within the area and to expand it helped to ruin agriculture in the Doab and were partly responsible for the revolts of his nobles. Despite these disclaimers, however, the formation of a Muslim political system designed for centralization and the introduction of an Islāmic ruling group and of Muslim culture were extremely important developments in the history of the subcontinent.

THE MUSLIM STATES OF SOUTH INDIA, C. 1350–1680

Sultanate rule in most of South India existed for only a few years and was firmly established only in the Deccan's northernmost province, whose capital was Daulatābād. The forced withdrawal of the sultanate forces from the Deccan between 1330 and 1347 was partly the result of resistance offered by Hindu chiefs and some Muslim nobles. Members of those two groups then established the two strongest states of the South—the Muslim-ruled Bahmanī kingdom and the Hindu-ruled Vijayanagar empire.

The Bahmanī dynasty. The Bahmanī kingdom, which lasted from 1347 to roughly 1527, arose out of a revolt by a group of Muslim nobles against the Delhi sultan, Muḥammad ibn Tughluq, in 1345. The founder of the Bahmanī dynasty, Ḥasan Gaṅgū, ascended the throne at Daulatābād in 1347 as ʿAlāʾ-ud-Dīn Bahman Shāh and soon moved his capital to Gulbarga, centrally located on the Deccan Plateau, where it remained until about 1423. The natural focus for the Bahmanī state was the Deccan

Revolt by Muslim nobles

Plateau tableland. It was bounded on the north by the Vindhya Range and Mālwa; on the west by the Western Ghāts and the steep drop to the coastal plain beyond; on the south by the Tungabhadra River, beyond which lay the much disputed no-man's-land, called the doab, between the Krishna and Tungabhadra rivers, where so many battles were fought with the rulers of the kingdom of Vijayanagar; and on the east, less definitively, by the lower end of the gradually sloping plateau. Much of the political and military history of the Bahmanī sultanate can be described as an attempt to gain effective control of the Deccan Plateau and then, less successfully, to expand outward from it. The period of consolidation was followed by a much longer period of intermittent warfare against Mālwa and Gujarāt in the north, Vijayanagar in the south, and Orissa in the east.

Bahmanī consolidation of the Deccan. Bahman Shāh spent most of his reign consolidating a kingdom in the area that had formerly been subject to the Delhi sultanate government at Daulatābād and in strengthening his hold over those Muslim nobles who chose to remain in the Deccan rather than to join Muḥammad ibn Tughluq in North India. He adopted the four territorial divisions established by Muḥammad ibn Tughluq for his own administration and established departments and appointed officials similar to those of the Delhi sultanate. Working outward from his capital, he was able to establish his authority over the western half of the Deccan Plateau and to impose an annual tribute upon the Hindu state of Warangal, which had also emerged from the breakup of the Deccan portion of the Tughluq empire. Often, however, the tribute was not paid, and a number of wars were fought over the question of whether the Bahmanīs could maintain a superior position in relation to their eastern neighbour in the following years.

Bahman Shāh was succeeded by his son Muḥammad Shāh I (reigned 1358–75), who began the struggle with Vijayanagar that was to outlast the Bahmanī sultanate and continue, as a many-sided conflict, into the 17th century. The period 1350–1500 saw at least 10 wars, most of which were concerned with control over the Tungabhadra-Krishna doab. The doab had been an area of contention long before the foundation of either the Bahmanī kingdom or Vijayanagar. During Muḥammad Shāh's reign institutional and geographical consolidation proceeded to the point where a solid foundation for the kingdom was established. In his wars against Vijayanagar and Telingāna (Warangal), Muḥammad Shāh effectively made use of newly organized artillery to defeat an army much larger than his own. His two wars with Vijayanagar gained him little, but his attack on Telingāna in 1363 gained him a large indemnity plus possession of the town of Golconda and its dependencies; and in 1365 his rapid response to a rebellion by the governor of Daulatābād and some Marāṭhā and other chieftains of Berār and Bāglanā led to a quick victory. The Sultan was spared further internal or external threats to his kingdom, and he spent the last 10 years of his reign consolidating his hold over the territories in his possession. His well-laid foundations were soon disturbed, however, when his son and successor, ʿAlāʾ-ud-Dīn Mujāhid (reigned 1375–78), was assassinated by his cousin Dāʾūd while returning from a campaign in Vijayanagar. Dāʾūd was in turn murdered by ʿAlāʾ-ud-Dīn's partisans, who then set Dāʾūd's brother Muḥammad II (reigned 1378–97) on the throne and blinded Dāʾūd's son. These political difficulties enabled Vijayanagar to take away part of Goa and other territory along the western coast; but the rest of Muḥammad II's reign was peaceful, and the Sultan spent much of his time building his court as a centre of culture and learning.

Political and cultural tendencies

Several political and cultural tendencies were becoming noticeable at this time and had significant effects on the development of the Bahmanī state and its successors. Although the state had been organized by a group of dissident nobles from the Delhi sultanate, differences in both the culture and the membership of the nobilities developed, largely because of differences in recruiting patterns. Soon after the foundation of the Bahmanī state, large numbers of Arabs, Turks, and particularly Persians began to immigrate to the Deccan, many of them at the invitation of Sultan Muḥammad I, and there they had a strong influence on the development of Muslim culture during subsequent generations. The new settlers (*āfāqīs*) also had a political effect, as they soon began competing successfully for important positions within the political hierarchy. The original rebels from the Delhi sultanate and their descendants, who came to be called *dakhnīs* (*i.e.,* Deccanis—from the Deccan), thought of themselves as the old nobility and thus resented the success of the newcomers. The situation was not unlike that of the Delhi sultanate, in which a party of established nobles had tried to protect their privileged position against newcomers who were developing claims to power. Thus, the distribution of high offices among Persian newcomers by Sultan Ghiyās-ud-Dīn (Muḥammad II's oldest son, who ruled for about two months) in 1397 was seen as a threat by the old nobles and Turks and was probably a major reason for his assassination. Later, the addition of Hindu converts and Hindus to the nobility complicated the situation further, as it had in the north, but the division between *dakhnīs* and *āfāqīs* (hereafter called newcomers) was most significant and contributed to the disintegration of the Bahmanī state.

Muḥammad II's peaceful reign was followed by a year of succession disputes caused both by party conflicts and by dynastic rivalries. When Muḥammad's cousins Aḥmad and Fīrūz finally gained control, Fīrūz succeeded as Tāj-ud-Dīn Fīrūz. His reign (1397–1422) was a period of notable cultural activity in the Bahmanī sultanate, as well as one of continued development of the trend toward wider political participation. Noted for his intelligence and learning, Fīrūz nurtured the greatest centre of Muslim culture in India at a time when the Delhi sultanate was rapidly dissolving. Perhaps in an effort to balance the continuing influx of Persians, as well as to strengthen his own position as a ruler who was above all the nobles and who recognized the realities of political power, Fīrūz gave a number of high offices to Hindus (Brāhmaṇs) and married several Hindu women, including the daughter of the King of Vijayanagar. Thus, the parallel with the earlier development of the Delhi sultanate nobility continued. The fact that Hindus were becoming politically more significant at a time when the military rivalry with Vijayanagar was renewed suggests a political rather than a religious motivation for that rivalry.

Fīrūz stopped an invasion in the north by the Gond raja of Kherla in Madhya Pradesh and conducted two moderately successful campaigns against Vijayanagar. The first brought him a tribute payment and temporary possession of part of the Andhra country, while the second ended with his marriage to the Vijayanagar king's daughter, another tribute payment, and the establishment of an apparently amicable relationship between the two rulers. The peace lasted for only 10 years, however, and a third war (1417–20) ended in a disastrous defeat for Fīrūz by the united forces of Vijayanagar and Fīrūz's former allies, the Velama faction of the Reddi ruling group in Andhra. Although no territory was lost, the defeat so weakened Fīrūz's position that, when the question was raised of whether Fīrūz's son or his brother Aḥmad should succeed him, most of the army supported Aḥmad; and Fīrūz, who was 70 years old and dying, was forced to abdicate in favour of him.

One of the first acts of the new sultan, Shihāb-ud-Dīn Aḥmad I (reigned 1422–36), was to move the capital from Gulbarga to Bīdar, which was surrounded by more fertile ground and had become more centrally located, now that some territory had been gained to the southeast, in Telingāna. Perhaps, also, the move signified Aḥmad's expansionist ambitions, for in 1425 he defeated and killed the Velema ruler of Warangal and finally annexed most of Telingāna, bringing his eastern border to the edge of Orissa. During the next decade, however, rebellions forced Aḥmad to allow local chieftains to rule as tributaries throughout much of the area.

Expansionist ambitions of Aḥmad I

External and internal rivalries. Although the Bahmanī state had been threatened from the north earlier, it was during Aḥmad's reign that conflicts first broke out with the northern neighbours, Mālwa and Gujarāt. The break-

down of centralized authority within the Delhi sultanate and the consequent rise of provincial kingdoms meant that new rivalries could develop on a regional basis, and the Bahmanī sultans found themselves contending with two of the successor states of the Delhi sultanate in an arena where their expansionist ambitions had some chance of success. A border dispute with Mālwa led to a Bahmanī victory and a short-lived recognition of the chieftainship of Kherla as a Bahmanī protectorate. Aḥmad then forged an alliance with another northern neighbour, Khāndesh, and, on the pretext of giving aid to a Hindu chieftain who had revolted against Gujarāt, he sent unsuccessful expeditions into Gujarāt in 1429 and 1430. The latter defeat was especially significant; it partly stemmed from rivalries between the Deccani officers and the newcomers from the Middle East, a friction that appears to have become gradually more intense from this point until the breakup of the Bahmanī sultanate.

During the last year of his reign, Aḥmad named his eldest son as his heir and gave him full charge of the administration; he parcelled out the provinces among his other sons, exacting from them promises that they would be loyal to the new ruler. Perhaps the scheme was useful; even though Sultan ʿAlāʾ-ud-Dīn Aḥmad II (reigned 1436–58) had to face a rebellion by one of his brothers, a precedent was set for a rule of primogeniture, which seemed to alleviate the problem of succession disputes for the rest of the century. Unfortunately for later Bahmanī rulers, rivalries among the nobility were to prove just as detrimental to the fortunes of the dynasty as family disputes were in many other dynasties of the period.

Aḥmad II proved to be a weaker ruler than his father had been, and during his reign the conflicts among the nobles intensified. Two short wars with Vijayanagar in 1436 and 1443–44 were confined to Krishna-Tungabhadra doab and signified little except the arrival of a new power, the Hindu Gajapati king of Orissa, who allied himself with the Bahmanī ruler in the second campaign. Perhaps more significant in its ultimate effect was the Bahmanī victory over Khāndesh in 1438. The force in that campaign was composed exclusively of newcomers, who had convinced the Sultan that Deccani treachery had been responsible for the defeat in Gujarāt in 1430. The newcomers thereby gained considerable influence with the Sultan but at the same time intensified the resentment of the Deccanis, who retaliated in 1446 by massacring a large number of them, with the malleable Sultan's tacit permission. Later, when the Sultan was convinced that the newcomers had been unjustly killed, he punished many of the responsible Deccanis and promoted the surviving newcomers. During the last years of his reign, Aḥmad had to face a rebellion in Telingāna led by his son-in-law, the most noteworthy aspect of which was the military and diplomatic skill displayed by a newly arrived noble, Maḥmūd Gāwān, who had persuaded the rebels to desist and the Sultan to pardon them.

Three sultans reigned between 1458 and 1481, but the most notable personality of the period was the Persian émigré Maḥmūd Gāwān, who was a leading administrator during the reigns of Humāyūn (reigned 1458–61) and his son Niẓām-ud-Dīn Aḥmad III (reigned 1461–63) and was chief minister under Muḥammad III (reigned 1463–82). During Maḥmūd Gāwān's ascendancy, the Bahmanī state achieved both its greatest size and greatest degree of centralization, and yet, partly because of the attempts at centralization and partly because of the continuing rivalry between the Deccanis and newcomers, the period ended with Maḥmūd Gāwān's assassination and the rapid dissolution of the effective power of the Bahmanī state.

Military successes in the west

Militarily, the Bahmanī forces usually were successful. After Maḥmūd Gāwān's installation as chief minister in 1463, a series of expansionist campaigns resulted in the subjugation in the west of most of the Konkan, including the important port of Goa, and the Bombay Carnatic. The frontier with Mālwa in the north was maintained more or less as it was, while an alliance with Vijayanagar proved effective in defeating Orissa in 1470. A later campaign in the east produced some minor advantages for the Bahmanīs in Telingāna.

As chief minister, Maḥmūd Gāwān attempted to enhance the power of the crown through a series of administrative reforms and political manoeuvres. Up to the 1470s the kingdom had been divided into four provinces, centring around the cities of Daulatābād, Māhūr, Bīdar, and Gulbarga, respectively. The governors of the four provinces had control over almost all aspects of civil and military administration within their territorial jurisdictions. Administration was thus decentralized from the beginning, but the relative power of the provincial governors as compared with the centre potentially became even greater as the state expanded and each of the four provinces grew larger. To decrease the power of the governors, Maḥmūd Gāwān created eight provinces in place of the four, reduced the military control of the governors by making the commanders of all forts but one in each province responsible directly to the Sultan instead of to the governor, and tightened central control over the employment and payment of troops within the provinces. In addition, he introduced a system of measurement and valuation of agricultural land and created a large block of crown land within each province. Maḥmūd Gāwān also tried to balance important appointments between Deccanis and newcomers—a policy that suggests not only his desire to reduce disputes among the nobility but perhaps also his intention to enhance the power of the chief minister, who, although a newcomer, had raised himself above party conflicts.

Unfortunately for Maḥmūd Gāwān and for the Bahmanī dynasty, party strife had developed to such an extent that a group of Deccani nobles who were motivated by hostility toward the chief minister as a newcomer, as well as by dislike of his efforts toward centralization, falsified evidence to make Maḥmūd Gāwān appear a traitor and convinced Muḥammad III to execute him in 1481. All the newcomers and some of the Deccani nobles disapproved of the execution and sided with Yūsuf ʿĀdil Khān, previously Maḥmūd Gāwān's chief supporter. Most of the newcomers returned to their provinces and refused to come to the capital, and the Sultan was left with only the support of the conspirators. When he died in 1482 (of grief over his error in judgment, the chronicles report), the leader of the conspirators, Malik Nāʾib, was able to make himself regent for Muḥammad's minor son, Shihāb-ud-Dīn Maḥmūd (reigned 1482–1518).

Bahmanī decline. Maḥmūd's reign began with an abortive attempt to assassinate Yūsuf ʿĀdil Khān, which resulted in the Khān's agreement to retire to Bijāpur and leave Malik Nāʾib and the conspirators in charge at Bīdar. Now the lack of institutionalized central power and the detrimental effects of the conflict between Deccanis and newcomers rapidly became more apparent. Malik Nāʾib, never popular even with a number of the Deccanis, was put to death in 1486 by the Abyssinian governor of Bīdar, and the Sultan subsequently began to rely on the newcomers for support. An attempt on Maḥmūd's life in 1487 by a group of Deccanis strengthened the Sultan's reliance on the newcomers and led to the slaughter of a great many Deccanis. But by this time it began to become apparent that the power of the Sultan was less than that of several of his nobles, and, although he continued to be a valuable pawn for the provincial governors to try to control, his power to rule was nearly gone. The provincial governors and their followers could not be controlled, nor did they believe that maintaining the centralized Bahmanī state would any longer be in their best interests. Consequently, the governors were usually unwilling to aid the Sultan when he attempted to put down rebellions by other governors or by powerful nobles.

Rebellions by governors and provincial nobles

One of the first revolts was that of a Turkish noble, Qāsim Barīd, who defeated the army sent against him by the Sultan and then forced Maḥmūd to make him chief minister of the state. Qāsim Barīd's attempt to reimpose central authority was opposed by most of the chief nobles, however, who defeated him once and then refused to recognize his authority. Next, Malik Aḥmad Niẓām-ul-Mulk, the son of Malik Nāʾib, began to carve out a territory for himself by conquering Marāṭhā forts along the western coast. He defeated the two armies sent against him by the

Sultan, whom he forced to recognize his conquests, and in 1490 he assumed a practical independence and established his capital at Ahmadnagar. Yūsuf ʿĀdil Khān of Bijāpur and Fatḥ Allāh ʿImād-ul-Mulk of Berār had demonstrated their sympathy for Malik Aḥmad's activities and soon emulated him. Although the three governors still did not assume the insignia of royalty, it was clear by the end of 1490 that Sultan Maḥmūd and the chief minister, Qāsim Barīd, could not command any of them.

Successors to the Bahmanī. During the 1490s the rivalries intensified among the former provincial governors, other high nobles, and Qāsim Barīd, who was the effective head of the government at the Bahmanī capital. Each began to form temporary alliances and to fight battles with other nobles in order to enhance his own position. Gradually the five successor states to the Bahmanī sultanate took shape, as lesser nobles were defeated and their territories were incorporated by the provincial governors or retained by Bīdar. Bijāpur, Berār, and Ahmadnagar were joined by Golconda in 1512, when Quṭb-ul-Mulk of Golconda founded the Quṭb Shāhī dynasty. Although a Bahmanī sultan still remained as a puppet ruler until at least 1538, effective control of the Bīdar government passed into the hands of Qasīm Barīd's son Amīr Barīd upon his father's death in 1505, thus establishing what proved to be a dynastic claim for the Barīd Shāhīs of Bīdar.

Ironically, the conflict between Deccanis and newcomers, which had done so much to destroy the unity of the sultanate, was of little importance after 1492. The major rivalry of the next decade was between two newcomers, Qāsim Barīd and Yūsuf ʿĀdil Khān. (Qāsim Barīd, however, was supported by the Deccanis of Bīdar in his struggle with another Deccani, Malik Aḥmad of Ahmadnagar.) The shift resulted from the fact that there were no longer parties of nobles but rather semi-independent states whose rulers were attempting to establish and expand their authority. With territorial integrity the primary goal, alliances with any semi-independent rulers outside the territory who might threaten its sovereignty necessarily fell into the field of foreign relations, and they tended to assume the expedient quality of most such relationships.

Unity against Vijayanagar

One issue that occasionally united the Bahmanī successor states was the desire to profit at the expense of Vijayanagar. Perhaps in an attempt to regain some of his lost authority and with the idea that spoils and possibly territory could be gained from raids on Vijayanagar, which was suffering from disunity and weak central control, Sultan Maḥmūd II proposed in 1501 that a policy of an annual *jihād,* or holy war, against the Hindu kingdom be adopted by the Muslim nobles. A number of relatively successful raids were undertaken during the next few years, but in 1509 the new ruler of Vijayanagar, Krishna Deva Rāya, repulsed the Muslims with substantial losses. Vijayanagar then grew stronger, and alliances against it increasingly became a matter of necessity, although there were many occasions when a Muslim king would ally with Vijayanagar against another Muslim. Ultimately, the Hindu kingdom fell victim to the expansionist ambitions of Bijāpur and Golconda, when these kingdoms took advantage of Vijayanagar's defeat by an alliance of Muslim armies in 1565 at Tālikota.

In the struggle for control of the Deccan after the breakup of the Bahmanī sultanate, the two southernmost of the successor states, Bijāpur and Golconda, ultimately found themselves in the most advantageous position because they were farthest away from the growing power of the Mughal Empire in North India. During the latter part of the reign of the emperor Akbar (reigned 1556–1605), the Mughals began a southward movement that was to end with the annexation of Bijāpur (1686) and Golconda (1687). During the intervening period, the Mughal presence became increasingly important to the remaining Deccan kings, who struggled to maintain or expand their position within the Deccan while fending off the advancing Mughal juggernaut.

Balance of power alliances. During the 16th century the strongest and best organized of the Bahmanī successor states was Ahmadnagar, followed by Bijāpur and then Golconda. All three were much larger and more important than Berār and Bīdar, and all three either began with or soon came to accept the Shīʿah form of Islām (the religion of the Persian newcomers) as the official faith of their rulers. During the 16th century the three major states formed shifting patterns of alliances against each other, which sometimes also included Vijayanagar, while the two smaller Muslim states ranged themselves on one side or the other in order to protect their independence. The goal of military campaigns normally was to humble the adversary without doing irreparable harm, for all three major Muslim states feared the supremacy of any one state, and a tripartite division of territory seemed more likely to insure the continued independence of all. This principle appears to have operated, at least temporarily, even in the case of a united campaign against Vijayanagar in 1565, for the destruction of Vijayanagar's power was followed almost immediately by the formation of a series of alliances between Vijayanagar and one or another of the victorious Muslim kings.

The defeat of Vijayanagar opened the way for the southward expansion of Bijāpur and Golconda, and the terms of a 1571 treaty between Bijāpur and Ahmadnagar suggest that it was considered proper to restore a potentially unbalanced political situation by recognizing the right of Ahmadnagar to annex Berār and Bīdar in return for recognition of Bijāpur's right to annex an equivalent amount of territory from Vijayanagar. Ahmadnagar did not annex Bīdar, but it did acquire Berār in 1574. Bijāpur was unable to take full advantage of the opportunities for expansion to the south during the 1570s, partly because of factional disputes among the nobles. Thus, Ahmadnagar managed to retain a slightly superior position. The tide began to turn in the 1580s, however, with the establishment of a stable regency at Bijāpur, fortified by a series of marriage alliances with other royal lines in the Deccan and by the political deterioration of Ahmadnagar under the rule of the slightly mad Murtaḍā Niẓām Shāh. Murtaḍā's murder in 1588, by a son who was more insane than he, set off a chain of events that resulted in simultaneous invasions by Bijāpur from the south and by Murtaḍā's brother Burhān, who had the support of the Mughal emperor Akbar, from the north. Burhān defeated the army of Ahmadnagar, recalled the foreign nobles (as the newcomers of Bahmanī times were now designated) who had been expelled from the kingdom, and assumed the throne in 1591. Campaigns against Bijāpur and against the Portuguese at Chaul (just south of present-day Bombay), as well as bitter rivalry between the Deccani and foreign nobles, further weakened Ahmadnagar at a time when Akbar's growing interest indicated grave danger. The death of both Burhān and his son in 1595 was followed by increased factionalism and finally by civil war as rival claimants to the throne were put forward. When one party appealed for aid to the governor of Gujarāt, Akbar had an excuse to launch the campaign he had already been planning. The two wars that followed resulted in the Mughal acquisition of Berār, the capture of the ruler of Ahmadnagar, and the defeat and annexation of Khāndesh. A group of nobles, however, led by the Abyssinian Malik ʿAmbār, raised a member of the royal family to the throne at Daulatābād and continued to fight the Mughals.

Political deterioration of Ahmadnagar

Mughal conquests in the Deccan. Largely because of the efforts of Malik ʿAmbar, a former African slave, Ahmadnagar was able not only to halt the Mughal advance into the Deccan during the next few years but also even to recover some of the territory lost before 1600. An extremely able and ambitious man, Malik ʿAmbar instituted revenue reforms, probably modelled after the recently introduced Mughal reforms by Todar Mal under Akbar, strengthened his contingents of Marāṭhā light cavalry, and emphasized training in guerrilla warfare for them. These preparations, as well as the rivalries among the nobles stationed in the Mughal Deccan during the reign of Akbar's successor, Jahāngīr, enabled Malik ʿAmbar's forces to win several victories and to recover Ahmadnagar city; but the Mughals decisively defeated a combined army from Ahmadnagar, Bijāpur, and Golconda in 1615, and in 1621 Malik ʿAmbar was forced to cede an additional strip of territory to the Mughals. For the next five years Malik

ʿAmbar held his own, once against an alliance between Bijāpur and the Mughals, although usually in alliance with Bijāpur, whose aid was given more with an eye to gaining territory from Ahmadnagar as it disintegrated than with the idea of stopping the Mughals. Malik ʿAmbar's death in 1626, however, was followed rapidly by disunity among his supporters and by the defeat and capture of the last Niẓām Shāh at Daulatābād in 1633.

Final annexation of Ahmadnagar

The final annexation of Ahmadnagar led to a direct confrontation between the Mughals and Bijāpur, Golconda, and the Marāṭhā leader, Shāhjī Bhonsle, who had built up a force while in the service of Ahmadnagar and was now operating independently. Bijāpur, weakened by internal dissensions after the death of Sultan Ibrāhīm II in 1626, was unable to resist the Mughal invasion of 1636 and agreed to a peace treaty by which Bijāpur acknowledged Mughal supremacy, promised to pay an annual tribute and to allow Mughal arbitration of disputes with Golconda, and recognized the old boundary of Ahmadnagar as the limit of Mughal territory. The peace thus established lasted until the death of Sultan Muḥammad in 1656. Next, the Mughals established peace with Golconda on even harsher terms. Thus, although Bijāpur and Golconda continued as independent states for another 50 years and, in fact, continued to make some progress in the south, where Vijayanagar's authority had disintegrated, they were already subject to some terms of Mughal authority by 1636 and were marked for later conquest. Indeed, it was probably only the rivalry among the Mughal emperor Shāh Jahān's sons, his illness in 1656, and the subsequent war of succession that saved the Deccan kingdoms from extinction in the 1650s.

The only force in the south that did not appear ready to fall before the Mughals was the Marāṭhās. Marāṭhā soldiers had been serving under the Deccan sultans for some time, but it was the organization of Marāṭhā forces by Ahmadnagar's Malik ʿAmbar, together with the unsettled conditions caused by the Mughal wars, that appears to have given Marāṭhā commanders the opportunity to enhance their positions. Foremost among them was Shāhjī Bhonsle, who continued to control, sporadically, a number of forts in their home territory after the fall of Ahmadnagar. For a time Shāhjī was in the service of Bijāpur, but primarily he served himself, organizing his forces and preparing the way for an explosion of Marāṭhā power under the direction of his more famous son Śivājī after 1650. Before his death in 1680, Śivājī had developed a political organization and a fighting force that proved able to withstand another 27 years of grim warfare against the Mughals and finally to cause a Mughal retreat from the Deccan.

THE VIJAYANAGAR EMPIRE, 1336–1646

The brief period of Mughal rule in South India following the conquest of Bijāpur and Golconda climaxed a movement toward the expansion of Muslim rule that had begun more than three centuries earlier; but during most of that period the far south was controlled by the last great medieval Hindu kingdom, Vijayanagar.

Many details of the founding of Vijayanagar are not known. It is clear, however, that the kingdom arose out of the political confusion resulting from the Muslim incursions into South India during the first part of the 14th century. At that time many existing dynasties were overturned, and shifts in the balance of power occurred within the region. The powerful local landholders (many of whom were called Nāyaks), probably spurred both by unhappiness at the higher revenue demands imposed by the Delhi sultans and by the concurrent revival of Śaivism, were able to organize significant military actions soon after Muḥammad ibn Tughluq's departure from the Deccan in 1329.

Founding of the kingdom of Vijayanagar

Development of the state. The kingdom of Vijayanagar was founded by Harihara and Bukka, two of five brothers who had served in the administrations of both Kākatīya and Kampili before those kingdoms were conquered by the armies of the Delhi sultanate in the 1320s. When Kampili fell in 1327, the two brothers were captured and taken to Delhi, where they converted to Islām. They were returned to the Deccan as governors of Kampili for the sultanate with the hope that they would be able to deal with the many local revolts and invasions by neighbouring Hindu kings. They followed a conciliatory policy toward the landholders of the area, many of whom had not accepted Muslim rule, and began a process of consolidation and expansion. Their first campaign was against the neighbouring Hoysaḷa king, Ballāla III of Dōrasamudra, but it stagnated; after the brothers reconverted to Hinduism and proclaimed their independence from the Delhi sultanate, however, they were able to defeat Ballāla and thereby secure their home base. Harihara I (reigned 1336–56) then established his new capital, Vijayanagar, in an easily defensible position south of the Tungabhadra River. Successful revolts against Delhi by several Hindu rulers on the eastern coast of the Deccan and by the Muslim governor of Maʿbar in the extreme south opened opportunities for expansion, and the subsequent struggles among the Hindu and Muslim powers of the south, based more upon political and military than religious lines, resulted in the emergence of both Vijayanagar and its natural rival, the Muslim Bahmanī sultanate.

Conquests. In 1336 Harihara held uneasy suzerainty over lands extending from Nellore, on the southeast coast, to Bādāmi, south of Bijāpur on the western side of the Deccan. All around him new Hindu kingdoms were rising, the most important of which were the Hoysaḷa kingdom of Ballāla and the Andhra confederacy, led by Kāpaya Nāyaka. But Ballāla's kingdom was disadvantageously situated between the Maʿbar sultanate and Vijayanagar, and two years after Ballāla was killed by the sultan in 1343, his kingdom had been conquered by Bukka, Harihara's brother, and annexed to Vijayanagar. This was the most important victory of Harihara's reign; the new state now could claim sovereignty from sea to sea, and in 1346 a great celebration was held, attended by Harihara and his four brothers. In the same year Bukka was established as joint ruler, with his headquarters at Gatti.

Harihara's brothers made other, less significant conquests of small Hindu kingdoms during the next decade. But the foundation of the Bahmanī sultanate in 1347 created a new and greater danger, and Harihara was forced to lessen his own expansionist activities to meet the threat posed by this powerful and aggressive new state on his northern borders.

Administrative foundation of the state

During Harihara's reign the administrative foundation of the Vijayanagar state was laid. Borrowing from the Kākatīya kings he had served, he created administrative units called *stholas*, *nāḍus*, and *simas* and created officials to collect revenue and to carry on local administration, preferring Brāhmanas to men of other castes. The income of the state apparently was increased by the reorganization, although centralization probably did not proceed to the stage where salaried officials collected directly for the government in most areas. Rather, most land remained under the direct control of subordinate chiefs or of a hierarchy of local landholders, who paid some revenue and provided some troops for the king. Harihara also encouraged increased cultivation in some areas by allowing lower revenue payments for lands recently reclaimed from the forests.

Consolidation. Harihara was succeeded by his brother, who as Bukka I (reigned 1356–77) engaged in a number of costly and mainly unsuccessful wars against the Bahmanī sultans during the first decade of his reign. His efforts resulted in the establishment of the Krishna River as the boundary between the two kingdoms. The major accomplishments of Bukka's reign were the conquest of the short-lived sultanate of Maʿbar (Madura), 1370, and the maintenance of his kingdom against the threat of decentralization. During Harihara's reign the government of the outlying provinces of the growing state had been entrusted to his brothers—usually to the brother who had conquered that particular territory. By 1357 some of Bukka's nephews had succeeded their fathers as governors of these provinces, and there was a possibility that the state would become less and less centralized as the various branches of the family became more firmly ensconced in their particular domains. Bukka, therefore, removed his

nephews and replaced them with his sons and favourite generals so that centralized authority (and his own line of succession) could be maintained. But the succession of Bukka's son Harihara II (reigned 1377–1404) precipitated a repetition of the same action. A rebellion in the Tamil country at the beginning of his reign probably was aided by the disaffected sons and officers of Bukka's deceased eldest son, Kumāra Kampana, who were not ready to acknowledge Harihara's authority. Harihara was able to put down the rebellion and subsequently to replace his cousins with his own sons as governors of the provinces. Thus, the circle of power was narrowed once again. The question of succession to the throne had not been settled, however. On many occasions, the conflict resumed between the King and his lineal descendant, who tried to centralize the state, and the collateral relatives (cousins and brothers), who tried to establish ruling rights over some portion of the kingdom.

The temporary confusion that followed the assassination of the Bahmanī sultan 'Alā'-ud-Dīn Mujāhid in 1378 gave Harihara the opportunity to recapture Goa and some other ports on the western coast. During the next decade pressure increased for expansion against the Reddi kingdom of Kondavīdu in the northeast. The slight gains made in 1390–91 against an alliance of the Velama chieftain of Rājakonda and the Bahmanīs were more than offset when the Bahmanī sultan besieged Vijayanagar in 1398–99, slaughtered a large number of people, and exacted a promise to pay tribute. The tribute was withheld two years later, however, when Vijayanagar made alliances with the sultans of Mālwa and Gujarāt. Nevertheless, Harihara's reign was relatively successful because he expanded the state, maintained internal order, and managed to fend off the Bahmanī sultans. The control of ports on both coasts provided opportunities for the acquisition of increased wealth through trade.

Wars and rivalries. Harihara II's death in 1404 was followed by a violent succession dispute among his three surviving sons. Only after two of them had been crowned and dethroned was the third, Devarāya I (reigned 1406–22), able to emerge victorious. During his reign the involvement of Vijayanagar and the Bahmanī sultanate as backers of different claimants to the throne of the Reddi kingdom of Kondavīdu led to another direct confrontation between the two powers, another siege of Vijayanagar city, and to the eventual partition of Kondavīdu (1420) between Vijayanagar and the Velamas of Rājakonda, who had changed sides during the protracted struggle. This extensive involvement in Telingāna brought Vijayanagar into conflict for the first time with the kingdom of Orissa to the north. Although a war was temporarily averted, there began a rivalry that was to last for 125 years.

Rivalry with Orissa

Perhaps Devarāya's most significant achievement was his reorganization of the army. Realizing the value of cavalry and well-trained archers, he imported many horses from Persia and Arabia and hired Turkish bowmen, as well as troopers who were skilled in mounted warfare. Thus, although it appears that he was seldom able to best the Bahmanīs in the field, he had begun to narrow the strategic and technological gap between north and south and to build an army that would be better suited to warfare on open plains.

The short reigns of Devarāya's two sons, Rāmcandra and Vijaya, were disastrous. In a war against the Bahmanīs, many temples were destroyed, and Vijaya was forced to pay a huge indemnity. A combined invasion by the King of Orissa and the Velamas of Andhra resulted in the loss of the territories newly gained in the partition of the Reddi kingdom of Kondavīdu. Vijaya's son and successor, Devarāya II (died 1446), reconquered the lost Reddi territories and incorporated them into his kingdom, restored the power of his subsidiary Reddi allies at Rājahmundry (*c.* 1428), and, in 1443, defended them again against the incursions of the Velamas and the King of Orissa. Wars with the Bahmanīs in 1435–36 and 1443–44 over control of Raichūr and Mudgal forts in the Krishna-Tungabhadra doab ended inconclusively. Those campaigns, however, led to further improvements in Vijayanagar's military forces when Devarāya proclaimed that Muslims would be welcome in his service and hired a number of Muslim archers to instruct his Hindu troops. Devarāya also levied tribute from Ceylon and campaigned successfully in the Kerala country of the far south, where his victories over local chieftains suggest a process of consolidation. His reign saw both the greatest territorial extension and the greatest centralization of the first period of the history of Vijayanagar.

Decentralization and loss of territory. During the 40 years after Devarāya's death in 1446, the centralized power of the state declined, and a considerable amount of territory along both coasts was lost to the Bahmanī sultans and to the suddenly powerful Gajapati ruler of Orissa. During the 1450s and '60s Kapilendra, the great king of Orissa, together with his son Hamvira, conquered the Reddi kingdom of Rājahmundry and the Vijayanagar province of Kondavīdu, captured Warangal and Bīdar from the Bahmanīs, and sent a victorious army down the east coast as far south as the Kāverī (modern Cauvery) River.

Although the army of Orissa soon withdrew from the southernmost districts, the raid had a considerable effect upon Vijayanagar. It not only weakened the empire in the east but also indicated that provincial governors might have to fend for themselves if they expected to retain their territories. The fact that Devarāya's son Mallikārjuna (reigned 1446–65) was succeeded by a cousin rather than by his own son was another indication of lessened central control and of the failure of the King and his immediate family to secure their own future, as had been done by many of his ancestors when they removed their cousins from positions of power. The new ruler, Virūpāksha (reigned 1465–85), had been a provincial governor. His usurpation was not accepted by many of the provincial governors on the east and west coasts or by the direct descendants of Mallikārjuna, who retired to the banks of the Kāverī and ruled much of the southern part of the kingdom in a semi-independent fashion.

Beginning in 1470, the Bahmanīs, under the chief minister Maḥmūd Gāwān, began a campaign that succeeded in taking much of the west coast and the northern Carnatic (Karnatic) from Vijayanagar. The loss of Goa and other ports was especially disconcerting because it cut off not only an important source of trade and state income but the principal source of supply of Middle Eastern horses for the military as well. The death in 1470 of Kapilendra of Orissa temporarily relieved military pressure in the east; but it was a provincial governor, Narasiṃha Sāluva of Chandragiri, rather than Virūpāksha, who took advantage of the resultant civil war in Orissa to regain lost territory. Narasiṃha also aided the displaced king of Orissa to regain his throne from his brother and consequently became involved in a war with the Bahmanīs (1478–81), who supported the other claimant.

Reconquest of lost territory

Later dynasties. Beginning as a small chieftain about 1456, Narasiṃha had put together a large dominion by 1485 as a result of conquests in the south, as well as campaigns against Orissa; and, although nominally subordinate to Virūpāksha, he was performing more extensive military and administrative functions than was his superior. It is not surprising, when Virūpāksha was murdered by one of his sons, who was in turn murdered by his brother, that Narasiṃha Sāluva (reigned 1485–90) stepped in to remove the new ruler and to begin his own dynasty. Usurpation was easier than consolidation, however, and Narasiṃha spent his reign in relatively successful campaigns to reduce his vassals throughout the kingdom to submission and in unsuccessful attempts to stop the encroachment of the King of Orissa. Narasiṃha also opened new ports on the west coast so that he could revive the horse trade, and he generally revitalized the army. By 1490 the process of centralization was begun again, and both internal and external political circumstances soon would combine to create better opportunities than ever before.

Reconsolidation. Narasiṃha Sāluva left his kingdom in the hands of his chief minister, Narasa Nāyaka, by appointing him regent for his two young sons. The minister, in effect, was the ruler from 1490 to 1503. Court intrigues led to the murder of the elder prince by one of Narasa Nāyaka's rivals and to the capture and virtual impris-

onment of the younger by Narasa Nāyaka in 1492. The usurpation resulted in opposition from provincial governors and chiefs that lasted for the rest of Narasa Nāyaka's life. Early in his regency, however, he had the opportunity to take advantage of the beginning of the disintegration of the Bahmanī sultanate. He invaded the disputed Tungabhadra-Krishna doab in 1492–93 at the invitation of the Bahmanī minister, Qāsim Barīd, who was trying to subdue the newly independent Yūsuf ʿAdil Khān of Bijāpur. Narasa Nāyaka took the strategic forts of Raichūr and Mudgal; and, although they were lost again in 1502, the growing disunity of the emerging Muslim polities would provide many similar opportunities in the future.

Narasa Nāyaka also campaigned in the south to restore effective control, which had not existed in many areas since the raid from Orissa in 1463–64. He compelled most of the chiefs and provincial governors to recognize his suzerainty in both Tamil country and the Carnatic and nearly restored the old boundaries of the kingdom (the eastern districts were still held by Orissa). By 1503 Narasa had practically completed the process of reconsolidation with which Narasiṃha Sāluva had charged him. He also had made virtually certain, however, that his own line rather than that of his old master would continue to rule. It was during the reigns of his sons that Vijayanagar rose to new heights of political power and cultural eminence.

Narasa's eldest son and successor, best known as Vīra Narasiṃha (reigned 1503–09), ended the sham of regency. After ordering the by-then grown prince's murder in 1505, he ascended the throne and inaugurated the third dynasty of Vijayanagar. The usurpation again provoked opposition, which the new king spent most of his reign attempting to quell. He was successful except in subduing the rebellious chiefs of Ummattūr and Seringapatam in the south and in recovering Goa from the Portuguese, with whom, however, he was able to establish relations to obtain a supply of better horses.

Reign of Krishna Deva

Growth of power. Vīra Narasiṃha was succeeded by his brother Krishna Deva Rāya (reigned 1509–29), one of the greatest of the Vijayanagar kings. During his reign the kingdom became more powerful than ever before, and internal consolidation reached a new peak. Krishna Deva spent the first 10 years of his reign solidly establishing his authority over his subordinate chieftains and governors while fending off invasions from the northeast. After decisively defeating an invading coalition of Bahmanī forces (who by this time were virtually separated into five states) and capturing Raichūr fort, Krishna Deva took advantage of a quarrel between Bijāpur and the Bahmanī ruler to subdue both Gulbarga and Bīdar and to restore the imprisoned Bahmanī sultan to his throne in 1512. During the same period he conducted a successful campaign to subdue Ummattūr in the south, and a new province was established from it. From 1513 to 1516 Krishna Deva campaigned against the Gajapati ruler of Orissa, conquering all that king's territory up to the Godāvari and raiding as far as the Orissan capital at Cuttack. Orissa then sued for peace, and its king gave his daughter in marriage to Krishna Deva, who consequently returned to Orissa all the conquered territory north of the Krishna River.

While Krishna Deva was fighting in the east, Ismāʿīl ʿĀdil Shāh of Bijāpur had retaken Raichūr fort. In 1520 Krishna Deva decisively defeated Ismāʿīl with some aid from Portuguese gunners and recaptured Raichūr. In 1523 he carried the attack further and invaded Bijāpur, captured several forts, and tried again to create dissension among the Muslims by raising a Bahmanī claimant to the throne at Gulbarga. One result of these successful campaigns and of Krishna Deva's subsequent haughty behaviour was to point out vividly to the Muslim rulers the dangers posed by Vijayanagar, so that in years to come they thought more and more of concerted action against that kingdom. Krishna Deva's highly successful reign thus led to increased danger to his realm. During most of his reign Krishna Deva maintained a mutually advantageous relationship with the increasingly powerful Portuguese, whereby he retained access to trade goods, especially to horses from the Middle East, while the Portuguese were allowed to trade in his dominions.

Vijayanagar was visited by Magellan's cousin, Duarte Barbosa, between 1504 and 1517, and, even allowing for the exaggeration that was often present in European accounts of India, Barbosa's description suggests a prosperous and well-populated country. The capital city had many palaces, wide yet crowded streets, and a wide variety of trading enterprises. The King reportedly had an army of 100,000 men, although it is unlikely that most were present at the capital except at the beginning of a campaign, and, upon his death, said Barbosa, 400 or 500 women would burn themselves on his funeral pyre.

About 1524–25 Krishna Deva abdicated and had his young son crowned king. His son died shortly thereafter, however, reportedly poisoned by the jealous former chief minister. Krishna Deva imprisoned the minister and his family and dealt successfully with a serious rebellion three years later—when one of the minister's sons escaped—as well as with Ismāʿīl ʿĀdil Shāh's attempt to take advantage of Krishna Deva's troubles to recoup his position. Krishna Deva's death in 1529 ended the period of the kingdom's greatest military and administrative success.

Renewed decentralization. Krishna Deva had passed over his infant son and his young nephew and picked his half brother Achyuta Deva Rāya (reigned 1529–42) to succeed him. Following a brief succession dispute, Achyuta Deva Rāya was able to reach the capital from Chandragiri, where Krishna Deva had kept him and other princes confined, and to ascend the throne. Although he probably was not as dissolute a ruler as the Portuguese writer Alvar Núñez Cabeza de Vaca described him to be, the severe challenges he faced made a successful reign difficult. Krishna Deva's death had precipitated renewed attacks by Bijāpur, Golconda, and Orissa and a revolt by the King's minister, Vīranarasiṃha Sāluva, and the southern chieftains of Ummattūr and Tiruvadi. Achyuta dealt successfully with all the enemies until the late 1530s, when he was imprisoned by Rāma Rāya, the chief minister, with whom he had agreed to share power. Opposition to Achyuta's imprisonment by some of the nobles, combined with a revolt in the south, led to his release and the beginnings of civil war; but the new ruler of Bijāpur, Ibrāhīm ʿĀdil Shāh, after early attempts to create divisiveness in Vijayanagar, arbitrated a settlement between Achyuta and Rāma Rāya. Under the settlement it was agreed that Achyuta was king but that he would not interfere with Rāma Rāya's role within his own estate.

Achyuta's reign ended with about the same external boundaries of the kingdom as in 1529; but the struggle with Rāma Rāya plus the activities of other nobles and chieftains weakened the hold of the centre over some of the provinces. The process of decentralization had set in again, but now the strongman who would pull the kingdom together was already on the scene. Rāma Rāya brought himself to the undisputed pinnacle of power in 1542–43, when he defeated his rival in the succession struggle following Achyuta's death and crowned his own candidate, Achyuta's nephew Sadāśiva (reigned 1542–76). After seven or eight years, Rāma Rāya also assumed royal titles, but from the first Sadāśiva was kept under guard, and Rāma Rāya, together with his brothers Tirumala and Venkatādri, ruled the kingdom.

Rāma Rāya was able to control, although not to subdue entirely, rebellious nobles in the east and the extreme south. He also concluded a treaty with the Portuguese (1547), whose settlements had been expanding and who had caused no small amount of damage to indigenous settlements over the past few years. The treaty was broken in 1558, however, and Rāma Rāya then exacted tribute in compensation for damage to temples caused by the Portuguese.

Relations with the Muslim states. Most crucial during the period of Rāma Rāya's rule, however, were Vijayanagar's relations with the Muslim successor states to the Bahmanī sultanate. At least since Krishna Deva Rāya's time, Vijayanagar had usually competed on a more than equal basis and in the same system of state rivalries with the five Muslim states. Thus, an invasion from Bijāpur was repulsed in 1543; in 1548 Rāma Rāya aided Burhān Niẓām Shāh of Ahmadnagar in taking a fort from

The four-party treaty of the mid-16th century

Bīdar, but in 1557 Rāma Rāya allied himself with Bijāpur against the Niẓām Shāh and Golconda. The result of the last war was a collective treaty, by which any of the four parties, attacked unjustly by another, could call upon the other allies to stop the aggressor. When Ḥusayn Niẓām Shāh broke the treaty by invading Bijāpur in 1560, Vijayanagar and Golconda responded with an attack that not only resulted in Ahmadnagar's loss of the fort of Kalyāṇi to Bijāpur but also in an invasion of Bīdar and the defeat of its ruler by Rāma Rāya. Soon, however, the ruler of Golconda, Ibrāhīm Quṭb Shāh, allied himself with Ahmadnagar against Bijāpur, and Rāma Rāya allied Vijayanagar with Bijāpur to severely defeat the aggressors.

Decline of Vijayanagar. It is likely that the sultans of Golconda and Ahmadnagar, who had lost much at the hands of Rāma Rāya, were primarily responsible for the formation of an alliance that destroyed Vijayanagar's power forever. By 1564 at least four of the five sultans (Berār is questionable) had declared a holy war and had begun their march on Vijayanagar, which resulted early in 1565 in the disastrous defeat of the Vijayanagar forces in the Battle of Talikota and in the subsequent sack and destruction of much of the city of Vijayanagar. Rāma Rāya was captured and killed, but his brother Tirumala escaped to the south with the King and much of the royal treasure.

Military policies. Although Rāma Rāya's efforts toward centralization were not entirely successful, it was his military policies, probably through no fault of his own, that ultimately led to disaster. There were rebellions when he replaced many members of the old nobility with relatives and close associates, but they appear to have been no more serious than many another rebellion of previous periods under similar circumstances. Indeed, judging on the basis of the number and size of the military campaigns that Rāma Rāya was able to launch outside Vijayanagar in later years, it would seem that his internal control was relatively secure. Rāma Rāya has been criticized for allowing Muslims to hold important positions within his administration, and, although his final defeat at Talikota was at least partly attributable to the defection of two of his Muslim generals, the policy appears to have worked well up to that time. Rāma Rāya's early experiences as an official at the court of Golconda appear to have given him ideas for improving the Vijayanagar administration and army. As early as 1535 he had hired 3,000 Muslim soldiers from Bijāpur, and he later tried to make the Vijayanagar state apparatus more like that of the neighbouring Muslim states. In short, he was building a state that would be as competitive as possible in that time and place. It is likely that at first Vijayanagar's Muslim neighbours took a similar view of state relations and that Vijayanagar was seen as just another competing state. Rāma Rāya's military successes plus the knowledge that Vijayanagar was stronger than any one of the sultans led to the Muslim alliance against it. Despite a Muslim historian's claim that the alliance was formed because of Rāma Rāya's bad treatment of Muslims, there is little evidence to indicate that the principal motives were other than political. Furthermore, the subsequent behaviour of the sultans suggests that once Vijayanagar had been humbled they were willing to return to a system of shifting alliances among all the Deccan powers.

Loss of central control. The Battle of Talikota did not result in the destruction of the kingdom of Vijayanagar, although the capital city never fully recovered from the ravages it suffered. Rāma Rāya's brother Tirumala established a new headquarters at Penugonda and began to rebuild the army, but the defeat at Talikota had been so severe that he could no longer maintain control over many of the provinces. Much of the south and southeast was lost as the Nāyaks of Madura, Tanjore (Thanjāvūr), and Jinjī effectively asserted their independence. Rebellions and banditry arose in many areas. Tirumala successfully defended himself against a challenge by his nephew, who was aided by Bijāpur, and then joined with Ahmadnagar and Golconda in a campaign against Bijāpur. Tirumala accepted the new states of Nāyaks of the south, retained the allegiance of Mysore, Vellore, and Keladi, and set his three sons as governors of the three linguistic regions of his kingdom. In 1570 he had himself crowned and thus officially inaugurated the fourth and last Vijayanagar dynasty (although the puppet king Sadāśiva appears to have survived until 1576).

Rebellions and banditry

When Tirumala retired, his son Śrīranga I (reigned 1572–85) tried to continue the process of rebuilding while struggling to maintain his place among the Muslim sultanates. An invasion by Bijāpur was repulsed with the aid of Golconda, but invasions by Golconda resulted in the loss of a substantial amount of territory in the east and the loss of the wealth of the temple of Narasiṃha at Ahobalam, which was plundered by a Marāṭhā Brāhmaṇa officer in the service of Golconda. Śrīranga's difficulties partly stemmed from the lack of aid from his brothers, who ruled their separate regions, and partly from the dissensions of his nobles and the semi-independent status of some of them. Many nobles had apparently decided that it was no longer in their best interests to give full support to the larger state and that, in the absence of overwhelming power, the development of smaller, subregional states was both possible and potentially more profitable.

Śrīranga died without issue and was succeeded by his younger brother Venkata I (reigned 1585–1614), whose ability and constant activity, combined with a relative dearth of interference by the Muslim sultanates, prevented the further disintegration of centralized authority over the next 28 years. A series of wars between 1580 and 1589 resulted in the reacquisition of some of the territory that had been lost to Golconda in the east, but Venkata spent most of his time attempting to retain his hold over his rebellious chieftains and nobles. Most of the east and the Tamil south was in rebellion at one time or another, the most serious threat occurring in 1601, when the Nāyaks of Madura, Tanjore, and Jinjī came to the aid of the rebellious Lingama Nāyaka of Vellore. The Nāyaks were defeated, and Vellore later was captured, but Venkata's authority was not restored in the far south. The process of decentralization, although halted for a time, could not be reversed. In the northern areas that had been laid waste by invading armies, Venkata undertook a program of restoration by offering lower revenue payments. It would appear that by the time of his death in 1614 he had accomplished enough so that a revival of imperial power and prosperity was possible, but instead rivalries among the nobility rapidly led to further decentralization and to the diminution of the state.

Breakup of the empire. Venkata's nephew and successor, Śrīranga II, ruled only four months. He was murdered, along with all but one of the members of his family, by one of the two contending parties of nobles. A long civil war resulted and finally degenerated into a series of smaller wars among a number of contending parties. By about 1629 the authority of the surviving member of the dynasty, Rama Deva Rāya (reigned 1618–30), was recognized in a much-truncated kingdom along the eastern coast. Although some chieftains continued to recognize his nominal suzerainty and that of his successor for the next few years, real political power resided at the level of chieftains or provincial governors, who were carving out their own principalities. The fourth Vijayanagar dynasty had become little more than another competing provincial power.

Bijāpur and Golconda took advantage of the decline in Vijayanagar's strength to make further inroads into the south, while the King's own nephew Śrīranga allied himself with Bijāpur. In 1642 an expedition from Golconda drove the King from his capital at Vellore. Hearing that his uncle was dying, Śrīranga deserted Bijāpur and had himself crowned. Although he was able to play off Bijāpur and Golconda against each other for a time, he could not gain control over the provincial Nāyaks, who were by then virtually independent; and, when Bijāpur and Golconda finally struck at the same time, Śrīranga was helpless. A last appeal to his Nāyaks to come to the defense of Hinduism resulted instead in his defeat by their combined forces in 1645. Meanwhile, Bijāpur and Golconda advanced, with the blessings of the Mughal emperor at Delhi, who had suggested that they should partition the Carnatic between themselves. The Nāyaks realized the danger too late, and

Muslim conquest of the Carnatic

by 1652 Muslim sultans had completed their conquest of the Carnatic. Śrīranga retired to Mysore, where he kept an exile court until his death in 1672.

Administration of the empire. The Vijayanagar empire achieved a degree of political unity over a number of linguistic regions of South India for roughly two centuries and was, to some extent, consciously represented by its sovereigns as the last bastion of Hinduism against the forces of Islām. As with similar Muslim religio-political claims, however, this one often appeared to be more rhetorical than real. The shifting patterns of alliances among Vijayanagar and the sultanates, the occasions on which a rival party of nobles or a claimant to the throne of Vijayanagar would enlist the aid of a Muslim sultan, and the employment of both Hindus and Muslims in the sultanates and Vijayanagar suggests that rivalries were more political than religious. The various progressive reforms of the Vijayanagar army suggest also that efforts were made to transform at least one aspect of the state in order to make it more competitive with its Muslim rivals.

The administration of the kingdom sporadically achieved a relatively high degree of centralization, although centrifugal tendencies regularly appeared. The central administration had both a revenue and military side, but the actual business of raising taxes and troops was mostly the responsibility of the provincial governors and their subordinates. The central government maintained a relatively small body of troops, but it assigned a value to the lands held by the provincial governors and determined the number of troops that were to be supplied from the revenues of each province. This administrative plan led to the development of the Nāyak system, which worked well enough when the central authority was strong but provided territorial bases for the Nāyaks to build semi-independent holdings in times of imperial weakness. The imperial rulers were aware of the power of the provinces and tried to counter it by appointing members of the royal family as governors of the militarily more important (but not necessarily more lucrative) provinces. On the whole, however, the device was not successful because succession rivalries, as in the Muslim kingdoms to the north, tended to produce filial disloyalty to the throne and even rebellion.

Although exact figures are unavailable, the evidence suggests that the level of taxation was close to half of the produce in many areas. Much of the revenue collected did not go to the state, however, because various layers of local landholders took their share first. Although most revenue came from agrarian taxes, commercial and artisan taxes were levied as well. (P.B.Ca.)

The Mughal Empire, 1526–1761

THE ESTABLISHMENT OF THE MUGHAL EMPIRE

The significance of Mughal rule. The establishment of the Mughal Empire (1526) was an event of great national and international importance. The Mughals were not rigid Muslims, and although the first two rulers of the line—Bābur and Humāyūn (Homāyūn)—generally followed policies similar to those of the Delhi sultans, they did not strictly enforce Muslim law. Akbar, the third ruler of the dynasty, almost completely discarded, both in theory and practice, the principles of the Islāmic state; he declared himself to be the impartial ruler of both Hindus and Muslims and accorded them equal rights and privileges. He disestablished Islām as religion of the state and, contrary to the practice of the sultans, placed all religions in the land at par. This policy amounted to a political and social revolution. It generated forces that put the diverse sections and communities of India on the road toward becoming one people, enabling the country to achieve unprecedented unity and political, social, economic, and cultural progress.

Internationally, too, India not only began to regain its position among the nations of Asia and Europe but it actually became one of the richest and most powerful countries in the world. Persia, Transoxania, and Turkey recognized India's position as a great power, perhaps greater than each one of them individually. European nations, such as Portugal, England, France, and Holland, displayed eagerness to have friendly relations with India and to share its wealth through commerce.

Mughal origins. The foundation of the Mughal Empire was not a single or isolated event. The idea was inspired by Timur, who invaded India in 1398 but had no definite ambition to conquer and rule over it. Nevertheless, Timur nominated one Khiḍr Khān Sayyid as his governor of Multān and probably also as his representative in Hindustān. Khiḍr ruled in Timur's name and, after Timur's death in 1405, in that of his son and successor Shāh Rokh Mīrzā (Shāhrukh). Khiḍr Khān's son, however, severed connections with the Timurid dynasty and set himself up as an independent ruler in 1421. In 1503 a direct descendant of Timur, Ẓahīr-ud-Dīn Muḥammad Bābur, inspired by the story of Timur's invasion of India, formed the resolution of one day repeating the exploit.

Bābur. Bābur was the fifth descendant of Timur on the side of his father and the 14th descendant of Genghis Khan through his mother. He was a Chagatai Turk, so-called because his ancestral homeland, the country north of the Oxus, was the heritage of Chagatai Khan, the son of Genghis Khan. But he was commonly known as the "Mughal." On the death of his father in 1494, Bābur, at age 11, became king of Farghana (Fergana), which formed the northern part of Transoxania (now in the Uzbek S.S.R.). As a boy king he lost Farghana and passed through a period of adversity, having to fight his own paternal and maternal uncles. During 20 years of adversity and warfare Bābur trained himself by appropriating military tactics and modes of warfare of his adversaries—Turks, Mongols, Persians, and Afghans, and he made a scientific synthesis of these systems. He became a good general through a career of many defeats and through years of homeless wanderings, privations, and hardships.

Foiled in his design of recovering his ancestral dominion in Central Asia, Bābur turned his attention to India. His conquest of the Punjab and relations with the Afghan rebels Dawlat Khān and ʿĀlam Khān have been described above.

Having secured the Punjab, Bābur advanced toward Delhi, receiving encouraging offers of support from many Delhi nobles. He routed two advanced parties of Ibrāhīm Lodī's troops and met the sultan's main army, 40,000 (according to Bābur 100,000) strong, east of the town of Pānīpat in April 1526. Bābur claimed to have 12,000 men all told, but historians have placed the Mughal force at 25,000 or more, because a large number of Indian mercenaries had swelled it after Bābur's occupation of the Punjab. The first Battle of Pānīpat

After eight days of inaction, during which the opposing forces faced each other at a few miles' distance, Ibrāhīm, provoked by an unsuccessful night attack by Bābur's troops, moved forward on April 21. Seeing that the Mughal front line was defended by a formidable row of wagons and breastworks, Ibrāhīm's troops hesitated and lost the advantage of a shock charge. Bābur, taking advantage of this, ordered his men to attack. Ibrāhīm fought bravely, but his swords and arrows could not answer Bābur's guns, and he died with 15,000 of his men. The Delhi army then broke and fled, and Bābur occupied Delhi and Āgra within a few days. His success was the result of a scientific combination of cavalry and artillery, the rapidity of his movement, and his superior generalship.

Bābur, however, had yet to encounter his most powerful Indian adversary, Rānā Saṇga of Mewār, the head of the Rājput chivalry in the country, before he could feel himself secure. The accounts of Rājput bravery made the Mughals nervous; but Bābur raised their morale by a stirring speech and roused their religious feelings by taking a vow in the presence of his troops to behave like a true Muslim, by renouncing wine and by breaking his wine vessels in a dramatic fashion. A very severe contest took place at Khānua, near Fatehpur Sīkri, on March 16, 1527. Bābur made use of the same tactics as at Pānīpat two years before, and won a similar victory.

After defeating the Afghāns of eastern Uttar Pradesh and Bihār on the confluence of the Ganges and the Ghāgra on May 6, 1529, Bābur was in possession of the country from the Indus to Bihār and from the Himalayas to Gwalior

and Chanderi; he had already obtained possession of Multān, and therefore in the northwest corner of India, only Sindh (Thatta) remained beyond his jurisdiction.

Though a very intelligent, shrewd, and highly cultured prince, who left an imperishable record of his adventurous life and a lifelike description of India in his memoirs, the *Tūzuk-e Bāburī*, Bābur was not a constructive genius. He did not create a new system of administration or improve the existing institutions. He squandered the huge wealth piled up in the treasuries of Delhi, Āgra, and other places and died in 1530, before he could consolidate his newly won territories in India.

Humāyūn. Bābur bequeathed to his son Humāyūn a vast though unconsolidated empire and an empty treasury. He gave him the deathbed advice to behave kindly toward his brothers. A dutiful son, the new ruler assigned Kābul, Ghazna (Ghaznī), and Punjab to his next brother, Kāmrān, and made similar assignments for his other two brothers, ʿAskarī and Hindāl. These ambitious men wanted to rule over Hindustān. His cousins, known as the Mīrzās, were equally ambitious and disloyal. Humāyūn's greatest difficulty, however, came from two independent chiefs—Bahādūr Shāh of Gujarāt and Shēr Khān of Bihār—who coveted the throne of Delhi.

Humāyūn's expulsion from India

Humāyūn began badly by invading the Hindu principality of Kālinjar in Bundelkhand, which he failed to subdue. Next he picked a quarrel with Shēr Khān by unsuccessfully besieging the fortress of Chunār (1532). Thereafter he conquered Mālwa and Gujarāt, but he could not hold them. Leaving the fortress of Chunār unconquered on the way, Humāyūn proceeded to Bengal to assist Sultan Maḥmūd of that province against Shēr Khān. He lost touch with Delhi and Āgra, and because his brother Hindāl began openly to behave like an independent ruler at Āgra he was obliged to leave Gaur, the capital of Bengal. Negotiations with Shēr Khān fell through, and the latter forced Humāyūn to fight a battle at Chausa, 10 miles southwest of Baksar and a short distance from the Karamnāsa River (June 26, 1539), in which Humāyūn was defeated. He was compelled to cross the flooded Ganges with the help of a water carrier (Niẓām). He did not feel strong enough to defend Āgra, and he retreated to Bilgrām near Kannauj, where he fought his last battle with Shēr Khān, who had now assumed the title of Shēr Shāh. Humāyūn was again defeated and compelled to retreat to Lahore, fleeing from Lahore to Sind, from Sind to Rājputāna and from Rājputāna back to Sind. Not feeling secure even in Sind, he fled (July 1543) to Iran to seek military assistance from its ruler Shāh Ṭahmāsp. The Shāh agreed to assist him with an army on the condition that Humāyūn become a Shīʿah Muslim and return Qandahār to Iran in the event of his successful acquisition of that fortress. In 1544 Humāyūn captured Qandahār from ʿAskarī and then occupied Kābul. Kāmrān fled to Ghazna, and from there to Sind. Humāyūn again lost Kābul to Kāmrān in 1546, but he was able to recover it the next year. Kāmrān continued to give him trouble, but he was ultimately defeated and blinded by Humāyūn's orders.

Shēr Shāh and his successors. During Humāyūn's exile Shēr Shāh established a vast and powerful empire and strengthened it with a wise system of administration. He carried out a new and equitable revenue settlement, greatly improved the administration of the districts and the *parganas* (village units), reformed the currency, encouraged trade and commerce, improved communication, and administered impartial justice.

Sher Shāh's administration

Shēr Shāh died on May 22, 1545, and was succeeded by his son Islām Shāh (ruled 1545–53). Islām Shāh, preeminently a soldier, was less successful as a ruler than his father. On Islām's death his young son Fīrūz, a boy of 12, came to the throne; but he was murdered by his own maternal uncle, Mubāriz Khān, who captured the throne and assumed the title of Muḥammad ʿĀdil Shāh Sūr. His right to rule was disputed by two other descendants of Shēr Shāh, and a civil war resulted.

Restoration of Humāyūn. From Kābul Humāyūn watched the situation in India. He had been preparing since the death of Islām Shāh to recover his throne. In December 1554 he crossed the Indus and marched to Lahore, which he captured without opposition (February 1555). Meanwhile, Sūr's rivals, Sikandar and Ibrāhīm, were fighting each other near Āgra and failed to put up an effective resistance to the invader. The advance guard of Humāyūn's army occupied Jullundur, Sirhind, and Hissār without striking a blow. When the Mughals pushed toward Delhi after occupying the Punjab, Sikandar sent a large army to intercept them. A battle at Machhīwara on the Sutlej (May 15) ended in Humāyūn's victory. Sikandar himself was defeated at Sirhind on June 22 and fled to the hills of the Punjab. Humāyūn occupied Sirhind and captured Delhi and Āgra in July 1555. He thus regained the throne of Delhi after an interval of 12 years, but he did not live long after his restoration; he died as the result of an accident in Shermandal in Delhi (January 1556). His death was concealed for about a fortnight to enable the peaceful accession of his son Akbar, who was at the time away in the Punjab.

Humāyūn was not a great emperor, and was unlucky despite his name (which means "fortunate"), but his memory was perpetuated by his son, who proved to be one of the greatest rulers of the Mughal empire.

AKBAR THE GREAT

Extension and consolidation of the empire. Akbar (ruled 1556–1605) was proclaimed emperor at Kalanaur in Gurdāspur amid gloomy circumstances. Delhi and Āgra were threatened by ʿĀdil Shāh's minister Hemū, and Mughal governors were being driven from all parts of northern India. Akbar's hold over a fraction of the Punjab—the only territory in his possession—was disputed by Sikandar Sūr and was precarious. To make matters worse, Delhi and the surrounding areas had been in the throes of a devastating famine for about two years, and there was disloyalty among Akbar's own followers. But guided by his guardian, Bayram Khān, the new emperor decided to fight Hemū, who had defeated the Mughal governors of Āgra and Delhi (October 7, 1556), occupied those cities, and was advancing toward the Punjab to drive Akbar out of India.

Threats to the empire

They met in battle on November 5 on the historic field of Pānīpat. Hemū, at the head of 30,000 men, overthrew the Mughal right and left wings and launched an attack on the enemy centre, hurling 1,500 elephants against it. The Mughals were about to give way when Hemū was struck in the eye by an arrow and fell unconscious on his elephant. His army was seized with panic and fled in all directions. The Mughal victory was decisive. They occupied Delhi and Āgra, and Sikandar Sūr was compelled to surrender in May 1557. Muḥammad ʿĀdil Shāh was killed in 1557 at Monghyr. Ibrāhīm, the third Sūr pretender, had to seek asylum in Orissa. Thus within two years of Akbar's victory there remained no rival to contest his claim to the sovereignty of Hindustān.

Until 1560, the adminstration of Akbar's truncated empire was in the hands of Bayram Khān. Bayram's regency was momentous in the history of India. At its end the Mughal dominion embraced the whole of the Punjab and Multān; the territory of Delhi; the present Uttar Pradesh including Gwalior and as far east as and including Jaunpur; and in the present state of Rājasthān, Dholpur, Ajmer, Nāgaur, and Jaitāran.

Bayram Khān's regency

Akbar, however, soon became restless under Bayram Khān's tutelage. Influenced by his former nurse, Māham Ānaga, and his mother, Hamīda Bānū Baygam, he was persuaded to dismiss him (March 1560).

Four prime ministers of mediocre ability then followed in quick succession. Akbar was not yet keen to take the reins of government into his own hands, and it was only by the beginning of 1561 that the young ruler began giving some measure of attention to the affairs of government.

Although not yet his own master, Akbar took a few momentous steps during this period. He conquered Mālwa (1561) and marched rapidly to Sārangpur to punish Adham Khān, the captain in charge of the expedition, for improper conduct. Second, he appointed Shams ud-Dīn Muḥammad Atgah Khān as prime minister (November 1561). Third, he took possession of Chunār, which had always defied Humāyūn, at about the same time.

The most momentous events of 1562 were Akbar's marriage with a Rājput princess, daughter of Raja Bhārmal of Amber, and the conquest of Merta in Rājasthān. The marriage led to a firm alliance between the Mughals and the Rājputs. The princess became the mother of the future Jahāngīr, Akbar's heir and successor, and the alliance secured solid military Rājput support for the Mughal throne, contributing greatly to the empire's expansion and stability.

Mughal–Rājput alliance

In 1562 Akbar experienced a remarkable mental change. "On the completion of my 20th year," he said, "I experienced an internal bitterness and from the lack of spiritual provision for my last journey, my soul was seized with exceeding sorrow." A result of this change was the abolition, after April 10, 1562 (his 20th birthday), of the enslavement of prisoners of war and of the forcible conversion of them and their women and children to Islām. By the end of June 1562, Akbar freed himself completely from the influence of the harem. The harem party, headed by Māham Ānaga, her son Adham Khān, and some other ambitious courtiers, murdered the new prime minister, Atgah Khān (May 16). Adham then tried to force his entry into the emperor's room, probably to explain the cause of Atgah Khān's murder and to beg his forgiveness; but he was killed by Akbar. Akbar informed Māham Ānaga of the death of her son. She was then lying ill, and she died on June 24. Akbar forgave the conspirators and appointed a former prime minister, Munʿim Khān, to that office.

Thus from about the middle of 1562 Akbar took upon himself the great task of shaping his policies, leaving them to be implemented by his agents. He embarked on a policy of conquest, establishing control over Jodhpur, Bhātha (modern Rewa), and the Gakkhar country between the Indus and the Beās in the Punjab. Next he conquered the wild but prosperous region of Gondwana in the modern Madhya Pradesh. During this period he removed discrimination against the Hindus by abolishing pilgrimage taxes in 1563 and the hated *jizyah* (poll tax on non-Muslims) in 1564.

Rebellions against Akbar

In 1565 some of the principal Uzbek nobles combined together to raise the standard of rebellion. During this period Mīrzā Sulaymān of Badakhshan invaded Kābul more than once. Mīrzā Ḥakīm, Akbar's half brother, governor of Kābul, feeling unsafe, proceeded to the Indus to appeal to Akbar for help. There he was persuaded by some disloyal Mughals to try to seize Delhi. He crossed the Indus and proceeded to Lahore. When Akbar prepared to resist him, he fled back to Kābul. Mīrzā Sulaymān, who was still near Kābul, then patched up a peace with Mīrzā Ḥakīm and retreated to Badakhshan. Taking advantage of the situation, some of Akbar's other relatives tried to play the same game. But Akbar was not the man to tolerate insubordination, and the rebellions were crushed.

Subjugation of Rājasthān. Rājasthān occupied a prominent place in Akbar's scheme of conquest; without establishing his suzerainty over that region, he would have no title to the sovereignty of northern India. Unlike his father and grandfather, Akbar was well aware of the valour and sincerity of the Rājputs; his aim was to subordinate them and win their loyalty. In 1567 he invaded Chitor, the capital of Mewār, whose ruler had scornfully rejected the idea of recognizing Akbar as his overlord. Attacking Chitor's gigantic fort (October), which was well defended and provisioned and was considered impregnable, he had three extensive mines dug under cover of fire from his batteries. These exploded with a terrific noise, shattering the northern bastion of the fort from its very foundation and killing a large number of defenders. At the same time he continued hitting the fort with constant fire from heavy guns. The commander of the fort was killed, and the garrison temporarily lost heart. But the Rājputs donned yellow robes, determined to fight and die a glorious death; they flung open the gate and faced the Mughals sword in hand (February 24, 1568). Most of them were slain, and Akbar ordered the massacre of about 30,000 men. The fort fell into his hands. Chitor was constituted a district, and Āṣaf Khān was appointed its governor. But the western half of Mewār remained in the possession of Rānā Udai Singh.

The fall of Chitor

Soon after the fall of Chitor, Akbar besieged Ranthambhor, a formidable fortress; it surrendered in March 1569. Kālinjar surrendered without fighting in August 1569. In November 1570 Akbar held a conference of Rājput chiefs at Nāgaur, which was attended by the rulers of Amber, Jodhpur, Bīkaner, and Jaisalmer. All these entered into a friendly alliance with Akbar and recognized him as their suzerain. Almost the whole of Rājasthān thus submitted, except the Rānā of Mewār, who continued to offer resistance from his retreat in the hills. Būndi, Dūngarpur, Bānswāra, and Pratāpgarh submitted shortly thereafter.

Foundation of Fatehpur Sīkri. Akbar's two sons, Salīm and Murād, were born near the shrine of Shaykh Salīm Chishtī in Sīkri, and he formed the plan of converting that village into a city and making it his residence (November 1571). The hilly eminence, one mile long and a furlong broad, was levelled into a plain, and royal palaces were erected on it. Nobles and officers built mansions for themselves. A city wall with nine gates was erected on three sides, and the fourth side on the southwest was protected by an extensive lake many miles in area.

Conquest of Gujarāt. Akbar's next objective was the prosperous province of Gujarāt, an emporium of commerce with the countries of west Asia and Europe. He found little opposition and easily captured Ahmadābād (November 1572). From there he proceeded to Cambay and defeated Ibrāhīm Mīrzā, one of his kinsmen, at Sarnāl (December). Next he captured Surat after a brief siege in February 1573, and returned to Fatehpur Sīkri.

As soon as Akbar's back was turned, a rebellion broke out in Gujarāt. Ibrāhīm Mirzā, who had fled to Daulatābād, returned and besieged Ahmadābād. Akbar returned in person; he reached Ahmadābād on September 2, having marched 450 miles in 11 days. He crossed the Sābarmati against the advice of his cautious captains and attacked the enemy the same day. Ibrāhīm was defeated and taken prisoner. Gujarāt submitted finally, and Akbar returned to Fatehpur Sīkri on October 5. This expedition is rightly described as "the quickest campaign on record." The conquest of Gujarāt pushed Akbar's western frontier to the sea and brought him into contact with the Portuguese, who made peace with him.

Conquest of Bihār and Bengal. The conquest of Bengal, Bihār, and Orissa, ruled by Dāʾūd Khan, became essential in view of Akbar's new policy, which aimed at the political unity of India. Taking Patna from Dāʾūd, he subjugated Bihār in 1574, and then returned to Āgra. Munʿim Khān was left with instructions to conquer Bengal, which was occupied after the Battle of Tukra (March 3, 1575). Munʿim Khān made peace with Dāʾūd, who had fled to Cuttack and prepared to renew the contest. Dāʾūd was allowed to retain possession of Orissa, but he disregarded the treaty and tried to recover Bengal. Accordingly a fresh campaign was organized, and Dāʾūd was finally defeated and killed in a battle near Rājmahāl in July 1576. Bengal was then annexed, but a few of the local chiefs continued to give trouble for some years more.

Defeat of Dāʾūd Khan

Attempt to conquer Mewār. Udai Singh's son Pratāp, who became Rānā of western Mewār in 1572, continued to offer uncompromising resistance to the Mughals. Akbar was determined to win the remaining part of Mewār by peace or war. He sent four diplomatic missions to persuade Pratāp to recognize his suzerainty, and at the same time tried to isolate him. The missions failed, and in April 1576 Akbar deputed Mān Singh of Amber to capture the remnant of Mewār. A fierce battle took place on June 18, 1576, at the entrance of Haldīghat, a spur of the Arāvalli chain, near the village of Khamnaur on the southern bank of the Banās River. So desperate was the Rānā's charge that the Mughal advance guard and left wing were scattered and the right wing and centre were hard pressed. But Pratāp's army was small in number, and he had no reserve or rear guard to back up his initial success. In his attempt to break the Mughal centre and right wing, he hurled his war elephants against them; but arrows and bullets from the imperial side proved too much, and the Rānā's men lost heart and fled from the field. Mān Singh occupied Gogunda, but failed to secure possession of all Mewār. The Rānā eventually succeeded in recovering a good part of his ancestral territory before

his death in 1597. Akbar sent more than one expedition against Amar Singh, son and successor of Pratāp; but his kingdom remained unconquered during Akbar's lifetime.

Annexation of Afghanistan. In 1581, Akbar moved against his rebellious half-brother, Mohammad Ḥakīm of Kabul. Ḥakīm submitted, and Akbar appointed his half-sister, Bakht-un-Nisā' Baygam, governor of Kābul. Ḥakīm died in July 1585, and Kābul was then incorporated in the empire.

Submission of Ḥakīm

Annexation of Kashmir. The ruler of Kashmir, Yūsuf Shāh, declared his submission but did not present himself at the imperial court. Akbar, who was at Lahore in 1585 to safeguard his northwestern frontier from an expected Uzbek invasion and to crush a rebellion of the Pashtun tribes between India and Afghanistan, sent forces to reduce Kashmir to absolute submission early in 1586. Yūsuf Shāh recognized Akbar as his sovereign. But Akbar disapproved of the treaty, and Yūsuf was arrested. His son Yā'qūb escaped to Srīnagar and prepared to resist, but he was compelled to surrender. Kashmir was now annexed to the empire and became a district (*sarkār*) of the province of Kābul. Yūsuf Shāh was released after some time and placed in the Mughal service.

Conquest of Sind. Akbar had acquired the fortress of Bhakkar as early as 1574. He now coveted the southern part of Sind on the mouth of the Indus, without which his supremacy over northwestern India could not be complete. Moreover, he wanted to use Sind as the base of operations against Qandahār. In 1590 he directed 'Abd-ur-Raḥmān Khān-e Khānān, governor of Multān, to acquire the principality of Thatta (lower Sind) from its ruler, Mīrzā Jānī Beg. The latter was defeated and compelled to surrender his entire territory, including the fortresses of Thatta and Sehwān (1591); he too entered the imperial service.

Conquest of Orissa. Mān Singh, governor of Bihār, invaded Orissa in 1592. Its ruler, Nīṣār Khān, submitted after a feeble resistance and was confirmed as governor. But two years later he rebelled and seized the crown lands of Puri and Jagannātha. Mān Singh defeated and expelled him, and the province was annexed to the empire and became a part of Bengal.

Conquest of Baluchistan. Mīr Ma'ṣūm Khān was deputed in February 1595 to conquer Baluchistan, the only principality (except Assam) in northern India that had not yet acknowledged Akbar's authority. Ma'ṣūm attacked the fort of Sibi, northwest of Qandahār, and defeated the *amīr*s, who delivered the whole of Baluchistan, including Makrān, the region near the coast, into the hands of the Mughals.

Qandahār and the Deccan. In April 1595 the Persian governor of Qandahār, Moẓaffar Ḥoseyn Mīrzā, who was not on good terms with the authorities in Tehrān, peacefully delivered his province into the hands of Akbar's men. The acquisition of Qandahār completed Akbar's conquests in northern India and gave the empire a natural frontier in the northwest. The late 1590s were taken up with the Deccan campaigns described in the previous section, which resulted in the defeat of Ahmadnagar, and the annexation of Khāndesh and Berār.

Akbar's administration. Emboldened by the successful promulgation, in September 1579, of the *maḥẓar* (miscalled the infallibility decree), by which he assumed the right of interpreting Islāmic doctrine whenever there were conflicting opinions about the meaning of the Qur'ān, and sought to end the influence of the *'ulamā'* (religious divines) in state affairs, Akbar announced a new theory of kingship. He asserted that the king was God's representative on earth and the impartial ruler of all his subjects, irrespective of their religion. He must be absolutely tolerant to every creed, establish universal peace in his dominions, and work ceaselessly for the welfare of all classes of his people. Akbar thus considered himself to be above the Muslim law or the law of any other religion or community. He believed in the unity of leadership, secular and religious.

In trying to implement his theory of kingship, Akbar prescribed a daily routine for himself. He started work at sunrise, when he appeared before his people at the balcony of salute (*jharokha-e darshan*), and he continued working until midnight, with a few hours' gap for meals, rest, recreation, and prayers. He is said not to have slept for more than three hours a night, feeling that hard work was essential for the welfare of the people.

Akbar was wise enough to delegate much of the work to his ministers and officers, keeping in his hands the initiation of policy and the issuing of instructions and seeing that these were properly followed. He successfully exercised the functions of supervision and control over every department of administration.

Central, provincial, and local government. Akbar's central government consisted of four departments, each presided over by a minister: the prime minister (*Wakīl*); finance minister (dewan, or *wazīr*); paymaster general (*mīr bakhshī*); and chief sadr (*ṣadr uṣ-ṣudūr*), who was the chief justice and religious official combined. They were appointed, promoted, or dismissed by the emperor, and their duties were well defined.

The empire was divided into 15 provinces—Allahābād, Āgra, Oudh (Awadh), Ajmer, Ahmadābād, Bihār, Bengal, Delhi, Kābul, Lahore, Multān, Mālwa, Qhāndesh, Berār, and Ahmadnagar. Kashmir and Kandahār were districts of the province of Kābul. Sind, then known as Thatta, was a district in the province of Multān. Orissa formed a part of Bengal. The provinces were not of uniform area or income. There were in each province a governor, a dewan (revenue and finance officer), a *bakhshī* (military commander), a *ṣadr* (religious administrator) and *qāẓī* (agents who supplied information to the central government).

In all parts of the empire there were many subordinate states, whose rulers enjoyed varying degrees of power and prestige. These states were reckoned as so many districts and were attached to the *ṣūbah*s (provinces) within the boundaries of which they happened to be situated.

The provinces were divided into districts (*sarkār*s). Each district had a *fowjdār* (a military officer whose duties roughly corresponded to those of a collector); a *qāẓī*; a *kotwāl,* who looked after sanitation, police, and administration; a *bitikchī* (head clerk); and a *khazānedār* (treasurer).

Every town of consequence had a *kotwāl.* The village communities conducted their affairs through *pañcāyat*s (councils) and were more or less autonomous units.

Fiscal system. The main sources of Akbar's income, besides the land revenue, were from forests, irrigation canals, and fisheries, tributes from feudatory princes, a salt tax, and customs duties. The main features of the land revenue administration were measuring and classifying land into four categories, ascertaining the actual produce of the soil, and preparing separate schedules of rates for produce and of government levies for the fiscal *dastūr*s into which the empire was divided. The average produce and demands were based on the previous 10 years' produce and prices; this was known as the 10-year settlement. The government's share was one-third of the produce. Definite rules were laid down for the guidance of collectors. Akbar reformed the currency to make it one of the best then known in the world.

The army. Mostly foreign in personnel, Akbar's army was originally composed of Mongols, Persians, Turks, Uzbeks, and Afghans, who were granted large assignments of land in lieu of salaries. The cavalry was the most important branch, followed by the artillery; the infantry was of little consequence. In 1573 Akbar first introduced the branding of horses and thereafter the *manṣabdārī* system, which graded the list of the officers, or *manṣabdār*s (holders of ranks), ranging from the rank of 10 to that of 5,000, subsequently raised to 12,000. There had already been a kind of gradation system, but with Akbar's innovation promotion or demotion from one rank to another was a normal feature. The *manṣabdār*s were required to maintain a fixed number of horsemen, horses, elephants, camels, and carts appropriate to their individual ranks; to bring their contingents for periodical musters; and to get their horses branded.

The *manṣabdārī* system

There were 36 grades of *manṣabdār*s; besides *manṣabdār*s, there were *aḥadī* and *dākhilī* troops, who belonged to a higher rank and were paid higher salaries. The army also used war elephants. There was no navy, but there were boats commodious enough to carry a few elephants each.

The *manṣabdārī* system, though not free from defects, was an improvement over the earlier system of tribal chieftainship and the organization of levying contingents. The *manṣabdār*s and their troops were under strict discipline and control of the central authority and were well paid.

The Mughal imperial service was organized on bureaucratic principles and was military in character and outlook. There was no distinction between military and civil services. With the exception of the *qāẓī*s, the *ṣadr*s, and some other religious officers, all had to be enrolled as *manṣabdār*s. Their salaries were determined and their ranks in the service fixed by their numerical designations. There was no specialization of qualification in appointments, and members could be transferred from one duty to another of an entirely different nature. There were no separate educational or medical services, though teachers, physicians, and artists were employed as imperial servants.

Repeal of discriminatory laws

Judicial reforms. Akbar repealed discriminatory laws against non-Muslims and amended the personal laws of both Muslims and Hindus so as to provide as many common laws as possible. He decreed that a man should not marry more than one wife unless his first wife was barren, forbade marriage between cousins and near relatives, and ordered that boys were not to marry before the age of 16 and girls before that of 14. He raised the age of circumcision to 12. The *kotwāl*s in the cities were charged with the duty of enforcing these laws. The organization of judicial courts was allowed to remain as it had been under his predecessors, and decisions of the village *pañcāyat*s was recognized. Akbar reformed the mode of conducting investigation and insisted that the judges try to find out the truth by every possible device and that they "should not be satisfied with witnesses or oaths, but pursue them by manifold enquiries by the study of physiognomy and the exercise of foresight . . ."

Religious policy. Akbar was interested in the study of the great problems of life and death. Early in 1575 he established at Fatehpur Sīkri a religious assembly hall (ʿIbādat-khānah) where discussions on Islāmic religion, law, and tradition were held. He soon became dissatisfied with what he considered the shallowness of Muslim divines and threw open the meetings to learned men of other religions. He invited the Zoroastrian priest Mahyārjī Rānā from Navsāri in Gujarāt in 1578–79, three Christian missions from Goa (1580–82, 1591, 1595–1605), and three Jain missions (1582–84, 1592–93, 1593–95) and acquired a good knowledge of these religions. He had already established contact with Hindu pandits. A comparative study of religions convinced him that there was truth in all of them, but that no one of them possessed absolute truth. He therefore disestablished Islām as the religion of the state, extended equal patronage to all faiths, and permitted legitimate religious propaganda and conversion to all. He rejected the Islāmic doctrine of resurrection and judgment and brushed aside revelation. He adopted many Hindu and Parsi beliefs and customs, such as belief in the doctrine of transmigration of the soul and sun worship. He did not subscribe to the five fundamentals of Islām, and he believed in Muḥammad as a prophet but not as the only prophet or the last and greatest prophet. He prayed, unlike a Muslim, four times—at sunrise, noon, sunset, and midnight. But he did not publicly repudiate his belief in Islām; nor did he persecute it, as some modern scholars have assumed.

Dīn-e Ilāhī. Many years' close and comparative study of religions made Akbar feel the insufficiency of Islām for being the national religion of India with its vast Hindu population and its highly developed religions and cultural traditions. Nor did he find Hinduism, Jainism, Zoroastrianism, or Christianity suitable for that purpose. Moreover, the followers of each of these religions (except those of Hinduism) considered their respective faiths alone to be true and the rest unworthy of acceptance. There was consequently a great deal of discord in the religious life of India. This troubled Akbar, because national progress was being hampered on account of religious conflict and disharmony. So he sounded religious leaders of various communities and unfolded a scheme of an eclectic organization (Dīn-e Ilāhī) that would combine the merits of all religions and eliminate their defects. But it was opposed by Raja Bhagwantdās of Amber, and Akbar gave up the idea of pursuing it to its logical conclusion. Instead he created the Dīn-e Ilāhī on the model of a Ṣūfī order (Muslim mystical brotherhood). This new order had its own initiation ceremony and rules of conduct; otherwise, members were permitted to retain their diverse religious beliefs and practices. It was devised with the object of forging the diverse communities in the country's population into one people.

Religious conflict and harmony

Akbar had succeeded in bringing about the political, administrative, economic, and to some extent the cultural unification of northern India by introducing Persian as the only official language and by bringing about the fusion of Hindu and Muslim architecture, music, and painting and providing a common religious and secular literature. By introducing several social reforms, he attempted to make social unification a reality. But he believed that for this unification to be permanent it must be based on some kind of mutual religious-cum-cultural understanding and appreciation. In short, he wanted a common religious forum for at least the elite of the various sections of the Indian population. The time was not ripe for the successful implementation of an idealistic program of that type; but Akbar's motive, nevertheless, was exceptionally noble.

Akbar's significance. Akbar's idealism, natural gifts, and force of character as well as his concrete achievements entitle him to a high place among the rulers of mankind. He is considered the greatest ruler of the Mughal dynasty and one of the greatest in the entire history of India.

His last days were full of anxiety due to the death of his younger sons and the rebellion of his eldest son, Salīm, who held court at Allahābād as an independent prince. Akbar became furious, and proceeded toward Allahābād to chastise the prince, but he had to return on account of one mishap or another. He thought of appointing Salīm's son Khusraw as his successor; but Salīm submitted in time, and Akbar forgave him and formally appointed him his heir and successor. Akbar fell ill on October 3, 1605.

Adapted from C. Davies, *An Historical Atlas of the Indian Peninsula*, 2nd ed. (1959); Oxford University Press



India in 1605.

Physicians failed to diagnose the trouble correctly, and Akbar died at midnight of October 25–26.

THE EMPIRE IN THE 17TH CENTURY

The rebellion of Khusraw

Jahāngīr. Within a few months of his accession, Salīm (ruled 1605–27), who assumed the title of Jahāngīr, had to deal with a rebellion led by his eldest son, Khusraw. Khusraw was defeated at Lahore and was seized just after he had crossed the Chenāb, at the ferry of Shāhpur. He was brought in chains before the emperor at Lahore and partially blinded; his followers were given barbarous punishments. Jahāngīr imposed on the Sikh Gurū Arjun a fine of two lakhs of rupees for having blessed Khusraw. On the Gurū's refusal to pay, he was put to death; this act permanently estranged the Sikhs from the Mughals.

Khusraw's rebellion led to a few more risings, which were suppressed without much difficulty. Shāh ʿAbbās I of Persia, taking advantage of the unrest, besieged the fort of Qandahār (1606); but abandoned the siege, when Jahāngīr promptly sent an army against him.

Marriage to Nūr Jahān. One of the most important events of the reign was Jahāngīr's marriage to Mih-run-Nisāʾ. She had been appointed lady-in-waiting to Jahāngīr's stepmother; the emperor chanced to see her in a festival in March 1611, fell in love with her, and married her that May. The lady became his favourite wife and was honoured with the title of Nūr Jahān.

Nūr Jahān raised her father, brother, and other relations to high positions in the state. She enjoyed great influence and authority, formed a clique with her parents, her brother Āṣaf Khān, and Prince Khurram (third son of Jahāngīr), and became a power behind the throne; Khurram, later called Shāh Jahān, married Āṣaf Khān's daughter Arjumand Bānū Baygam, who was later given the title of Mumtāz Maḥal. Nūr Jahān exercised a strong influence on her husband and looked after him with unparalleled care and devotion. Under her influence Jahāngīr restrained himself from excessive drinking. She relieved him of much of the drudgery of administrative routine and anxiety. She enhanced the splendour of the Mughal court and ably seconded the efforts of her husband in patronizing learning and art and disbursing charity. But in politics and administration she encouraged factionalism. From 1611 to 1622 she supported the claims of Khurram to be the heir; but after her daughter by an earlier marriage, Lādlī Baygam, was wed to Prince Shahryār, Nūr Jahān began to back the latter and tried to reduce Khurram's influence. This drove Khurram into a rebellion against his father, and there was a prolonged civil war.

Nūr Jahān's influence

Subjugation of Mewār and Kāngra. In pursuance of his father's policy, Jahāngīr dispatched five successive expeditions to subdue Rānā Amar Singh of Mewār, but all these failed. In 1613 Khurram was given the supreme command of the army, and Jahāngīr himself marched to Ajmer to be near the scene of action. The Rānā was then obliged to open negotiations, and Jahāngīr was happy to accept the terms proposed (1615). Amar Singh recognized Jahāngīr as his suzerain and all his territory in Mughal possession was restored, including Chitor, which, however, was not to be fortified. The Rānā was not obliged to attend the imperial *darbār* (administrative council), but his son was to represent him; nor was he required to enter into a matrimonial alliance with the Mughal royal family. Jahāngīr naturally claimed success against the Rānā, whom his great father had failed to subdue. He also took credit for the conquest in 1620 of the hill state of Kāngra with the sacred shrine of Jwālāmukhī, which also had defied all previous Muslim invaders, including Akbar. Jahāngīr was right in boasting that except for Assam the whole of northern India was now in his possession.

War in the Deccan and the loss of Qandahār. Jahāngīr considered it his duty to complete the conquest of the Deccan, where, as was mentioned above, he was opposed by Malik ʿAmbar of Ahmadnagar. In 1620–21, Malik ʿAmbar was forced to surrender extensive territories to the Mughals; Bījāpur and Golconda also agreed to pay tribute.

Shāh ʿAbbās' siege of Qandahār

In 1622 ʿAbbās of Persia again besieged Qandahār, and Khurram (now Shāh Jahān) was directed to relieve that fortress. But the prince was planning a rebellion against his father and failed to take effective action. The fortress fell after 45 days' siege. Shāh ʿAbbās justified its capture on the plea that it belonged to Persia. Jahāngīr accused the Shāh of treachery and sent forces to recover the fortress, which could not be done successfully owing to Shāh Jahān's rebellion and the illness and death of Jahāngīr himself.

Rebellion of Shāh Jahān. Alienated by the hostility of Nūr Jahān, Shāh Jahān raised the standard of revolt in 1623. After failing to take Fatehpur Sīkri, in April, he retreated to the Deccan, thence to Bengal, and from Bengal back again to the Deccan, being pursued all the time by an imperial force. His plan to seize Bihār, Oudh, Allahābād, and even Āgra failed; and Mahābat Khān, the imperial general, inflicted defeat after defeat on the fugitive. At last Shāh Jahān submitted to his father unconditionally. Nūr Jahān, who was becoming suspicious of Mahābat Khān's fast-rising power and prestige, demanded that Shāh Jahān surrender Rohtās and Asīr, send Dārā and Aurangzeb, two of his elder sons, to court. Shāh Jahān complied without hesitation. He was forgiven and appointed governor of Bālāghāt. The three-year-old rebellion, which had caused a considerable loss of men and money, came to an end in April 1626.

Mahābat Khān's coup. Immediately on the conclusion of peace with Shāh Jahān, the imperious Queen decided to punish Mahābat Khān for his refusal to take orders from anyone but Jahāngīr and to support Shahryār's claim to the throne. She ordered Mahābat Khān to Bengal and framed charges of disloyalty and disobedience against him. Instead, Mahābat Khān proceeded to Punjab, where the Emperor was encamped. Jahāngīr refused to see him, and the Queen insulted him further by asking him to explain why he had affianced his daughter without previous permission. Mahābat Khān was enraged and decided to secure the person of the Emperor and destroy Nūr Jahān's power. He appeared before the imperial camp at the head of his powerful army, took Jahāngīr back to his own camp, and placed him under surveillance. Nūr Jahān then voluntarily surrendered to Mahābat Khān and was permitted to join Jahāngīr. They remained under the custody of Mahābat Khān for about two months, and they both apparently reconciled themselves outwardly to the general's domination; but Nūr Jahān hatched a plot to secure the release of the Emperor. One day Jahāngīr planned to review his troops and sent word to Mahābat Khān to keep his army at a little distance so as to prevent any possibility of a clash between the two forces. During the review Jahāngīr joined his men, escaping his captor. Thus outwitted, Mahābat Khān submitted to Jahāngīr. But the emperor's health was greatly impaired, and he died on November 7, 1627, near Bhimbar.

Capture of Jahāngīr

Evaluation of Jahāngīr. Although he had rebelled against his father, Jahāngīr cherished the loving memory of Akbar and held him in a great reverence. Mughal painting and music reached high standards of development under his patronage. He usually followed his father's policy of religious toleration and, although he did not have complete faith in the dogmas of any religion, he was certainly a liberal Muslim. He was a fairly successful ruler and a just administrator, and had the good sense to continue the administrative system and policies of his father.

Shāh Jahān. Shāh Jahān (ruled 1628–58) had already had his elder brother Khusraw strangled to death (1622); and under his instructions his father-in-law Āṣaf Khān slew all other royal princes, the possible rivals for the throne. On Shāh Jahān's accession Āṣaf Khān was appointed prime minister and Nūr Jahān was given an adequate pension.

Rebellions against Shāh Jahān

Shāh Jahān's reign was marred by a few rebellions, the first of which was that of Khān Jahān Lodī, governor of the Deccan, who was recalled to court after failing to recover Bālāghāt from Ahmadnagar. He rose in rebellion and fled back to the Deccan, and Shāh Jahān followed, and in December 1629 defeated Khān Jahān and drove him to the north, ultimately overtook him, and killed him in a skirmish at Shihondā in what is now the Bānda district of Uttar Pradesh (January 1631).

The next rebellion was led by the Hindu Orchha chief

Jujhār Singh in Bundelkhand. Jujhār was compelled to submit after his kinsman Bhārat Singh defected and joined the Mughals. Jujhār was pardoned on condition of surrendering a part of his territory and proceeding to the Deccan for imperial service. On his return to Orchha he attempted to seize Gondwana; but he was asked to surrender it and pay a fine. Jujhār refused to comply and recalled his troops from the royal service. After a prolonged campaign, he was defeated and murdered (1634). Two of Jujhār's sons were converted to Islām; a third was put to death because he refused to convert. Many Hindu temples in Bundelkhand were wantonly desecrated and demolished. Orchha was made over to Devī Singh, another relative of Jujhār Singh's, who had gone over to the Mughals; the Bundelās refused to acknowledge him as their leader, and trouble continued for some time.

Meanwhile in 1632 Shāh Jahān became involved in a dispute with the Portuguese of Hooghly in Bengal, who were indulging in piracy and who had stolen two slave girls belonging to the Mughals. The Portuguese were obliged to deliver up the slaves and pay a fine, and 400 of them were brought to Āgra, where they were imprisoned when they refused to embrace Islām.

In the Deccan, Shah Jahān's forces captured Daulatābād, the new capital of the Niẓāmshāhī kingdom. This, however, did not bring about the complete extinction of that kingdom; it kept the northern parts of what is now the Pune (Poona) district and the whole Konkan.

Settlement in the Deccan

Bijāpur submitted, acknowledged Mughal suzerainty, and once again agreed to pay an annual tribute. Golconda's sultan also agreed to pay tribute and to assist the Mughals against Bījāpur in case of a war with the latter (May 1636). Relations with the Marāṭhās (see below *The Marāṭhās*) also were settled, but not without fighting. For the time being Shāh Jahān was able to settle the affairs of the Deccan to his satisfaction.

Early in 1638 Shāh Jahān succeeded in recovering Qandahār, which had been lost during the last years of Jahāngīr's reign. The Persian governor, ʿAlī Mardān Khān, was on poor terms with the court of Tehrān; he intrigued with the Mughal officers and surrendered the fortress to Shāh Jahān. In February 1649 the Persian Shāh recaptured Qandahār after a brief siege. Shāh Jahān made three unsuccessful attempts (1649, 1652, 1653) to recover the fortress. The relations between India and Persia, therefore, became strained.

Central Asian policy. Following in the footsteps of his predecessors, Shāh Jahān hoped to conquer Samarkand, the original homeland of his ancestors. The brother of Emām Qolī, ruler of Samarkand, invaded Kābul and, in 1639, captured Bāmīān, which gave offense to Shāh Jahān. The Emperor directed his son Murād to invade Badakhshan; but Murād failed (1646) and was replaced by another son, Aurangzeb. The latter fought bravely, but the situation was complicated by an alliance between the Uzbeks and ʿAbbās II of Persia, and Shāh Jahān's Central Asian policy ended in failure.

War of succession, 1657–59. Shāh Jahān fell ill in September 1657, and rumours spread that he was dead. He executed a will bequeathing the empire to his eldest son, Dārā. His other sons, Shujāʿ, Aurangzeb, and Murād, who were grown men and governors of provinces, decided to contest the throne. Shujāʿ, governor of Bengal, was defeated by an army sent by Dārā at Bāhādurpur, five miles northeast of Vārānasi (February 24, 1658), and was pursued to the border of Bengal. Murād and Aurangzeb joined their armies and marched to Dharmat, 14 miles southwest of Ujjain. There they defeated an imperial force sent by Dārā (April 25).

Dārā made another effort to drive the two princes back to the Deccan, but he was beaten on June 8 at Samūgarh, east of the Āgra fort. Dārā retreated to Āgra, covered with humiliation, and fled to Delhi and thence to Lahore, Gujarāt, and Rājasthān. He was subsequently defeated in another battle at Deorai near Ajmer and fled to take shelter in the northwest frontier; but he was betrayed and executed on the trumped-up charge of heresy. Aurangzeb had already disposed of Murād (July 5) and had beaten off Shujāʿ, who fled to Arakan and was killed there; Aurangzeb thus became the sole victor. He then imprisoned Shāh Jahān in the Āgra fort and himself became emperor.

Victory of Aurangzeb

Shāh Jahān died a prisoner on February 1, 1666, at the age of 74. He was on the whole a tolerant and enlightened ruler, patronizing scholars and poets of Sanskrit and Hindī as well as of Persian. He systematized the administration, but he raised the government's share from one-third to one-half of the gross produce of the soil. He was fond of pomp and magnificence. He commissioned the famous peacock throne and erected many elegant buildings, including the lovely Tāj Mahal outside Āgra, a tomb for his queen, Mumtāz Maḥal.

Aurangzeb. Aurangzeb (ruled 1658–1707) transferred the capital from Āgra to Delhi and went through a hurried coronation there on July 31, 1658. Having gained the throne as the champion of orthodox Islām, he appointed a censor of public morals (*muḥtasib*) in every large city to put down such forbidden practices as drinking, gambling, and sexual immorality; he punished heretical opinions, blasphemy, and the omission of five daily prayers and the fast of Ramaẓān (Ramaḍān) by Muslims. Aurangzeb deliberately reversed the policy of his predecessors toward non-Muslim subjects by enforcing the principles and practices of the Islāmic state. He reimposed the *jizyah* on non-Muslims and saddled them with religious, social, and legal disabilities. To begin with he forbade their building new temples and repairing old ones. Next, he issued orders to demolish all the schools and temples of the Hindus and to put down their teaching and religious practices. He doubled the customs duties on the Hindus and abolished them altogether in the case of Muslims. He granted stipends and gifts to converts from Hinduism and offered them posts in public service, liberation from prison in the case of convicted criminals, and succession of disputed estates. He persecuted the Shīʿah and the Ṣūfīs (nonconformist Muslims) as well.

Aurangzeb's policy toward non-Muslims

This policy of fanaticism led to numerous rebellions. The Jāṭ peasantry of Mathura rebelled in 1669; a Satnāmī rising occurred in 1672; the Sikhs in the Punjab revolted under their Gurū Tegh Bahādur, who was brutally put to death in 1675. But the most prolonged uprising was the Rājput rebellion, caused by Aurangzeb's annexation of the Jodhpur state and his seizure of its ruler's posthumous son Ajit Singh with the intention of converting him to Islām. This rebellion spread to Mewār; and Aurangzeb himself had to proceed to Ajmer to fight the Rājputs, who had been joined by the Emperor's third son, Akbar (January 1681). But by a stratagem Aurangzeb managed to isolate Akbar, who fled to the Deccan and thence to Persia. The war with Mewār came to an end (June 1681) because Aurangzeb had to pursue Akbar to the Deccan, where the prince had joined the Marāṭhā king Sambhājī. But Jodhpur remained in a state of rebellion for 27 years more, and Ajit Singh occupied his ancestral dominion immediately after Aurangzeb's death.

Aurangzeb's conquest of Assam proved to be transitory, and his war with the frontier tribes sapped the resources of the empire. But in 1661 Palāmau, or the Rānchi region, was annexed to Bihār, and Mughal suzerainty was reestablished in Ladākh.

Aurangzeb spent the last 25 years of his reign in the Deccan, and invaded and annexed Bijāpur (1686) and Golconda (1687). Then he concentrated his might against the recently established Marāṭhā kingdom, but Marāṭhā resistance was so stubborn that even after nearly two decades of struggle he failed to completely subdue them (see below). The aged Emperor died on February 20 (old style March 3), 1707, and was buried at Khuldābād, four miles west of Daulatābād.

Aurangzeb possessed natural gifts of a high order. He had assiduously cultivated learning, self-knowledge, self-reverence, and self-control, and he exercised a curb over his tongue and temper. He was extremely industrious, methodical, and disciplined in habits and thoughts, and his private life was virtuous. But his narrow religious bigotry made him ill-suited to rule the mixed population of his empire. He treated all non-Muslims in the country as inferior people. His ambition was to convert India into a purely Muslim country; but in fact he destroyed rather

Assessment of Aurangzeb

than built. He was suspicious by nature. Besides killing his brothers, he imprisoned his three sons, the eldest of whom died in confinement. Similarly, his daughter Zīb-un-Nisā' ended her life in prison.

A recent attempt has been made to show that Aurangzeb was not as bigoted a Sunnī Muslim as has been represented, and that he granted land to a famous temple of Vārānasi. A few instances of this kind—which to be sure were due to political reasons—would not make Aurangzeb appear tolerant, when abundant, unimpeachable contemporary evidence proves that he destroyed thousands of Hindu temples and schools and forcibly converted Hindus to Islām. The idea that he pursued these policies only in wartime is a pure invention of his modern apologists and finds no confirmation in contemporary writings.

MUGHAL DECLINE IN THE 18TH CENTURY

Bahādur Shāh I (Shāh ʿĀlam). Aurangzeb's eldest son had died in prison. His second son, Muʿaẓẓam Bahādur Shāh, or Shāh ʿĀlam (ruled 1707–12), aged over 63 years, gained the throne after defeating and disposing of his two brothers, Āʿẓam and Kām Bakhsh. He recognized Ajit Singh, in rebellion since birth, as the Maharājā of Jodhpur and made peace with other Rājputs. Then he tried to settle his scores with the Sikhs, who also had rebelled in the Punjab under their leader Bandā Singh Bahādur. After a good deal of initial success, Bandā was defeated and driven into the hills of Jammu. But when Bahādur Shāh died in February 1712, Bandā recovered Sādhaura and Longarh and continued his depredations as before.

Bahādur Shāh was a mild and generous man. Though possessed of dignity of behaviour, he proved to be a weak ruler. He was fond of compromise even in important political and administrative matters. But he followed his father's policy of intolerance, retained the *jizyah,* and did not appoint Hindus to high posts. His short reign was on the whole tolerably successful.

Jahāndār Shāh. There were three wars of succession among Bahādur Shāh's sons; and the eldest son, Jahāndār Shāh (ruled 1712–13), triumphed. He was an incompetent ruler who neglected the business of government. He bestowed his favours on his concubine Lāl Kowr and appointed her low relatives to high posts. The *wazīr* Ẕūlfiqār Khān sulked at home, thwarted in administrative business by musicians and buffoons. There was confusion all around.

The Sayyid brothers

Jahāndār's claims were disputed by his nephew Farrukhsiyar. He defeated Jahāndār Shāh with the help of the Sayyid brothers (ʿAbdullāh Khān and Ḥusayn ʿAlī Khān) and put him to death on February 11, 1713. Jahāndār was the first ruler of the Mughal dynasty to prove utterly incapable of administering the affairs of the empire and to behave like a foolish upstart.

Farrukh-siyar. Farrukh-siyar (ruled 1713–19) was a young man of 30 at the time of his accession. He was utterly weak, thoughtless, and devoid of physical and moral courage; and his reign was one long attempt at perfidious conspiracy against his *wazīr* ʿAbdullāh Khān and his brother Ḥusayn ʿAlī Khān—the two Sayyid brothers who had been instrumental in raising him to the throne.

The Sayyids reduced Ajit Singh of Jodhpur (May 1714). Another expedition defeated Bandā, the Sikh leader, who was put to a cruel death at Delhi (June 1716). But the Jāṭ leader Churāman could not be suppressed and peace had to be made with him. By this time the quarrel between the Emperor and the Sayyid brothers had come to a head. Ḥusayn ʿAlī Khān brought a Marāṭhā force to Delhi, deposed the Emperor, and placed Rafīʿ-ud-Darajāt on the throne (February 1719). This boy king was suffering from consumption, and he was removed from the throne. His elder brother Rafīʿ-ud-Dawlah was crowned king on June 16, 1719, with the title of Shāh Jahān II. Rafīʿ-ud-Dawlah too was a sickly youth, and he died on September 17, 1719.

Muḥammad Shāh. On September 28 the Sayyids crowned Rawshan Akhtar (ruled 1719–48), a grandson of Bahādur Shāh, emperor under the title of Muḥammad Shāh. He was a weak and inexperienced prince, and the Sayyids retained all power in their hands. The powerful Niẓām al-Mulk I, Mughal viceroy of the Deccan, who was hostile to the Sayyids, captured Khāndesh and defeated and killed two nephews of the Sayyid brothers. Therefore, Ḥusayn ʿAlī Khān and the Emperor proceeded to the Deccan to put down Niẓām al-Mulk. On the way Ḥusayn ʿAlī fell victim to a conspiracy (October 1720). Muḥammad Shāh then turned back toward Delhi and defeated and imprisoned the elder Sayyid, ʿAbdullāh Khān (November).

Muḥammad Shāh appointed Niẓām al-Mulk *wazīr* of the empire, but the latter found it difficult to function in the midst of youthful, pleasure-loving courtiers and returned to the Deccan (December 1723). Qamar-ud-Dīn Khān was, therefore, appointed *wazīr.* The Niẓām practically became independent in the Deccan (October 1724). Another adventurer, Saʿadat Khān, who was appointed governor of Oudh in September 1722, laid the foundation of the virtually independent Shīaʿh dynasty of that province. Bengal, Bihār, and Orissa, which constituted a viceroyalty during the reign of Bahādur Shāh, became practically independent at this time. The Marāṭhās diverted their attention to the north, beginning in 1731 annual raids into the Mughal territory in northern India.

Invasion by Nāder Shāh

The central government had become so weak that the country was invaded in 1739 by a Persian adventurer, Nāder Shāh, who had established himself as the ruler of Iran. He captured Qandahār, Ghazna, and Kābul and crossed the Indus at Attock (December 1738). He overpowered the governor of Lahore and defeated Muḥammad Shāh near Karnāl (February 1739). Nāder then marched to Delhi and ordered a general massacre in which 30,000 people were slain. He left Delhi on May 26, laden with a booty amounting to 30 crores of rupees (300,000,000) in cash, along with jewels, gold and silver plate, and other valuable articles, including the famous peacock throne of Shāh Jahān. He deprived the empire of the province of Kābul, which he annexed to Iran.

The invasion paralyzed Muḥammad Shāh and his court and heaped humiliation upon the country. But the Emperor failed to take a lesson from it. Nothing was done to reorganize and reform the administration. Marāṭhā raids on Mālwa, Gujarāt, Bundelkhand, and the territory north of these provinces continued as before. The Emperor was compelled to appoint the Marāṭhā peshwa, Bālājī Bājī Rāo, as governor of Mālwa. The province of Katēhr (Rohilkhand) was seized by an adventurer, ʿAlī Muḥammad Khān Ruhēla, who could not be suppressed by the feeble government of Delhi. The Punjab was invaded by Aḥmad Shāh ʿAbdālī, a lieutenant of Nāder Shāh who became king of Kābul after Nāder's death (June 1747); he sacked Lahore but was defeated by a Delhi army at Mānupur near Machhīwara and compelled to retreat (March 1748). Muḥammad Shāh died in April 1748.

Aḥmad Shāh. Upon Muḥammad Shāh's death, his son Aḥmad Shāh (ruled 1748–54) crowned himself king. He was a young man of 21 who possessed no qualities of leadership. His *wazīr*s, Ṣafdar Jang of Oudh, and Mīr Bakhshī Saādāt Khān, were also mediocre. Ṣafdar Jang tried to crush the Ruhēla and Bangash Pathans, who had usurped Rohilkhand and Farrukhābād, but suffered a defeat (September 1750). He undertook a second expedition against Rohilkhand and defeated the Ruhēlas with Marāṭhā assistance (April 1751). He returned to Delhi to find that the Emperor had made peace with ʿAbdālī, who had invaded northwestern India a third time, and ceded to him the Punjab and Multān (March 1752). The *wazīr,* therefore, had the Emperor's favourite adviser murdered, which led to a complete breach between the Emperor and his *wazīr* and resulted in a civil war between them. Ṣafdar Jang was defeated and left for Oudh (November 1753). Power passed into the hands of ʿImād-ul-Mulk, a grandson of Niẓām-ul-Mulk, who entered into an alliance with the Marāṭhās and with their help became *wazīr.* He deposed the Emperor on June 2, 1754, and placed ʿĀlamgīr II (a son of Jahāndār Shāh) on the throne.

Restoration of Muslim rule

ʿĀlamgīr II. ʿĀlamgīr (ruled 1754–59) was as inept as his predecessor. During his reign the fourth ʿAbdālī invasion took place. The invader was provoked by ʿImād-ul-Mulk's seizure of the Punjab, which had been ceded to ʿAbdālī in 1752. After driving the *wazīr*s' agent from Lahore (November 1756), ʿAbdālī was pressed by Shāh Walī Allāh, the most important Muslim Ṣūfī saint of

northern India, to move down to Delhi and help restore effective Muslim rule in the country by crushing the Marāṭhās, who were dominating all northern India, and the Jāṭs, who had brought the entire country from Bharatpur to Ballabgarh under their rule. ʿAbdālī stayed in India for a few months, married his son Timur to the daughter of ʿĀlamgīr II, and sent an army against Surajmal Jāṭ. He sacked Mathura and massacred a large number of innocent pilgrims. He then entrusted to the Ruhēla captain Najīb-ud-Dawlah (Najīb Khān) the protection of ʿĀlamgīr II. Najīb proved to be an exacting and rigorous jailer, who meted out rough treatment to ʿĀlamgīr II. There was no love lost between him and the *wazīr*, ʿImād-ul-Mulk, who had once again regained power with Marāṭhā assistance.

The Marāṭhā influence was now reestablished at the capital, and their leader Raghunāth Rāo marched to the Punjab and drove away ʿAbdālī's son and agent (April 1758). ʿImād-ul-Mulk expelled Najīb to his *jāgīr* (Sahāranpur and Najībābād) and drove the Emperor's eldest son, ʿAlī Gawhar, from Delhi. The prince fled to Sahāranpur, from there to Oudh, and thence to Allahābād. The Marāṭhā power had now reached its zenith. Their only potential enemy in northern India was Najīb-ud-Dawlah. Therefore Dattājī Sindhia besieged him at Shakartāl, 18 miles west of Muzaffarnagar, and the siege dragged on through the rainy season of 1759. Najīb appealed to his kinsmen of Rohilkhand—to Shujāʿ-ud-Dawlah of Oudh and Aḥmad Shāh ʿAbdālī of Kābul—to save him from destruction. Shāh Walī Allāh once again besought ʿAbdālī to come and uproot the Marāṭhās and the Jāṭs, and promised him heavenly assistance in this sacred mission. So ʿAbdālī undertook his fifth invasion of India. The *wazīr* ʿImād-ul-Mulk, alarmed at the news, had ʿĀlamgīr II assassinated on November 29, 1759. He proclaimed Muḥī-ul-Millat, grandson of Kām Bakhsh, as emperor under the title of Shāh Jahān III. After this he marched toward Shakartāl to join Dattājī Sindhia, who now raised the siege and proceeded toward Lahore to fight the invader.

ʿAbdālī entered Punjab in August 1759 and drove away the Marāṭhā governor. From Lahore he proceeded toward Delhi, sending a detachment of his army to fight Dattājī Sindhia, who was marching toward Sirhind. He defeated and killed Dattājī at Barārī Ghāt, 10 miles north of Delhi, and then occupied Delhi.

Meanwhile, the peshwa had dispatched a strong army under his cousin Sadāshiv Rāo Bhāo to drive away the invader and reestablish Marāṭhā supremacy in northern India. Bhāo was joined by Surajmal of Bharatpur; but in view of the Marāṭhā record in the north, all Rājput and other Hindu chiefs declined to join him. His attempt to win over Shujāʿ-ud-Dawlah of Oudh also failed, and the latter joined the invader, who claimed his cause as one of Islām. Surajmal too deserted the Marāṭhās because of differences over general policy and the mode of warfare to be followed. Bhāo expelled ʿAbdālī's agent from Delhi, deposed Shāh Jahān III, and proclaimed the accession of ʿĀlamgīr II's son Shāh ʿĀlam II, who was in the east, planning to invade Bihār. ʿĀlam had already proclaimed himself emperor after his father's death; he made Shujāʿ-ud-Dawlah his *wazīr*.

Accession of Shāh ʿĀlam II

At the end of the rainy season the Marāṭhās marched from Delhi toward Sirhind in order to occupy the Punjab and cut off ʿAbdālī's retreat. On October 17 Bhāo captured the fort of Kunjpurā, six miles northeast of Karnāl, and proceeded to Sārhind. ʿAbdālī, who had spent the rainy season of 1760 at Sikandarābād, crossed the Yamuna at Bāghpat, 25 miles north of Delhi. On learning of the Marāṭhā movement toward the Punjab, he deputed a contingent of his troops to pursue Bhāo. The latter thereupon turned back and encamped at Pānīpat. ʿAbdālī arrived there three days later. The Marāṭhās fortified their camp and decided to fight.

The armies lay facing each other for two months, with daily skirmishes occurring between their scouts. The final battle was fought on January 14, 1761, in which the Marāṭhā army, which had suffered two months' starvation and lack of equipment, was completely routed. Bhāo died, fighting bravely to his last breath. The peshwa's son Vishwās Rāo, who was the generalissimo of the army, was also slain along with most of the notable officers and chiefs; Mahādājī Sindhia and Malhār Rāo Holkar alone among the notables escaped from the field. The third battle of Pānīpat shattered the dream of a Marāṭhā empire for all India. It cleared the way for the British, who had already usurped the position of making and unmaking nawabs of Bengal, Bihār, and Orissa, to the sovereignty of northern India. Aḥmad Shāh ʿAbdālī was not anxious to follow up his victory and to occupy and rule India. His troops clamoured and demanded arrears of pay, compelling him to return to Kābul. He recognized Shāh ʿĀlam as emperor and ʿImād-ul-Mulk as *wazīr*. He delivered charge of Delhi to Najīb-ud-Dawlah. ʿAbdālī's attempt to conclude a peace with the peshwa and Surajmal of Bharatpur failed, and he left Delhi for Kābul in March 1761.

Shāh ʿĀlam being away in Bihār, the throne of Delhi remained vacant from 1760 to 1771. During most of this period Najīb-ud-Dawlah was in charge of the administration of the capital and the dwindling empire. He acted like a virtual dictator and carried on a ceaseless warfare with the Jāṭs and the Sikhs but failed to crush either of them. The nominal empire at that time consisted of Delhi and a small area around it. (A.L.S.)

The Marāṭhās

EMERGENCE OF THE MARĀṬHĀ EMPIRE

Mahārāshtra: the land and its people. Mahārāshtra, homeland of the Hindu Marāṭhās, which the great Marāṭhā leader and hero Śivājī (1627–80) made the base of an empire, extends from the Arabian Sea on the west to Sātpura Mountains in the north. In the 17th century, Mahārāshtra was surrounded by the independent Muslim kingdoms of Bijāpur and Golconda in the south, by the Mughal Empire in the north, and by a number of European settlements on the west. The Muslim states of the south were in a process of disintegration, and the Mughal Empire, though at its largest during Śivājī's time, declined rapidly after the death of the last notable Mughal emperor Aurangzeb in 1707. The vacuum created by the eclipse of the Mughal Empire remained unfilled until the beginning of the 19th century, when the British began to consolidate their position. In this period between the decline of the Mughals and the establishment of British dominance, the Marāṭhās carved out their empire. A number of earlier ruling dynasties in the region had given it political unity, and the teachings and writings of a number of saints in Marāṭhī, a regional language, gave rise to a cultural and religious unity.

The career of Śivājī. Śivājī's father, Shāhjī Bhonsle, a *jāgīrdār* (holder of the right to collect the land tax from an assigned area), had been prominent in resisting the aggressiveness of the Mughal Empire, which then dominated India. Śivājī's mother, Jijā Bāī, developed in him high moral qualities and a great faith in religion, and all accounts agree in evaluating Śivājī as a born leader of men. In 1647 Śivājī, at the age of 20, assumed the administration of his *jāgīr* and, with the help of a number of neighbouring chieftains scattered over the Sahyādri Ranges, undertook the task of carving out a kingdom for himself in western Mahārāshtra. Fortunately for Śivāji, the neighbouring kingdom of Bijāpur was in decline. Its ruler, Muḥammad Shāh, suffered from a protracted illness and his kingdom was weakened by internal dissensions. Śivājī took advantage of the situation by building up an impregnable chain of strong fortresses, and by 1653 he had set up a kingdom that extended over the entire territory between the Bhīma and Nīra rivers.

Foundation of the Marāthā state

With a strong base in the Sahyādri Ranges, Śivājī turned his attention to the territory west of the Sahyādri. Between 1654 and 1660 he extended his kingdom on the west coast. Many willingly joined him; opponents, such as the Morays of Javli, were ruthlessly crushed. By 1657 Śivājī was raiding the Mughal territories in the direction of Junnar and Ahmadnagar and the Bijāpur territories in north Konkan. The coastal regions were being occupied at this time by European powers. The English set up a factory at Rājapur in 1648 and the Dutch settled down at Vengurla and the Portuguese at Chaul. It seems that Śivājī was aware of the danger of European power growing on

the west coast of India and decided to build up a strong naval power; he established a number of naval fortresses on the western coast over a period of some 30 years.

It was during one of his encounters with Bijāpur forces in 1659 that Śivājī killed Afẓal Khān, one of the leading generals of Bijāpur, who had been deputed to feign friendship with Śivājī and kill him. With alluring promises of confirming Śivājī in the possession of territories he had conquered from Bijāpur, Afẓal Khān, through his ambassador, invited Śivājī to meet him at the base of Pratāpgarh fort, 28 miles west of Wai. Apprised of Afẓal Khān's intentions, Śivājī reciprocated with an equally feigned fidelity and, in a physical embrace, killed him. He then followed up with a sharp attack on Afẓal Khān's army and the capture of a vast booty. Śivājī's army then poured into south Konkan and the Kolhāpur district and captured the fort of Panhāla. By this time, Shāyistah Khān had been appointed the Mughal governor of Deccan and, in conjunction with Bijāpur forces, was trying to bring pressure upon the Marāṭhās. Having driven the Marāṭhās from the neighbourhood of Poona (Pune), Shāyistah Khān was resting comfortably in a house in Poona, in which Śivājī had passed his early youth, when Śivājī made a surprise attack on him. Shāyistah Khān escaped with the loss of a thumb, but one of his sons and a large number of his followers were killed. Such adventures spread Śivājī's fame far and wide and made him feared by his adversaries. His sack of Surat, the largest trade mart on the west coast (January 1664), was an act of still greater audacity.

Relations with the Mughals. Having thus failed to check Śivājī, Aurangzeb, the Mughal emperor, turned to other tactics. He appointed Jai Singh, his ablest Hindu Rājput general and diplomat—a man of infinite tact and patience—to deal with him. Jai Singh succeeded in convincing the Bijāpuris of Mughal strength and generosity and won them over to his side. This intensified the Mughal thrust against Śivājī. Attacking the important Marāṭhā fort of Purandhar, Jai Singh forced Śivājī to sue for peace. Jai Singh seems to have decided, as a first step in his Deccan policy, to neutralize the Marāṭhās and then suppress the kingdom of Bijāpur. This policy suited the immediate interests of Śivājī also, for he was content to allow the Mughal armies to break the back of Bijāpur resistance. In June 1665 a treaty was signed at Purandhar between Śivājī and Jai Singh under which Śivājī was to surrender to the Mughals 23 out of 35 forts conquered by him in Bijāpur and to receive Bijāpur territories worth more than double that amount in lower Konkan and Bālāghāt. Śivājī was also to accept Mughal suzerainty and send his son Sambhājī as his representative to the Mughal court.

Aurangzeb, who distrusted Śivājī, did not accept Jai Singh's proposals with enthusiasm. Jai Singh's subsequent failure to deal effectively with Bijāpur discredited his policies and enabled Aurangzeb's advisors to attribute the failure to lack of support, even to treachery, on Śivājī's part. Under these circumstances, Jai Singh decided to arrange a personal meeting between Śivājī and the Mughal emperor, and, induced by Jai Singh, the Marāṭhā leader agreed to visit Aurangzeb at Āgra. But the visit misfired, and Jai Singh's entire Deccan policy collapsed. Treated as a hero by his countrymen, Śivājī did not expect to be treated as a mere local potentate, as he was in the eyes of the custodians of procedural propriety at the Mughal court. Finding that others had precedence over him in the audience hall, he flew into a rage, swooned, and was moved, at the Emperor's orders, into an anteroom. His behaviour shocked Aurangzeb, but, fearing that Śivājī's followers might ally themselves with Bijāpur and its supporters, he restrained himself and put Śivājī under simple confinement, from which he made a dramatic escape concealed in a fruit basket.

Śivājī at Aurangzeb's court

Returning from Āgra, Śivājī for some time avoided provoking the Mughals, limiting his energy to consolidation of his power in the immediate vicinity of his homeland. Nor did he repudiate the suzerainty of the Mughal Empire that he had accepted by the Treaty of Purandhar, and he was soon confirmed by the Mughals in his title of raja (king). After more than three years of truce with the Mughal emperor, Śivājī resumed his military adventures in 1670. He soon regained the whole territory of north Konkan, which had been ceded to the Mughals in the Purandhar treaty.

Śivājī's second offensive. On October 3, 1670, Śivājī carried out the second sack of Surat and exacted heavy tribute from the local merchants. On October 17, he invaded Bāglāna and defeated Dā'ūd Khān at Dindori. This was followed by invasions of Khāndesh and Berār. Realizing the gravity of the situation, Aurangzeb sent reinforcements to the Deccan, but, despite the great concentration of the Mughal power in that area, the Marāthās were able to capture the twin forts of Mulher and Salher early in 1672. The Mughal government soon revised its strategy, appointing the capable Bahādur Khān, as sole commander in the Deccan. This led to effective results. Bahādur Khān set up his headquarters at Padgaon, north of the Bhīma River, a place of great strategic importance.

Excluded from Mughal territory by Bahādur Khān, Śivājī now turned toward Bijāpur, where the ruler's death in December 1672 had thrown everything into confusion. Śivājī recaptured Panhālā from Bijāpur early in March 1673 and followed it up by the capture of Sātāra in July. The Mughal preoccupation with the Khyber Afghans in 1674, Aurangzeb's personal move to the northwestern frontier, and the removal of some Mughal forces from the Deccan again emboldened Śivājī to make inroads into Mughal territories. Crowning himself at Raigadh on June 6, 1674, with great pomp and ceremony, Śivājī again moved into Bāglāna and Khāndesh. Fearing that the Mughal effort to contain the Marāṭhās on the one side and control the Bijāpur kingdom on the other was merely leading to an increase in Śivājī's strength, Bahādur Khān revived Jai Singh's strategy and decided to concentrate on the suppression of Bijāpur and patch up his relations with Śivājī. Śivājī knew that once the Mughals succeeded in destroying Bijāpur it would be his turn next, and he was not confident of being able to withstand a frontal attack by the Mughals on his territories. He devised a new strategy; deciding to build up a second base for Marāṭhā power, in Karnātaka, inaccessible to the Mughals.

Śivājī's coronation

Conquest of Karnātaka. Śivājī thought that by allying himself with the kingdom of Golconda he would be able to exercise greater pressure on Bijāpur, and, while the Mughals captured its northern territories, he would be able to conquer for himself the southern and southeastern parts. A joint scheme for the conquest of Bijāpur territories in the south was finally evolved in a personal meeting between Śivājī and the Golconda ruler at Hyderābād. Śivājī also tried to win over to his side some Marāṭhā leaders allied with Bijāpur. His greatest success in the Karnātaka area lay in the conquest of Jinjī, which he made his principal seat of government in the south. Pressed by the Mughal armies, Bijāpur turned to Śivājī for help and confirmed him in the possession of all the territory that he had conquered from it. It was a personal tragedy for Śivājī that, at a time when he appeared to be at the height of his glory, his own son Sambhājī turned against him and joined the Mughals. Sambhājī soon fell out with the Mughals, however, and escaped to Panhālā, where he was kept in confinement. On his way back from Karnātaka, Śivājī met Sambhājī at Panhālā and tried his best to win him over, but without much success. Śivājī breathed his last on April 4, 1680, at Raigadh, at the age of 53.

Śivājī's achievements. Śivājī left behind him a well-organized army, a well-settled administration, and a well-endowed treasury. During the lifetimes of his immediate successors, however, the high principles on which he had set up his administration fell into disuse, his armies became instruments of aggression and vandalism, and his treasury was depleted. His kingdom was broken up once in his own lifetime and on several occasions thereafter, but the character of the political system that he established lasted for nearly 150 years.

Śivājī set up his state by welding a scattered group of clans into something resembling a nation, and infused it with an ideal. He was a deeply religious man who



The Marāṭhā kingdom at the death of Śivājī (1680).

Adapted from C.C. Davies, *An Historical Atlas of the Indian Peninsula* (1959); Oxford University Press

respected saintly character, irrespective of caste, community, or religion, but he did not allow religious or moral considerations to interfere with his statecraft and empire building. It was this combination of saintliness and political shrewdness, of gentleness, finesse, and wisdom, of military and administrative abilities, that gave Śivājī a unique position in Indian history.

Śivājī borrowed the pattern of his administration from the Mughals and also tried to bring about a number of reforms. He was against the *jāgīr* system and ruthlessly confiscated the land given to people under earlier regimes. He tried to abolish the numerous middlemen who had come between the cultivators and the government. He believed in making appointments on the basis of merit, which was an innovation. In the sphere of military administration, Śivājī's main emphasis was on the building of forts. He was a strict disciplinarian, severe punishment being inflicted on any soldier who violated the rules laid down by him. Unlike the Mughal armies, Śivājī's armies were extremely austere. His camp was generally without pomp, without women, and without baggage. His effort to raise a strong and independent navy of his own was but one example of his grasp of the situation.

It was the torch of revolt lighted by Śivājī that finally engulfed the entire country and consumed the mighty Mughal Empire, rendering completely ineffective all the efforts Aurangzeb made to save it.

Śivājī was succeeded by his son Sambhājī. It was at this time that one of Aurangzeb's sons, Prince Akbar, deputed to deal with the Rājputs (a group occupying a large area in the northeast), revolted against the Mughal emperor and took shelter with Sambhājī. An alliance between the Rājputs and Marāṭhās would have been very dangerous to the Mughal Empire, and Aurangzeb was quick to suppress the revolt of the Rājputs and march into the Deccan to take charge of the anti-Marāṭhā campaigns. Both Bijāpur and Golconda being in an advanced stage of dissolution, the powerful Mughal armies conquered them in quick succession, in September 1686 and October 1687, respectively. The Mughals then concentrated their strength against Sambhājī, who was captured by Aurangzeb's forces. He was dressed as a buffoon and forced to walk for miles along public roads lined with thousands of spectators. Despite this humiliation and even under torture, he refused to yield before Aurangzeb. His body was hacked to pieces along with those of a number of his followers, and their heads, stuffed with straw, were exhibited in the chief towns of the Deccan to the accompaniment of drum and trumpet.

Sambhāji's capture by the Mughals

The Marāṭhā war of independence. For Aurangzeb, however, it proved to be a Pyrrhic victory. The atrocities he had committed in the south aroused the Marāṭhā nation as nothing had done before. Rājarām, Sambhājī's younger brother, was quickly proclaimed *chhatrapati* (the Marāṭhā royal title). The Mughal forces continued to advance. A group of new Marāṭhā leaders took direction of civil and military affairs in an attempt to save the situation. Guerrilla warfare was employed and rendered the Mughal conquests useless. In the meantime, Rājarām was moved for security to the town of Jinjī in the far south, which resisted a Mughal siege for eight years. On Rājarām's death in March 1700, his widow, Tārā Bāī, took up the leadership of the entire movement. In the meantime (1689) Shāhū, the eldest son of Sambhājī, and his talented mother, Yesūbāī, had been captured by Aurangzeb and were moving with his armies from place to place.

The Deccan wars proved to be the undoing of Aurangzeb. Lasting for over 20 years, they were immensely costly, and officers who had marched with the Emperor in their youth to the south became old and worn out with no victory in sight. Aurangzeb's death on February 20 (old style March 3), 1707, marked the end of the effort. While the Mughal Empire stood on the verge of collapse, the Marāṭhās were filled with new hope and vigour.

Bālājī Viswanāth as peshwa. On Aurangzeb's death, Shāhū was able to escape from the Mughal camp and was crowned as *chhatrapati* by the Marāṭhās. His succession was challenged by Rājarām's widow, Tārā Bāī, and he had to spend his energies in internecine quarrels for many years. Still, he was fortunate in having as his main supporter Bālājī Viśvanāth, a Brahmin, who was first appointed as *senākarte* (organizer of the army) and rose to be the peshwa (prime minister). It was with the help of Viśvanāth that Shāhū was able to secure the support of Kānhoji Angria, a partisan of Tārā Bāī and leader of the Marāṭhā navy. While Angria dealt with the Siddis and later with the Portuguese and the English, the Marāṭhā armies, taking advantage of the weakness of the Mughal Empire, made inroads into the Mughal-held territories of Mālwa, Gujarāt, Bāglāna, and Khāndesh. A number of battles were fought and both sides claimed victory. The Marāṭhās, however, held onto some of their gains. Distracted by the Marāṭhā attacks on one side and beset by court intrigues on the other, Ḥusayn ʿAlī, the Mughal governor of Deccan, sought negotiations with Shāhū.

Intervention in Mughal affairs

Guided by Shankraji Malhar, a shrewd Marāṭhā diplomat, Ḥusayn ʿAlī established contacts with Shāhū and signed an agreement with him, under which: (1) Shāhū was recognized as the lawful ruler of territories and forts that had at one time belonged to Śivajī; (2) the Marāṭhās were confirmed in their legal possession of territories recently conquered by them in Khāndesh, Berār, Gondwana, Karnātaka, and other places; and (3) they were allowed to collect certain taxes from all the six Mughal *ṣūba*s (districts) in the south, in return for which they were to serve the Mughal emperor with a contingent of 16,000 troops and pay him an annual tribute. Then, as has been described above, a Marāṭhā army was quickly collected under Khande Rāo Dābhāde and marched with Ḥusayn ʿAlī to Delhi. With the balance of power in their favour, the Sayyid brothers replaced Farrukh-siyar with their protégé, Muḥammad Shāh (1719).

Bālājī returned in triumph to Sātāra, where he died shortly thereafter. He was a capable administrator, as is testified by the new methods of revenue collection, based on *sardeshmukhī* and *chauth,* which he set up for the Marāṭhā state. The *sardeshmukhī* gave its holder (*jāgīrdār*) a right to 10 percent of the taxes collected in his region. The *chauth* was collected from semiconquered territories, where the Marāṭhās kept the peace but took no responsibility for internal welfare and administration. Marāṭhā leaders could make inroads into distant territories and, if they succeeded in establishing control as *jāgīrdār,* could treat those territories as their *vaṭan*s (domains), to which

they could make hereditary claims. Each owed allegiance to Shāhū, who, in turn, owed allegiance to the Mughal emperor. The question of allegiance, however, was academic. The Marāṭhā Empire was highly decentralized, and each *jāgīrdār* claimed virtual ownership of the territory he controlled.

Bājī Rāo and the expansion of Marāṭhā power. Bālājī Viśvanāth was succeeded as peshwa by his son, Bājī Rāo, who, both as a commander and as a diplomat, far outstripped his father. His immediate concern was with the Niẓām al-Mulk, the Mughal viceroy in the Deccan, who was in revolt against Muḥammad Shāh. In 1720 the Niẓām had wrested Asīrgarh and Burhānpur from the Mughals and defeated powerful Mughal armies in the battles of Khandwa and Bālāpur. Evading a direct confrontation with the Niẓām, Bājī Rāo decided to turn his attention to Karnātaka in the far south and Mālwa and Gujarāt to the north of the Niẓām's territories. Bājī Rāo personally led two successful expeditions in Karnātaka in 1725 and 1726, and reestablished Marāṭhā rule there. Taking advantage of the absence of the Marāṭhā forces, however, the Niẓām had moved his armies in the direction of Junnar and Poona. Bājī Rāo adopted diversionary tactics and, quickly moving from Karnātaka, threatened the base of the Niẓām's power in Burhānpur. The Niẓām's forces, quickly retracing their steps, were caught unawares at Pālkhed, 20 miles west of Aurangābād, where they found it impossible to extricate themselves. On March 6, 1728, the Niẓām was forced to enter into an agreement with the Marāṭhās, under which he ratified the Marāṭhā claims accepted earlier by the Sayyid brothers, agreed to pay up the arrears of *chauth* and *sardeshmukhī,* and recognized Shāhū as the legitimate ruler over the Marāṭhā dominions.

Expeditions to Karnātaka

Following his victory against the Niẓām, Bājī Rāo turned his attention to Mālwa, Bundelkhand, and Gujarāt. His brother, Chimnaji, defeated a Mughal army led personally by the governor of Mālwa at Amjhera near Dhār on November 29, 1728. He then moved into Gujarāt. Bājī Rāo, had, in the meantime, moved on to the support of Chhatrasāl Bundela, then fighting against Muḥammād Khān Bangash, governor of Allahābād, and inflicted severe defeats on Qayām Khān, son of Muḥammad Khān Bangash, and then on Bangash himself.

The rapid successes won by the Marāṭhās in the heart of the Mughal Empire presented a serious problem to the rulers at Delhi. The Emperor now appointed Muḥammad Khān Bangash as the governor of Mālwa. He defeated a Marāṭhā force near Ujjain but was soon defeated by another Marāṭhā army, the Marāṭhās continuing their advance toward Delhi.

Determined to show the Mughal emperor "flames and the Marāṭhās at the gates of the capital," Bājī Rāo reached Delhi in March 1737, throwing the capital and its environs into a state of panic and confusion, but, he retreated when a Mughal army moved out of Delhi to meet him. In sheer desperation, the Mughal emperor once again turned toward the Niẓām for help. The Niẓām attempted a pincer movement against the Marāṭhās, trying to crush them with a large force under his son, Nāṣīr Jang. This was perhaps the largest army put forth against the Marāṭhās since the days of Aurangzeb. The Niẓām, however, proved no match for the superior military craftsmanship of Bājī Rāo and was defeated by him near Bhopāl. Mālwa was now formally ceded to the Marāṭhās under the imperial seal. The Niẓām agreed to withdraw completely from the territory between the Jumna (Yamuna) and Narmada, and a substantial sum of cash was to be paid by him to the Marāṭhās out of the imperial treasury.

The Niẓām's defeat

After Nāder Shāh's sack of Delhi in 1739 (see above), Shāhū decided that the most realistic way of extending Marāṭhā power was to adopt a policy of resuscitating the falling Mughal Empire and controlling its administration, rather than attempting to take up the throne. "If we attempt the other course (of aggression)," he told his advisers,

> it would involve us in enmity with all our neighbours, with the consequence that we should be exposed to unnecessary dangers and court a crowd of troubles all round.

Meanwhile the Marāṭhās had become engaged in a struggle with the Portuguese on the western coast. The policy of raising a strong line of naval defenses on the west coast in order to protect the base of the Marāṭhā Empire had been laid down initially in the 1670s by Śivājī. In the meantime, the Portuguese were also busy setting up their fortresses in the region. The effort was resumed by the Marāṭhās as early as 1719 but was taken up in earnest only in the summer of 1737. In April 1739 the Marāṭhās attacked Portuguese-held Bassein. Bassein surrendered on May 5, 1739, and a peace was finally concluded with the Portuguese. Then, when Marāṭhā power was at its zenith, Bājī Rāo died on April 28, 1740, at the age of 40.

Assessment of Bājī Rāo

Bājī Rāo's death closed one important era in Marāṭhā history. Next only to Śivājī in military genius, he had transformed the Marāṭhā state into the major power in India. Under his leadership, the Marāṭhās, once "a cloud no bigger than man's hand," spread all over India. Bājī Rāo is often criticized for his policy of carrying the Marāṭhā power to Delhi at the cost of its consolidation in the Deccan itself, where it was faced by powerful challenges—from the Niẓām al-Mulk in the east and the Portuguese on the west. "Strike, strike at the root, and the branches will fall of themselves," he is supposed to have said. Shāhū, who would have preferred to respect Mughal authority and wished to concentrate on strengthening Marāṭhā power in the south, was unwilling or unable to control his peshwa. Bājī Rāo's eye was certainly fixed on the revenues of Mālwa and Gujarāt, with the help of which he hoped to bring about a consolidation of the Marāṭhā kingdom in the south. But, while struggling for the establishment of the Marāṭhā domination over Mālwa and Gujarāt, Bājī Rāo constantly kept before him the broader political objective of establishing Marāṭhā hegemony over the Deccan. This is borne out by the texts of several agreements he made with the Niẓām, in each one of which Marāṭhā claims for *chauth* and *sardeshmukhī* over the Deccan are reiterated, and by several of his transactions with the Mughal emperor, in which he demanded the right of consultation in the appointment of a governor for Deccan.

THE MARĀṬHĀ CONFEDERACY

Bājī Rāo I was succeeded as peshwa by his son Bālājī Rāo, then only 19. His period as peshwa, 1740–61, was marked by the inability of the Marāṭhā government to control its own chieftains.

The Marāṭhās, engrossed as they were in distant campaigns, did not seem to realize that Nāder Shāh's invasion of 1739 had opened a window in the northwest of India, through which his successors were keeping a constant watch on the affairs at Delhi. Aḥmad Shāh ʿAbdālī, who succeeded Nāder Shāh in 1747, maintained close relations with the Ruhēlas and the Bangashes, the chief enemies of the Delhi Empire. Ṣafdar Jang, the *wazir* (chief minister), with the approval of the Emperor, sought Marāṭhā help, while the Pashtuns turned to ʿAbdālī. The Marāṭhās under Malhār Rāo Holkar and Jayāppā Sindhia, assisted by the Jāṭs, inflicted severe defeats on the Pashtuns and received from the Emperor half of Rohilkhand as their prize—their first territorial conquest beyond the Jumna. The Pashtuns, in sheer desperation, turned toward ʿAbdālī for help, thus involving the Marāṭhās in a clash between the two northern empires.

Conflict with Aḥmad Shāh ʿAbdālī

The Marāṭhās might still have kept out of it if they had not, in their search for glory, decided to follow ʿAbdālī after he was returning from the sack of Delhi in 1757. Defeating Najīb Khān at Āgra and securing the whole of the Doab, the Marāṭhā armies marched in the direction of Lahore. In March 1758 they defeated the Governor of Sirhind, and by April 9 they had expelled ʿAbdālī's agents from Lahore. Tukōjī Holkar and Sabaji Sindhia subjugated the whole of Punjab and planted the Marāṭhā banner on the fort of Attock. They claimed that they had brought the whole of India, from Attock to Cuttack and from Kashmir to Kanniyākumāri (Cape Comorin) under their control, little realizing how uncertain this control was. In 1759, ʿAbdālī, shocked by the Marāṭhā aggressiveness, returned to India. By October he had taken possession of Lahore and was moving toward Delhi.

Aroused to a supreme effort, the peshwa then dispatched Sadāshiv Rāo with an army against ʿAbdālī. The subsequent campaign and Sadāshiv Rāo's defeat at Pānīpat in January 1761 has been described above. Bālājī Rāo died a few months later.

Challenge and survival. Marāṭhā power in the north soon revived under Mādhav Rāo I (1761–72). By April 1770, two able commanders, Ramchandra Ganesh and Visājī Krishna had not only captured Āgra and Mathura but had entered the territory of Aḥmad Khān Bangash and, crushing the joint opposition of Bangash and the Ruhēlas, were able to reestablish the Marāṭhā position in Rohilkhand. On February 10, 1771, Delhi was taken and Jawan Bakht, the son of Shāh ʿAlām II, was placed on the Mughal throne as a Marāṭhā client.

The Marāṭhās now had a new rival in the British, however, who had made excellent use of the Marāṭhā absence from northern India to establish their own power there. At Baksar (Buxar) in October 1764 they defeated the combined armies of the Emperor, the Wazir, and the Ṣūbedār. In 1765, they acquired supreme power from Calcutta to the Doab.

The First Marāṭhā War

By 1772 the British had fully exploited the weaknesses of the local governments in Bengal and Karnātaka by consolidating their power in these regions, and were ready to move against the Marāṭhās, who were distracted by internal disputes following Mādhav Rāo's death. Capturing Broach in 1772 and Salsette a little later, the British obtained a number of commercial, military, and political advantages on the Bombay coast. They also secured a strong position for themselves in Gujarāt in 1775, but then withdrew under instructions from British East India Company headquarters at Calcutta.

The war was soon resumed, this time by the Calcutta government itself. Using the visit of a French agent at Poona to raise the bogey of a Marāṭhā alliance with European powers, Warren Hastings, the governor general, made new demands upon the Poona government and, on the latter's failure to implement them, started an offensive against them simultaneously on the diplomatic and military fronts. Madhya Pradesh, Gujarāt, and Konkan were selected for large-scale military operations, but each was tied up with a diplomatic objective. Col. Thomas Goddard, the commander of the British armies marching through Bundelkhand, made an effort to detach Kandoji Bhonsle, one of the Marāṭhā leaders, from his allegiance to the Poona government. The effort failed, which put Goddard's expedition into a difficult situation. Another British army ascended the Western Ghāts from the side of Bombay and clashed with the Marāṭhās at Wadgaon. Here also they were more interested in pushing diplomatic thrusts than in pressing military victories. Their plan of campaign in Konkan and Gujarāt was directed at bringing pressure upon Mahādājī Sindhia.

The anti-British coalition

The British aggression against the Marāthās was met by a spirit of resistance all over the country; and by 1780 Nānā Fadnavis, representing the central government, was able to organize an alliance against the British, consisting of the Marāthā confederacy, the Niẓām, and Hyder (Haidar) Ali. The Niẓām, by constantly writing to Nānā at Poona and Diwakar Pandit at Nāgpur, had played a leading role in setting it up, whereas Nānā Fadnavis was responsible for bringing in Mahādājī Sindhia, Tukōjī Holkar, and other Marāṭhā leaders. Even those who stood outside the alliance, such as the Nawab Wazīr of Oudh and the Nawab of Bengal, gave it their silent approbation. The Siddis of Janjira, the traditional enemies of the Marāṭhās, joined the combination by a separate agreement. The Portuguese, the French, and the Dutch were sympathetic. A detailed plan of war was drawn up by the allies: Bhonsle was to attack the English in the north, Hyder Ali was to bring them under control on the Madras coast, the Poona forces were to oppose them in Gujarāt and Konkan, and the Niẓām was to attack them on the east coast.

The weakness of the alliance—at both the diplomatic and the military levels—soon became clear. The British could bribe Bhonsle and induce him to permit a free passage to British armies marching under Gen. Eyre Coote through Orissa in the direction of Madras. The Niẓām not only did not send his armies in the direction of Orissa, as he had agreed to do, but kept himself busy in exploiting the opportunity opened to him by the great military pressure that the Marāṭhā armies were exerting in Gujarāt. Besides the political weaknesses of the alliance, it suffered from severe military limitations. Its members, with the solitary exception of Hyder Ali, were incapable of taking any military offensive against the English.

It was Hyder Ali, with his Europeanized armies, who was able to inflict some serious reverses on the English in Karnātaka. Reaching Madras, Eyre Coote found the British there ill prepared either to repel or resist Hyder Ali's armies. Coote wanted the Bombay government to send some military help to Madras, which it was impossible for them to do, since the military pressure on the Bombay side also was growing. This induced the British to greater diplomatic efforts. On September 10, 1780, a second treaty was signed between the British and Fateh Singh of Gujarāt, under which Fateh Singh agreed to give them military support. Bhonsle then agreed (April 6, 1781) to support the British in their war against Hyder Ali. The growing strength of Hyder Ali's offensive on the Coromandel and the Malābar coasts compelled the British first to reduce their pressure on Mālwa and the Western Ghāts and, later, to withdraw themselves from these regions. It was, however, the difficult military situation on the west coast that forced them to open negotiations with the Marāṭhās. On May 17, 1782, a treaty was signed between the British and the Marāṭhās at Salbai. With the solitary exception of Salsette and its adjacent islands, the British were forced to restore every inch of territory they had acquired during the course of a long and nerve-racking war. The weaknesses of the Marāṭhā state, however, had been exposed, and the British could hope to break up the entire political system in stages.

The Treaty of Salbai

During the period of the first Marāṭhā War, the imperial forces at Delhi were being looked after by Mīrzā Najaf Khān. On Najaf Khān's death in April 1782, the Emperor decided to invite the Marāṭhā leader Mahādāji Sindhia to Delhi. Over British opposition, Mahādāji was appointed the *vakīl-e muṭlaq,* the highest executive office in the empire. Mahādāji's task in Delhi was not easy. He had to deal with recalcitrant Mughal chiefs and rebellious Rājput leaders. He sustained a major military defeat at Lalsot in 1787 at the hands of the Rājputs, who had been joined by Hamdani and his Mughal troops. In this difficult situation, Mahādāji frantically sought help from Poona, but received little. Tukōjī Holkar and ʿAlī Bahādur, who were sent with Marāṭhā armies to the north, proved obstacles to his plans rather than a help. Nevertheless, while the British were repairing their losses and consolidating their gains in the war with the Marāṭhās, Mahādāji was able to strengthen his position at Delhi.

Visiting Poona in 1792, he found the atmosphere full of mistrust and suspicion. He succeeded in gaining the confidence and respect of the young Peshwa, but his sudden death on February 12, 1794, at the age of 67, threw all his schemes out of shape. A victory that the Marāṭhā armies won against the Niẓām at the Battle of Kharda in March 1795 could be described only as "the vanishing flicker of the Marāṭhā glory."

The Battle of Kharda

Decline of Marāṭhā power. Following the death of Mahādāji Sindhia, the decline of the Marāthā Empire became rapid. Six months after the victory over the Niẓām at Kharda, Peshwa Savai Mādhav Rāo II died under tragic circumstances, by accident or suicide. The succession devolved upon Bājī Rāo II, a thoroughly incompetent and unscrupulous man but the only legitimate claimant in the peshwa family. Nānā Fadnavis opposed him with all his strength but was finally made to yield under the pressure of Daulat Rāo Sindhia, who was now the most powerful chieftain. The state of uncertainty at Poona and differences among its leaders gave the British governor general, Lord Wellesley, a chance to intervene. He brought increasing pressure on Bājī Rāo to accept an alliance with the British. Nānā was inclined to conciliation, but Bājī Rāo attempted to organize an anti-British confederacy. After Nānā Fadnavis died in 1800, all restraints upon Bājī Rāo were removed.

He instituted a reign of terror in Poona; people suspected of the slightest treachery were ruthlessly selected and brutally killed.

In 1795, a war of succession had started in the Holkar state. Daulat Rāo supported Kāśī Rāo. Encouraged by Nānā, Kāśī's three brothers challenged his claims to the throne. Malhār Rāo being killed in a fracas, Jaswant Rāo collected a roving band of adventurers around him and pursued his family's wars against Sindhia and tried to force the Peshwa to side with him.

Bājī Rāo, fleeing from Jaswant Rāo in 1802, went to Wadgaon, then to Sinhagadh, and ultimately into the arms of the British. Boarding a British ship at Suvarnadurg, he reached Bassein on December 16. On December 31, 1802, he signed a treaty with the British at Bassein, which was to prove the death warrant of the Marāṭhā Empire. Under the Treaty of Bassein, (1) the friends and enemies of one were to be treated as friends and enemies of the other, (2) the British were to protect Bājī Rāo's territory as their own, and (3) a subsidiary force of 6,000 regular infantry, with field artillery, was to be permanently maintained by the British at Poona.

The British were determined to place Bājī Rāo back in power, under their own tutelage, at Poona. On August 7, 1803, Sir Arthur Wellesley opened war upon Sindhia and Bhonsle by an attack on Ahmadnagar. This was followed by the Battle of Assaye, which proved to be a stiff one, the victory costing a great deal to the British. It was followed by the capture of Burhanpur, Asirgadh, and Gāwīlgarh. Earlier, on September 14, the British, under General Burkin, had already defeated another army of Sindhia and, entering Delhi, had taken the Emperor under their protection. The fighting against Sindhia and Bhonsle soon spread to Gujarāt, where the British captured Baroda, Broach and Champaner, and to Orissa, where they captured Jagannāth and Cuttack. Bhonsle and Sindhia were forced to sign separate treaties with the British—Bhonsle on December 17, 1803, under which he ceded to the British the province of Cuttack with all its forts, and to the Niẓām, western Berār up to the Wardha river; and Sindhia on December 30, 1803, by which he ceded to the British the Doab region, parts of Bundelkhand, Broach, and some districts of Gujarāt, the fort of Ahmadnagar, and the Ajantā region up to the Godāvari. He was also asked to renounce his control over the Emperor, relinquish all his claims over the Peshwa, the Niẓām, and the Gaekwar, and to recognize the independence of all those feudatories who had made separate engagements with the British.

The Second Marāṭhā War

The British preoccupation with Bhonsle and Sindhia gave Jaswant Rāo Holkar an opportunity to intensify his military raids. Forcing Bhonsle and Sindhia to sign peace treaties, the British turned toward Holkar and warned him of the consequences of his actions. Holkar, it seems, was living in a world of illusions. He did not seem to understand the change that had taken place in Indian politics. With the Peshwa under their control, with the Niẓām, Sindhia, Bhonsle, and Gaekwar reduced to the position of subordinate allies, and with complete control over Bengal, Oudh, and Karnātaka, the British had become the paramount power in India, whereas all the Indian powers were divided against each other. In 1804, it was sheer foolhardiness for any Indian chieftain to think of resisting them. With the help of his vast, disorganized, and disparate armies, Holkar could elude the British for some time. Plundering Pushkar and Jaipur, he quickly moved on to Mandsaur. He defeated a British force at Banās, sending it into retreat to Āgra. Emboldened by his successes, he captured Mathura and fell upon Delhi. Encountering stiff resistance there, he moved into the Doab, ruthlessly devastating the British territories. Defeated at Farrukhābād, November 17, 1804, he retreated to Dīg, where he was supported by the Jāṭ Raja of Bharatpur. On December 13, Dīg was captured, but the mud walls of Bharatpur successfully resisted several stormy attacks of the British armies and repulsed them with heavy losses. In the meantime, Sindhia, regaining his confidence, had repudiated the treaties he had signed with the British less than a year earlier and marched against the latter with a large army and 180 pieces of heavy artillery. His army soon joined the 60,000 soldiers under Holkar, and a number of leading Marāṭhā *sirdār*s collected at Sabalgarh to think out ways and means of driving the British out of the country. Wellesley's rough-and-ready methods had brought about a severe reaction, and the decision makers in London were forced to think of alternative policies.

Realizing that their aggressive policies had created a difficult situation for them in India, the British government quickly changed tactics and slowed down their program of expansion. They now sent Lord Cornwallis once again to India and asked him to restore Indian politics to the point at which he had left them. Cornwallis started with pacifying Sindhia and signing a treaty with him under which the British gave up their recent alliances with the Rājput princes of Udaipur, Jodhpur, Kota, and others in Mālwa and Mewāt and undertook to enter into no engagement with Sindhia's tributaries, nor to interfere with Sindhia's conquests from Holkar between the Tāpti and Chambal. Holkar kept up the fight against the British. Deserted by Sindhia, he proceeded to Ajmer and requested the Raja of Jodhpur to join him. Frustrated there, he moved toward Patiāla, hoping to receive the support of the Sikhs and Afghans. Both Ranjit Singh and the Shah of Afghanistan gave him some encouragement, but, when Holkar found Ranjit Singh secretly entering into a pact with the British, he was left with no alternative but to sign a treaty with them. Completely in control of the situation in India, the British could afford to be liberal and, in return for his surrendering all claims to his territories north and west of Chambal, they guaranteed to him the territories to the south and east of it. All his ambitions of defeating the British having been frustrated, Holkar settled down at Indore, brooding deeply over the national disaster. For some years his sole hobby was to work day and night in a gun factory he had set up at Bhānpura. In October 1808, he went completely mad, and three years later he died, at the age of 30, after a short career of nine years, packed with heroic deeds and terrible disasters. He was succeeded by Malhār Rāo Holkar, a boy of 10 years, and his wife Tulsi Bāī carried on the administration.

Appointment of Cornwallis

End of the empire. Lord Wellesley was able to establish the British paramountcy in India, but it was at a terrible cost. He had captured power all over India but had refused to take up responsibility. Bājī Rāo, Daulat Rāo, the Niẓām, and the Nawab Wazīr of Oudh were all like gilded puppets, dancing to the tune of the British but incapable and unwilling to exercise any control over administration. These puppet rulers were provided with British subsidiary forces for their protection, but their own armies, now disbanded and converted into roving Pindari bands, wandered aimlessly through the length and breadth of the country, plundering and destroying whatever came their way. The whole country was thrown into chaos. Within five miles of Poona neither life nor property was safe. The wave of anarchy gradually began to spread to the areas under the direct control of the British and forced them to find a solution.

It was specifically with the object of bringing peace to a country in the worst throes of lawlessness and disorder that Lord Hastings was appointed governor general of India in 1813. On the death of Raghuji Bhonsle in March 1816, he forced his successor, Mudhojī Appa Sāhib, to sign a subsidiary treaty. In June 1817 Bājī Rāo was asked to sign another treaty with the British by which he surrendered to them whatever phantom of overlordship over Indian chiefs he still possessed. He was forced to give up all his territory outside Mahārāshtra, withdraw all his *vakīl*s (envoys) from foreign courts, and agree not to maintain any correspondence or communication with them. This amounted to a public declaration of the dissolution of the Marāṭhā confederacy. Bājī Rāo agreed outwardly but, smouldering with anger and frustration within, started inducing other Marāṭhā chiefs to rise against the British. In September 1817, while Bājī Rāo attacked the Poona residency, Appa Sāhib attacked the residency at Nāgpur. Both were foolhardy attempts and ended in failure. Bājī Rāo escaped from Poona and moved toward Nagpur,

The Third Marāṭhā War

with a view to joining Appa Sāhib, but, learning that he was being carried as a prisoner to Allahābād, entered the forests of Gondwana and remained in hiding there for two years, supported by the disbanded armies of Poona and Nāgpur along with a number of Pindaris.

Hastings had, in the meantime, entered into separate treaties with Kota, Būndi, Bharatpur, Udaipur, Jodhpur, Jaipur, and other protégés of Sindhia. Isolated from his supporters and surrounded by the British armies, Sindhia was forced to sign a treaty on November 5, 1817. Thereafter, the British armies moved through the length and breadth of central India, suppressing the Pindaris with a heavy hand. Forcing Amīr Khān, one of the Pindari leaders in collusion with Holkar, to surrender, and conferring upon him the rulership of Tonk, the British attacked Holkar's armies. Malhār Rāo, a boy of 16, and his sister, Bhimā Bāī, a widow of 20, rode through the ranks imploring their men not to surrender. But it was an unequal fight, and Holkar was forced to sign the Treaty of Mandsaur on January 6, 1818, under which he ceded to the British territory north of Būndi and south of Sātpura and agreed to maintain a subsidiary force and disband his armies. Bājī Rāo maintained a solitary resistance for some more time, frantically appealing to all and sundry, including the ruler of Burma, for help against the British. Pursued by British troops, he moved from pillar to post between November 1817 and May 1818 and finally surrendered, shamelessly begging for liberal terms. Offered a safe asylum and a liberal pension, he decided, finally, to settle down at Bithur near Kānpur, symbolizing in his misery the end of the Marāṭhā Empire. (S.P.V.)

India and European expansion, c. 1500–1858

EUROPEAN ACTIVITY IN INDIA, 1498–C. 1760

Earlier links between Europe and India

When the Portuguese navigator Vasco da Gama landed at Calicut in 1498, he was restoring a link between Europe and the East that had existed many centuries previously. The first known connection between the two regions had been Alexander the Great's invasion of the Punjab, 327–325 BC. In the 2nd century BC Greek adventurers from Bactria had founded kingdoms in the Punjab and the bordering Afghan hills; these survived into the late 1st century. This territorial contact in the north was succeeded by a lengthy commercial intercourse in the south, which continued until the decline of the Roman Empire in the 4th century AD. Trade with the East then passed into Arab hands, and it was mainly concerned with the Middle Eastern Islāmic and Greek worlds until the end of the Middle Ages. The only physical contact came from occasional travellers, such as the Italians Marco Polo and Niccolò dei Conti, and the Russian Afanasy Nikitin in the 15th century; and these were few because of commotions within the tolerant Arab-Islāmic world created by successive incursions of Turks and Mongols. For Europe in 1498, therefore, India was a land of spices and of marvels handed down from imaginative Greek authors. For Muslims, Europe was the land of Rūm, or the Greek empire of Constantinople (Turkish after 1453); and for Hindus, it was the abode of the foreigners called Yavanas, a corruption of the Greek word Ionian.

Renewed contact with Europe

The Portuguese. The Portuguese were the first agents of this renewed contact because they were among the few European nations to possess both the navigational know-how and the necessary motivation for the long sea voyage. During the 15th century the land routes for the Indian trade—via the Red Sea and Egypt or across Persia, Iraq, Syria, and Turkey—had become increasingly blocked, mainly by Ottoman action. The surviving Egyptian route was subject to increasing exploitation by a line of middlemen, ending with the Venetian monopoly of the European trade, and in 1517 it likewise passed under Ottoman suzerainty. The motive for finding a new route was therefore strong; this task fell to the Portuguese, partly because the stronger Spaniards were absorbed in discovering the New World (a by-product of the same search for an Eastern route). The Portuguese further inherited crusading zeal from wars against the Moors in Portugal and North Africa. Finally, they had learned navigational techniques from the Genoese, who were disgruntled at their exclusion from the Mediterranean carrying trade.

When Vasco da Gama arrived in Calicut, he hoped to find Christians cut off by Muslim action, to deal a blow at Muslim power from their maritime rear, as it were, and to corner the European spice trade. He found his Christians in the Syrian Christians of Cochin and Travancore, he found the spices, and he found Muslim Arab merchants entrenched at Calicut. It was his successors, Francisco de Almeida and Afonso de Albuquerque, who established the Portuguese Empire in the East. Almeida set up a number of fortified posts; but it was Albuquerque (governor 1509–15) who gave the empire its characteristic form. He took Goa in western India in 1510, Malacca in the East Indies in 1511, and Hormuz (Ormuz) in the Persian Gulf in 1515, and he set up posts in the East Indian Spice Islands. The object of these moves was to establish for Portugal a strategic command of the Indian Ocean, so as to control the maritime spice trade and to ruin the Ottoman-controlled Middle Eastern Muslim world by cutting off its trade. Hormuz dominated the Persian Gulf; an attack on Aden was intended to do the same for the Red Sea. While Malacca was the nerve centre for the spice-producing islands of Indonesia and the exchange mart for the trade with the Far East, Goa, not Malacca, was the capital because of Portuguese concern with the Ottomans of the Middle East.

The Portuguese reliance on sea power

The Portuguese method was to rely on sea power based on fortified posts and backed by settlements. Portuguese ships, sturdy enough to survive Atlantic gales and mounted with cannon, could easily dispose of Arab and Malay shipping. The bases enabled the Portuguese to dominate the main sea-lanes; but Portugal, with less than 1,000,000 people and involved in Africa and South America as well, was desperately short of manpower. Albuquerque turned his fortresses into settlements to provide a resident population for defense. Intermarriage was encouraged. At the same time, Christianity was encouraged through the church. Goa became an archbishopric. St. Francis Xavier started from Goa on his mission to the south Indian fishermen. The Inquisition was established in 1560. The new mixed population thus became firmly Catholic and provided a stubborn resistance to attacks.

A lack of resources precluded any attempt to establish a land empire. Portugal's control of the Indian Ocean—its period of empire—lasted through the 16th century. During this time it attained great prosperity. Goa acquired the title of Golden, and it became one of the world's wonder cities. Trade with Europe was a royal monopoly, and, in addition, a system of licenses for all inter-Asian trade enriched the royal exchequer. Inter-Asian trade was free to individual Portuguese; and it was the profits of this, combined with trimmings from the royal monopoly, that gave them their affluence.

The three marks of the Portuguese Empire continued to be trade, anti-Islāmism, and religion. The Portuguese early considered that no faith need be kept with an infidel, and to this policy of perfidy they added a tendency to cruelty beyond the normal limits of a very rough age; the result was to deprive them of Indian sympathy. In religion the Portuguese were distinguished by missionary fervour and intolerance. Examples of the former are the Madura mission of Roberto de Nobili (1577–1656), nicknamed the White Brahmin, and the Jesuit missions to the court of the Mughal emperor Akbar (ruled 1556–1605). Of the latter, there was the Inquisition at Goa and the forcible subjection of the Syrian church to Rome at the Synod of Diamper in 1599.

The Portuguese thus had few friends in the East to help them in a crisis, and in 1580 the Portuguese kingdom was annexed to Spain; thenceforth until 1640 Portuguese interests were sacrificed to those of Spain. Because of the Spanish failure to quell a Dutch rising in The Netherlands, and after the defeat of the Spanish Armada in 1588, the route to the East was opened to both English and Dutch. The Dutch arrived first; under their blows in Indonesia and those of the English in India the Portuguese ascendancy crumbled, though they retained Goa until 1961.

This first modern impact on India has left distinct though

not extensive traces. The first is the mixed population of Luso-Indians, or Goanese, along the western coast of India and in Ceylon and with them, a lingua franca in the ports and markets. Then came Roman Catholicism, which today has some 12,000,000 followers and an array of churches, convents, and colleges all over India. More tangible traces include imported articles such as tobacco, potatoes, pineapples, tomatoes, papayas, cashew nuts, and two varieties of chilies.

The Dutch. In the race to the East after the Spanish obstacle had been removed, the Dutch, having larger resources, were the first to arrive. Their first voyage was in 1595, helped by the local knowledge of Jan Huyghen van Linschoten, who had worked for six years in Goa. J. Van Neck's voyage in 1598 was so profitable (400 percent for all eight ships) that the die was cast for a great Eastern adventure. The Dutch objective was neither religion nor empire, but trade; and the trade in mind was the spice trade. They were monopolists rather than imperialists. Empire came later, as a safeguard for monopoly.

Dutch objectives in India

The Dutch, therefore, went directly to the East Indies (Indonesia), the main source of spices, and only secondarily to south India, for pepper and cardamoms, and to Ceylon for these and cinnamon. From 1619 their headquarters were fixed at Batavia (Djakarta) in Java, from which they developed a series of outlying stations in the East Indian islands (*e.g.*, Celebes and the Moluccas) and intermediate ones like Cape Town in South Africa, and Ceylon for supply. This was the work of the governor general J.P. Coen (served 1618–23; 1627–29), and the whole system may be said to have been completed under the governor general J. Maetsuyker (served 1653–78). Monopoly was the keynote of 17th century Dutch activity in the East; empire followed in the 18th.

The Dutch system demanded the control of the Eastern seas, and this meant the elimination of European rivals. Those in possession were the Portuguese. The Dutch succeeded with superior resources and better seamanship, but the Portuguese, though defeated, were not destroyed. Ousted from most strongholds, the Portuguese retained their capital, Goa, in spite of blockades and sieges. The second European obstacle was the English, who followed the Dutch to the East Indies; no match for the Dutch in resources, they were virtually excluded from the East Indies with the Dutch seizure of their factory at Amboina (modern Ambon) in 1623 (see below *The British, 1600–1740*).

It remained for the Dutch to organize their trade, which was operated through the Dutch East India Company, a complicated organization dominated by the maritime state of Zealand. Much larger than the English company, it had the character of a national concern. Dutch sea power, more efficient than that of the Portuguese, secured monopoly conditions in the islands and sea-lanes. It was only in land areas like Travancore that resort had to be made to competition. But there remained the problem of exchange, for the Dutch, like the English, were short of exchange goods. Textiles were needed to buy spices in Indonesia; and silver, to buy textiles (cotton or silk) in India and China. To work the spice monopoly the Dutch developed an elaborate system of Eastern trade from the Persian Gulf to Japan, the ultimate object of which was to secure the goods with which to secure the spices without recourse to scarce European resources. It was this trade that brought the Dutch to India at Surat, the Coromandel coast (Negapatam), Bengal, and up-country at Āgra.

The British, 1600–1740. The English venture to India was entrusted to the East India Company, which received its monopoly rights of trade in 1600. The company included a group of London merchants attracted by Eastern prospects, not comparable to the national character of the Dutch company. Its initial capital of £50,000 was less than one-tenth of the Dutch company's. Its object, like that of the Dutch, was to trade in spices; and it was at first modestly organized on a single-voyage basis. These separate voyages, financed by groups of merchants within the company, were replaced in 1612 by terminable joint stocks, which covered operations over a term of years. Not until 1657 was a permanent joint stock established.

Formation of the English East India Company

The company's objective was the spices of the East Indies (Indonesia), and it went to India only for the secondary purpose of securing cottons for sale to the spice growers. The British East Indian venture met with determined Dutch opposition, culminating in the massacre of Amboina in 1623, when the Dutch seized the English factory there and executed its factors. No redress was ever obtained, though the Dutch occupation was not recognized until 1667.

In India the English found the Portuguese enjoying Mughal recognition at the western Indian port of Surat. Portuguese command of the sea nullified the English embassy to the Mughal court in spite of its countenance by the emperor Jahāngīr. The English victory at Swally in 1612 over the Portuguese, whose control of the pilgrim sea route to Mecca was resented by the Mughals, brought a dramatic change. The embassy of Sir Thomas Roe (1615–18) to the Mughal court secured an accord (in the form of a firman, or grant of privileges), by which the English secured the right to trade and to establish factories in return for becoming the virtual naval auxiliaries of the empire. This success, with England's exclusion from Indonesia by the Dutch in the same period, determined that India, not the Far East, should be the chief theatre of English activity in Asia.

There followed through the 17th century a period of peaceful trading through factories operating under Mughal grants. This held good for Surat and later for Hooghly (1641) in Bengal. In the south, the factory at Masulipatam (1611) was moved to the site of Madras, granted by a Hindu raja (1640); it shortly (1647) came under the control of the sultans of Golconda and thence passed to the Mughals in 1687. The only exception to this arrangement was the island port of Bombay; Bombay was independent, but its trade was small because the Marāṭhās, soon locked in combat with the Mughals, held the hinterland.

The trade the company developed differed radically from that of the Dutch. It was a trade in bulk instead of in highly priced luxury goods; the profits were a factor of volume rather than scarcity; it worked in competitive instead of monopolistic conditions; it depended upon political goodwill instead of intimidation. The English trade became more profitable than that of the Dutch because the smaller area covered and the lack of armed forces necessary to enforce monopoly reduced overhead charges. But it encountered its own difficulties. The Indians would take little other than silver in exchange for their goods, and the export of bullion was an offense to England's reigning mercantilist political economy. Lack of military power meant management of Asian governments instead of their coercion. Lack of home dominance meant compromise and hazard of fortune.

The English company's trade

To solve the silver problem the English developed a system of country trade not unlike that of the Dutch, the profits of which helped to pay for the annual investment of goods for England. Madras and Gujarāt supplied cotton goods, and Gujarāt supplied indigo as well; silk, sugar, and saltpetre (for gunpowder) came from Bengal, while there was a spice trade along the Malabar coast from 1615 on a competitive basis with the Dutch and Portuguese. Opium was shipped to the Far East, where it later became the basis of the Anglo-Chinese tea trade. The merchants lived in "factories" or in a collegiate type of settlement where life was confined, colourful, and often short.

The company had many difficulties in England. There was mercantilist disapproval and mercantile jealousy of monopoly; there was danger from government at a time of political commotion. King Charles I encouraged the rival Courteen Association (1635), and Oliver Cromwell allowed virtual free trade until 1657. Under the later Stuarts the company prospered, only to have its hopes dashed by a war in India and by the Whig revolution of 1688. The Whigs promoted a new company in 1698, which, however, failed to oust the old one after some years of struggle. In 1702 the government insisted on a merger, which was completed in 1708–09 under the name of The United Company of Merchants trading with the East Indies. This was the body that 40 years later launched on the sea of Indian politics.

Difficulties for the company in England

A way of harassing the company besides attacks on the export of bullion was to limit the sale of cotton goods in England. In 1700 the sale of Asian silks and printed or dyed cottons was forbidden, but trade continued for re-export to the Continent. After 1700 the company found a new profitable line in the China tea trade, whose imports increased (1706–50) from 54,000 to 2,300,000-odd pounds.

In India, the company suffered a serious setback when it resolved, under Sir Josiah Child's inspiration, to resort to armed trade and attack the Mughals. The emperor Aurangzeb was too strong, however, and the venture (1686–90) ended in disaster. Out of this fiasco, however, came the foundation of Calcutta by Job Charnock in 1690—a mud flat that had the advantage of a deep anchorage—and the age of fortified factories surrounded by satellite towns. These were the answers, with Mughal consent, to increasing Indian insecurity. The Madras factory was already fortified, and Ft. William in Calcutta followed in 1696. The company thus had, with independent Bombay, three centres of Indian power.

For the next half century the company confined its relations with the Mughals, who had now spread to the deep south beyond Madras, to disputes over rights and terms of trade at local levels. Fresh privileges were obtained in Delhi, and these they were content to argue about rather than fight for. The factors were learning the art of Indian diplomacy as they had formerly to learn the arts of Indian commercial management.

The French. The French had shown an interest in the East from the early years of the 16th century, but individual efforts had been checked by the Portuguese. The first viable French company was launched by the minister of finance Jean-Baptiste Colbert, with the support of Louis XIV, in 1664. After some false starts, the French company acquired Pondicherry, 85 miles (137 kilometres) south of Madras, from a local ruler in 1674. It obtained Chandernagore (modern Chandernagar), 16 miles north of Calcutta, from the Mughal governor in 1690–92. At first the French initiatives suffered from the mixing of grandiose political and colonial schemes with those of trade. Under the care of François Martin, from 1774, the company turned to trade and began to prosper.

French acquisitions

The progress of the settlements was interrupted, however, by events in Europe. The Dutch captured Pondicherry in 1693; when the French regained it under the Peace of Ryswick (1697), they had gained the best fortifications in India but had lost their trade. By 1706 the French enterprise seemed moribund. The company's privileges were let to a group of Saint-Malo merchants from 1708–20.

After 1720, however, came a dramatic change. The company was reconstituted, and over the next 20 years its trade was expanded, and new stations were opened. Mauritius was finally settled in 1721; Mahé in Malabar and Kārikāl on the eastern coast were acquired in 1725 and 1739, respectively. Chandernagar was revived. Though a company in form, the French company remained under the close supervision of the government, which nominated the directors and, from 1733, guaranteed fixed dividends.

In spite of the company's growth and its fostering by government, its sales in Europe in 1740 were only about half those of the English East India Company. Its trade was large enough to be worth seizing but not great enough to rival that of the English.

Other enterprises in India included a Danish East India Company, which operated intermittently from 1616 from Tranquebar in south India, acquiring Serāmpore (modern Shrīrāmpur) in Bengal in 1755, and the Ostend Company of Austrian Netherlands merchants from 1723, a serious rival until eliminated by diplomatic means in 1731. Efforts by Swedes and Prussians proved abortive.

The Anglo-French struggle, 1740–63. In 1740 India appeared to be relatively tranquil. In the north the Persian Nāder Shāh's invasion (1739) had proved only to be a large-scale raid. In the Deccan, the viceroy, Niẓām al-Mulk, provided some measure of stability. In western India the Marāṭhās were dominant. However, there was competition between Marāṭhās, Mughals, and local rulers for political supremacy in the Deccan. There was a sense of impending change in the air; the Mughal emperor was sickly, the Niẓām was aged, the Marāṭhās were active and ambitious.

It was on this scene that events in Europe precipitated an Anglo-French struggle in India. The War of the Austrian Succession began with Frederick II the Great of Prussia's seizure of Silesia in 1740; France supported Prussia, and from 1742 England supported Austria. The stage thus set, the English decided that the French Indian trade was too powerful to be left alone; the neutrality of previous years was therefore abandoned. Both sides depended on sea power for success, but it was the French who moved first—with an improvised fleet from Mauritius, Comte Mahé de La Bourdonnais drove the British in alarm to Bengal and captured Madras after a week's siege in September 1746. Quarrels between La Bourdonnais and the Governor of Pondicherry, Joseph-François Dupleix, marred this unexpected success, but an English attack on Pondicherry was repelled. Then the Treaty of Aix-la-Chapelle (1748), which ended the war, returned Madras to the English in exchange for Cape Breton Island in North America.

It would thus appear that the status quo had been restored. In fact the situation had radically changed. Madras was now recognized as British by European treaty, and this was accepted by one of the rival Indian chiefs. The French had grown in prestige as skillful soldiers and in power by detachments of the French fleet left behind on La Bourdonnais' departure. Above all, the astute Dupleix had seen the opportunity offered for exploiting the new French reputation in the confused politics of the region. For some years there had been a disputed succession to the governorship of the Carnatic (Karnātaka), itself a dependency of the Niẓām of Hyderābād. The Niẓām had installed a new Carnatic nawab in 1743, but the dispute smouldered on between the partisans of the two rival families, who looked impartially to Marāṭhās, Mughals, and the Europeans for help.

Dispute over the Carnatic succession

In 1748, on the morrow of Aix-la-Chapelle, an occasion for French interference occurred with the death of the aged Niẓām. There was a disputed succession between his second son and a grandson, Muẓaffar Jang. Dupleix, encouraged by his easy repulse of the Carnatic nawab from the walls of Madras, decided to support both Muzaffar and the claimant to the Carnatic nawabship, Chanda Sāhib. Dupleix's reward for success would be the means of ruining the English trade in south India and an indefinite influence over the affairs of the whole Deccan. At first fortune favoured him. The Carnatic nawab was killed in the Battle of Ambur (1749), which demonstrated convincingly the superiority of European arms and methods of warfare over those of contemporary India. The threatening invasion of the new niẓām, Nāṣir Jang, ended with the Niẓām's murder in December 1750. French troops conducted Muẓaffar Jang toward Hyderābād; and when Muẓaffar in turn was murdered three months later, the French succeeded in placing the late Niẓam's third son, Ṣalābat Jang, on the Hyderābād throne. Thenceforward, in the person of the skillful Marquis de Bussy-Castelnau, Dupleix had a kingmaker at the centre of Muslim power in the Deccan.

The English response to these dramatic successes was to support for the Carnatic nawabship the son of the late nawab, Muḥammad ʿAlī, who had taken refuge in the rock fortress of Trichinopoly (modern Tiruchchirāppalli). They had already interfered in the affairs of Tanjore (modern Thanjāvūr) and were no strangers to Indian politics. The French supported Chanda Sāhib for the nawabship. There thus developed what was really a private war between the two companies.

Bussy-Castelnau was established at Hyderābād, with the revenues of the Northern Sarkārs (six coastal districts) to support his army. In the south the French had only Muḥammad ʿAlī to remove. But from 1751 Dupleix's star began to wane. Robert Clive, a discontented young British factor who had left the countinghouse for the field, seized the fort of Arcot, political capital of the Carnatic, with 210 men in August 1751. This daring stroke had the hoped-for effect of diverting half of Chanda Sāhib's army to its recovery. Clive's successful 50-day defense permitted

Downfall of Dupleix

Muḥammad ʿAlī to procure allies from Tanjore and the Marāṭhās. The French were worsted, and they were eventually forced to surrender in June 1752. Dupleix never recovered from this blow; he was superseded in August 1754 by the director Charles Robert Godeheu, who made a not unfavourable settlement with the British.

The French gained but a brief respite; the Seven Years' War in Europe, in which Britain and France were once more on opposite sides, broke out in 1756. Both sides sent armaments to the East. The first British force was diverted to Bengal, so that the French general Thomas Arthur Lally had an advantage on his arrival in 1758. Lally was brave but headstrong and tactless; after taking Fort St. David, he lost time and credit marching to Tanjore, where he forfeited Indian sympathy by executing temple Brahmins. Then his attack on Madras (1758–59) miscarried, while Clive's troops from Bengal defeated the French garrison of the Northern Sarkārs. When Sir Eyre Coote arrived with reinforcements, the British defeated Lally decisively at Wandiwash in January 1760. Bussy-Castelnau, who had been recalled from Hyderābād, was captured; and Lally retreated to Pondicherry, where, after an eight-month siege made tense by bitter recrimination, he surrendered in January 1761. The French threat to British power in India had ended.

Reasons for the British victory

This defeat could be partly blamed on Lally, but there were also other more vital causes. An overriding factor was the British command of the sea. Lally could get no allies for lack of money and no money for lack of supply from France. The British could supply Madras from both Britain and Bengal. The French company was under the control of the French government, and the company suffered from the vicissitudes of its politics.

European military superiority. The supremacy in Indian politics, which seemed to come so suddenly to the Europeans in India, also requires explanation. There was the matter of arms. The Mughals imported their cavalry tactics from Turkestan and their artillery from Turkey. Their firearms remained slowfiring and cumbrous, so that they were outclassed both in quick firing and in range by the 18th-century musket and the cannon landed from the fleets. Infantry could fire three times instead of once before the heavy Mughal cavalry could close, and this destroyed at a blow the Indians' traditional dominance. Moreover, besides this technical advantage, the Europeans also had the advantage of discipline. Troops with loyalty guaranteed by regular pay were more than a match for the personal retinues or mercenary soldiers of the Indian chiefs, however brave they might be individually. A chronic problem with Indian armies at that time was the lack of means to pay them; campaigns would be diverted for collecting revenue for this purpose (when Europeans later trained Indians in the European manner, their advantage increased; discipline removed the uncertain factor of personal leadership, and regular pay removed the Indian general's bugbear of mutiny). A further advantage was civil discipline; the European forces were directed by men themselves under discipline, who were without hereditary connections or ties (though to modern eyes they often seemed refractory or disloyal, by contemporary Indian standards they were regularity itself). Indian loyalty was to leaders who might be killed, to relatives who might back the wrong side, to governments that might (and often did) fail to pay or be overthrown. On the Indian side, whatever the situation, someone was nearly always looking over his shoulder thinking of the chances of a change of leadership or a successful coup. Thus the European possessed not only an expertise denied to the Indians but also a spirit of confidence, a tenacity, and a will to win that was rare in the Indian forces of the time.

Causes of the revolution

Revolution in Bengal. The revolution in Bengal was the product of a number of unrelated causes. The imminence of the Seven Years' War prompted the British to send out Clive with an armament to Madras in 1755. Succession troubles in Bengal combined with British mercantile incompetence to produce a crisis at a moment when the French in south India were still awaiting reinforcements from France.

ʿAlī Vardī Khān—the governor, or nawab, and virtual ruler of Bengal—died in April 1756, leaving his power to his young grandson Sirāj-ud-Dawlah. The latter's position was insecure because of discontent among both his Hindu and Muslim officers and because he himself was both headstrong and vacillating. On an exaggerated report that the British were fortifying Calcutta, he attacked and took the city after a four-day siege, on June 20, 1756. The flight of the British governor and several councillors added ignominy to defeat. The survivors were held for a night in the local lockup, known as the Black Hole; many were dead the next morning.

News of this disaster caused consternation in Madras. An armament preparing to oust Bussy-Castelnau from the Deccan was diverted to Bengal, giving Clive a force of 900 Europeans and 1,500 Indians. He relieved the Calcutta survivors and recovered the city on January 2, 1757. An indecisive engagement led to a treaty with Sirāj on February 9, which restored the company's privileges, gave permission to fortify Calcutta, and declared an alliance.

This was a decisive point in British Indian history. According to plan Clive should have returned to Madras to pursue the campaign against the French; but he did not. He sensed both the hostility and insecurity of Sirāj's position and began to receive overtures to support a military coup. The chance of installing a friendly and dependent nawab seemed too good to be missed. Having taken this decision, Clive chose the right candidate in Mīr Jaʿfar, an elderly general with much influence in the army. In so acting Clive was probably influenced by the example of Bussy-Castelnau at Hyderābād; for six years Bussy-Castelnau had maintained himself with an Indo-French force, sustaining the niẓām Ṣalābat Jang and maintaining French influence in the largest south Indian state with outstanding success. This system, of a "sponsored" Indian state, controlled but not administered, was the one Clive had in mind for Bengal.

Clive's plans for Bengal

The prospects for success seemed good. The event, however, proved otherwise, and there were reasons for this not realized at the time. The chiefs were so lacking in vigour that they made little resistance to British encroachments. External danger could only come from one direction and source—the Mughal authority—and that was at the moment in dissolution. While Bussy-Castelnau had no French merchants to satisfy, the British merchants in Calcutta were ready and eager to exploit the situation. And, because the British company's government was made up entirely of merchants, it is easy to understand why the sponsored state of 1757 became the virtually annexed state of 1765.

Before breaking with Sirāj, Clive took the French settlement of Chandernagar, which the Nawab left to its fate lest he might need British help to repulse an Afghan attack from the north. The actual conflict with Sirāj, at Plassey (June 23, 1757), was decided by Clive's resolute refusal to be overawed by superior numbers, by dissensions within the Nawab's camp, by Mīr Jaʿfar's failure to support his superior, and by Sirāj's own loss of nerve. Plassey was, in fact, more of a cannonade than a battle. It was followed by the flight and execution of Sirāj, by the occupation of Murshidābād, the capital, and by the installation of Mīr Jaʿfar as the new nawab.

Clive now controlled a sponsored state, and he played the part with great skill. His position was prejudiced at the outset by the Nawab's failure to find the expected hoarded treasure with which to fulfill his financial promises to the British. The Nawab therefore looked for financial support toward his Hindu deputies, with whom saving was a second nature. Clive had therefore to intervene repeatedly. In 1759 he defended Patna from attack by the heir to the Mughal throne, ʿAlī Gawhar (later Shāh ʿĀlam II), who hoped to strengthen his position in the confused world of Delhi politics by acquiring Bihār. Clive also had to deal with the Dutch, who, hearing of Mīr Jaʿfar's restiveness and alarmed by the growth of British power in Bengal, sent an armament of six ships to their station at Chinsura on the Hooghly. Though Britain was at peace with Holland at the time, Clive manoeuvred the Dutch into acts of aggression, captured their fleet, defeated them on land, and exacted compensation. They retained Chinsura but could never again challenge the British position in Bengal.

Clive left Calcutta on February 25, 1760, at the height of his fame and aged only 34, looking forward to an English political career. The Nawab was completely dependent on the British, to whose trade it seemed that the rich resources of Bengal were now open. But the prospect was less brilliant than it looked; and for this, and for the troubles that ensued in the next few years, Clive had a direct responsibility. Two measures undermined the plan of a sponsored state, leading to the company's bankruptcy on the one hand and to the virtual annexation of Bengal on the other. The first of these was an understanding with Mīr Jaʿfar, not mentioned in the actual treaty, that the private internal trade of the East India Company's servants should be exempt from the usual tolls and customs duties (the company's trade with Europe was already exempt from duty under concessions made in 1717). The company's monopoly covered the actual trade with Europe and the articles purchased in India to make up its "investment"; but its servants were free to trade otherwise within Bengal and overseas to Asian countries. It was this private trade that, in the virtual absence of salaries, gave most of them their subsistence and some their fortunes. In Bengal this private trade was subject to internal duties. Mīr Jaʿfar's concession, expressed in the form of *dastaks*, or passes, gave them immense advantages over their Indian competitors. From free trade many passed to intimidation; agents were employed who used the British name to terrorize the countryside and infringe the company's monopoly.

Problems in a sponsored state

The second measure was the taking of presents. This was not forbidden by the company and was, in fact, a recognized custom; but it unlocked the floodgates of corruption. The treasure of Murshidābād was reputed to be worth £40,000,000 sterling. On the strength of this rumour, large amounts were paid to the armed forces and to the company leaders. In addition, Clive obtained a further Mughal title and then claimed a revenue assignment, or *jāgīr,* for its upkeep, which was worth £30,000 a year. In the context of contemporary values these grants equalled nearly one-quarter of the average annual Bengal revenue, or about one-seventeenth of the then annual revenue of Great Britain. With such a vigorous opening of the floodgates, it is not surprising that the other servants of the company asked for more almost as a matter of right and that the company's directors in London, with relatives and connections on the spot, preferred verbal denunciations to any resolute or sustained action. The effects became speedily apparent when in fact the £40,000,000 sterling turned out to be £1,500,000, so that (as Clive later admitted to a parliamentary enquiry), the Nawab had to sell jewels, goods, and furniture to meet his obligations. The results of these measures unfolded themselves in the next decade and continued to be felt for a generation.

THE EXTENSION OF BRITISH POWER, 1760–1856

The period of disorder, 1760–72. The departure of Clive signalled the release of acquisitive urges by the company's Bengal servants. These urges were so strong that the governor, Henry Vansittart (served 1760–64), found himself unable to control them. Under the company's constitution, he had only one vote in a council of up to a dozen and could be overruled by any knot of determined men. During these years a body of English merchants, long separated from British standards and social restraints, suddenly found themselves with real but undefined authority over the whole of a large and rich province. It is not surprising that they thought mainly of getting rich quickly.

The first step was the deposition of the nawab Mīr Jaʿfar on the grounds of old age and incompetence. His supplanter, Mīr Qāsim, reasserted his authority. He paid £200,000 in gratuities, of which Vansittart received £50,000 besides an allowance of £18,000 from the company and private trade on his own account. In addition, he ceded the districts of Burdwān, Midnapore, and Chittagong. Both sides wanted power, and both sides were short of money. The Nawab had lost substantial land revenue and the lucrative tolls on the British merchants' private trade; the company was receiving no remittances from Britain, because the directors considered that Bengal should pay for itself. A clash was inevitable.

Mīr Qāsim removed his capital to distant Monghyr, where he could not be so easily overseen, asserted his authority in the districts, and raised a disciplined force under an Armenian officer. He then turned to the company and negotiated a settlement with Vansittart, by which the company's merchants were to pay an ad valorem duty of 9 percent, against an Indian merchant's duty of 40 percent. At this the Calcutta Council revolted, reducing the duty to 2.5 percent and on salt only. The breach came in 1763 when Mīr Qāsim, after defeat in four pitched battles, murdered his Indian bankers and British prisoners and fled to Oudh. The next year Mīr Qāsim returned with the emperor Shāh ʿĀlam II and his minister Shujāʿ-ud-Dawlah to be finally defeated at Buxar (Baksar). This conflict rather than Plassey was the decisive battle that gave Bengal to the British.

These events had been viewed with growing alarm in London. The news of the Mīr Qāsim campaign coincided with the victory of Clive's faction in the company over that of Lawrence Sulivan. Clive used it to appoint himself governor with power to act over the head of the council; he intended an administrative reformation and a political settlement. He arrived in May 1765 to find that the British victory at Buxar had placed Shāh ʿĀlam in his hands but had created a situation of deep confusion in other respects. Mīr Jaʿfar had been restored and had died; his second son had succeeded at the cost of £139,000 in presents. The British merchants and their agents were the unresisted predators of the Bengal economy, and no one knew the next step to take.

Clive's return to India

Clive acted with extraordinary vigour. Within four days of arrival he had set up a Select Committee; and, when he left less than two years later, he had effected another revolution. Turning to India's political situation, Clive had to decide where to stop. No one barred his way to Delhi, and he could at that moment have turned the whole Mughal Empire into a company-sponsored state. But he realized that Delhi was easier to have than to hold. He fixed his frontier at the borders of Bihār and Oudh. Shāh ʿĀlam was given the districts of Kora and Allahābād, and he settled in the latter city, with a tribute (or subsidy) of £260,000 from Bengal (nearly 10 percent of its estimated revenue). Shujāʿ received back Oudh, with a guarantee of defense, in return for paying the troops involved and a cash indemnity. These two were to be buffers between the company and the Marāṭhās and possible marauders from the north.

Clive's next step was to settle Bengal's own status. The Mughal emperor still had much influence, though little power; his complete discountenance might therefore have done the company more harm than good. Clive's solution was to obtain from Shāh ʿĀlam the dewanee, or revenue-collecting power, in Bengal and Bihār (the company was thus the imperial dewan [Hindī *dīwān*] for those two provinces). The Nawab was left in charge of the judiciary and magistracy, but he was helpless because he had no army and could only get money to raise one from the company.

Clive's system of dual government

This was Clive's system of "dual government." The actual administration remained in Indian hands; and for superintendence Clive appointed a deputy dewan, a Persian officer, Muḥammad Riẓā Khān, who was at the same time appointed the Nawab's deputy. The chain was thus complete. The company, acting in the name of the Emperor and using Indian personnel and the traditional apparatus of government, now ruled Bengal. Their agent was Riẓā Khān; the success of the experiment turned on his efficiency and the extent of the governor's support.

Within the company, Clive enforced his authority by accepting some resignations and enforcing others. Presents of more than 4,000 rupees were forbidden, and those between that figure and 1,000 rupees were only to be received with official consent. The regulation of private trade was more difficult, for the company paid virtually no salaries. Clive formed a Society of Trade, which operated the salt monopoly, to provide salaries on a graduated scale; but the company directors disallowed this on the ground of expense, and two years later they replaced it by commissions on the revenue, which cost the company more. Finally, Clive dealt with overgrown military allowances with

equal vigour, overcoming a mutiny headed by a brigade commander. He used a legacy from Mīr Ja'far to start the first pension fund for the Indian army.

Clive left Calcutta in February 1767. His work—diplomatic, political, and administrative—was a beginning rather than a complete settlement. But in each direction, instead of looking back to the past, it reached out to the future. This creative period exacted a heavy price. Clive was pursued to England by his enemies, who launched a parliamentary attack, which, though triumphantly repulsed in 1773, led to his suicide the following year. The year 1765 marks the real beginning of the British Empire in India as a territorial dominion.

Enrichment of the company's servants

It is worth noting how the company's servants so enriched themselves at this time that they undermined the economy of Bengal, and those who returned to England became a byword for ostentation. Apart from the great political prizes already mentioned, it must be remembered that all of the company's servants were engaged in private trade on their own account. Their new authority and the company's power enabled them to exploit their trade with little hindrance. They had the means of using intimidation (through their agents) against Indian rivals like the indigo growers, and Indian police, customs, revenue, and judicial officials. Presents and bribes were the price Indians had to pay for freedom from harassment. They were able through their connection with the administration to arrange virtual monopolies for particular articles in particular districts, fixing a low purchase price as well as a high selling price. They could arrange commissions on revenue collection, mercantile transactions, and any form of commercial activity. What was not done through agents could be arranged through intermediaries, who also, of course, had their own compensation. Thus a man could make a fortune, lose it in England, return for another, lose it again, and return for a third. It is significant that from the time of Clive's second governorship lamentations increased that the opportunities for quick fortunes were slipping away.

The Company Bahadur. The regime that Clive established in 1765 was really a private dominion of the East India Company. It was not a British colony, and it fitted into the very flexible structure of the dying Mughal Empire. The structure of the administration was Mughal, not British, and its operators were Indian, personified by the deputy nawab Muḥammad Riẓā Khān. It was a continuation of the traditional state under British control, and it can be aptly described by the company's popular title, the Company Bahadur, the Valiant, or Honourable, Company. This Company Bahadur state continued through Warren Hastings' time and in essence until the early 19th century, although Lord Cornwallis (governor general, 1786–93) substituted largely British for Indian personnel. The revenue was collected by the officers of the deputy nawab; the law administered was the current Mughal (Islāmic) criminal code, with the traditional personal codes of the Hindu and Muslim communities; the language of administration was Persian. Only the army broke with the past, with its British officers, its discipline, and its Western organization and tactics.

The arrival of Warren Hastings

It was this state that Warren Hastings inherited when he became governor of Bengal in 1772. His 13-year rule can be divided into his internal administration, his dealings with his council, and his foreign policy. Hastings inherited a state that in the five years since Clive's departure had stepped back toward the corruption from which Clive had rescued it. But Hastings was armed with authority by the directors, so that the first two years of his government were a period of real reform. He first dealt with the *dastaks,* or free passes, the use of which had crept in again since Clive's departure; they were abolished, and a uniform tariff of 2.5 percent was enforced on all internal trade. Private trade by the company's servants continued, but within enforceable limits. The Bengali began to experience some security and a settled order, if not yet an equitable society. Next, the company "stood forth as *dīwān,*" taking over the responsibility for the revenue collection from Riẓā Khān, who was arraigned for corruption; the charges could not be proved, however, even with the approving support of the British authorities. Hastings substituted British for Indian collectors working under a Board of Revenue. In a way this was a retrograde step, for the new collectors were often as corrupt as their predecessors and more powerful; but the change gave legal power to those who already wielded it in fact, and in the future their irregularities could more easily be dealt with than could the surreptitious dealings through the old Indian collectors. Finally, Hastings instituted a network of civil and criminal courts in place of the deputy nawab's. The same law was administered by British judges, who were often incompetent, but a model was provided into which Western ideas and practices could later be fed.

These changes held good through the period of Hastings' rule and may be said to have provided a viable though not yet very competent or equitable state. Criminal and personal law cases were virtually in the hands of Indian assessors to British judges who did not know Persian; revenue administration was distorted by the collectors' desire both for gain and increased returns. Hastings was least successful in his revenue administration, for which he never advanced beyond a condition of trial and error; a five-year settlement made in ignorance proved unsuccessful, and he was finally reduced to annual settlements, which meant hit-and-miss arrangements with the traditional *zamīndār*s (land revenue collectors).

Hastings was personally incorrupt, but he had to tolerate a good deal in others and to resort to extensive jobbing to placate his supporters both in Bengal and in London. He left a personal legend behind him, but his administration was disorderly as well as strong. A reason for this can be found in his relations with his council. Under the Regulating Act of 1773 Hastings became governor general of Fort William in Bengal, with powers of superintendence over Madras and Bombay. He was also given a Supreme Court, administering English law to the British and those connected with them, and a council of four, appointed in the Regulating Act. The leading council member, Sir Philip Francis, hoped to succeed him; and because Hastings had no power of veto, Francis was able with two supporters to overrule him. For two years Hastings was outvoted, until the death of one member enabled him to use his casting vote. But the struggle continued until Francis returned to London in 1780, to continue his vendetta there. The struggle culminated with charges against Hastings of malversation by an Indian entrepreneur, Nand Kumar, and with the latter's conviction before the Supreme Court of perjury and his execution under English law. The episode exposed the moral weakness of the council majority, which failed to reprieve Nand Kumar, and convinced the Indians of Hastings' overriding power. This struggle, lasting for years, left Hastings triumphant but also embittered; he had not only to deal with the opposition in Calcutta, which never ceased, but also with the constant threat of supersession in the involved politics of London at that time. This strain probably accounts for the acts that formed important items in Hastings' subsequent impeachment—these were the dunning of Raja Chait Singh of Vārānasi and his deposition in 1781; and the pressuring of the baygums (princesses) of Oudh for the same reason. Hastings' financial difficulties at the time were great, but such actions were harsh and high-handed.

Hastings' relations with his council

The impeachment of Warren Hastings (1787–95) at the behest of Edmund Burke and the Whigs, which followed his return from India and ended in his acquittal but retirement, was a kind of very rough justice. Hastings had saved for the company its Indian dominions, and he was relatively incorrupt. But the charges served notice that the company's servants were responsible for their actions toward those they governed, and for these actions they were answerable to Parliament. Hastings was so identified with the company's rule that he was the inevitable target for any such assertion of principle.

The company and the state. During the first half of the 18th century the East India Company was a trading corporation with a steady annual dividend of 8–10 percent, offering its servants prospects of a modest fortune through private trade, along with great hazards to health and life. It was directed in London by 24 directors—elected annually

by the shareholding body, the Court of Proprietors—who worked through a series of committees.

The Bengal adventure from 1757 turned the two courts—of directors and proprietors—into political bodies, because they now controlled a great Eastern state. Shares became political counters, the purchase of which might secure votes needed to change the company's policy. A second result was the return to Britain of the company's servants with fortunes; their ostentation and lack of restraint earned them the title of nabobs. These events soon produced reactions. The shareholders wanted to share in this new wealth, in the guise of increased dividends; and the directors wanted the company as well as its servants to benefit from this wealth. Two processes were thus set in motion—one a rising pressure for increased dividends and the other an attempt by the company to discipline its servants and to secure some profit for itself. Broadly speaking, it was the success of the first and the failure of the second that provoked state intervention in the company's affairs.

State intervention in company affairs

The close personal connection between the "direction" and the company's servants themselves weighed heavily and eventually stultified the directors' efforts. It produced an infirmity of purpose, which led to the return to Bengal by one faction of servants dismissed for irregularities by another—a factionalism epitomized by the struggle between Clive and Sulivan for control of the company. These developments occupied the 1760s, drastically reducing the prestige of the company. On the side of discipline, alarm at the overruling of Vansittart and the wars against Mīr Qāsim and Shāh ʿĀlam led to the dispatch of Clive as governor in 1765. As the effect of Clive's measures diminished after his return to England in 1767, three "supervisors" were dispatched to Bengal in 1769 with plenary powers, but they were lost at sea. Then Warren Hastings was appointed in 1772 with a reform mandate. But it was too late, for bankruptcy was now knocking at the door.

The company had hoped for large profits from Clive's first control of Bengal. The hopes then shortly dashed were revived by his second governorship. Clive believed that he had secured a revenue surplus of £2,000,000 for the company. On the strength of these expectations, the company's dividend was raised to 12.5 percent in 1767; in the same year the first signs of parliamentary opposition were bought off by the offer of £400,000 a year to the state in return for undisturbed possession of Bengal. As the expectations withered, this became a financial millstone that compelled the company in 1772 to ask for the loan of £1,000,000 to avert bankruptcy. This opened the floodgates of parliamentary criticism, leading to committees of inquiry and revelations of malpractices, to Clive's suicide (1774), and to the beginning of state intervention.

The Regulating Act

In 1773 the British government gave a loan of £1,400,000 to the company; but its price was the Regulating Act, passed the same year. The act sought to "regulate" the affairs of the company, both in London and India. In London, the qualifications fee for a vote was raised from £500 to £1,000, and the directors' terms were extended from one to four years, with a year's gap before re-election. This ended the jobbing of votes for the control of policy by private interests and gave continuity of policy to the direction. In India, a governor generalship of Fort William in Bengal was established, with supervisory control over the other Indian settlements and Warren Hastings as its first incumbent. Hastings was given four named councillors, but future appointments were to be made by the company. Finally, a Supreme Court with a chief justice and three judges was set up.

The Regulating Act was a first step toward taking the political direction of British India out of the hands of the company and of securing a unified overall control. But it had serious defects, which bedeviled administration in Bengal and made India (despite British preoccupation with the American Revolution) a leading subject of political controversy over the next 20 years.

The governor general possessed no veto in his council. With three political councillors from Britain, each ready to take Warren Hastings' place, this led to his virtual supersession by the majority for two years and to a paralysis of the executive. Hastings used the energy in fighting his council that should have gone to reforming Bengal. The superintending power added responsibility with little power to enforce it. The Supreme Court decided to administer English law (the only law it knew) and to apply it not only to all the English in Bengal but also to all Indians connected with them; in practice this meant those Indians in Calcutta, and it led to such grave abuses as the hanging of Nand Kumar for an offense not recognized as being capital in any Indian code.

In 1780 the company's privileges ran out; but this was during the crisis of the American Revolution, so that a decision was delayed until 1784. Charles James Fox's radical measure to transfer the control of British India to seven commissioners was defeated by the King's influence in the House of Lords, but the next year the matter was settled for more than 70 years by the prime minister William Pitt's India Act of 1784. Its essence was the institution of a dual control. The directors were left in charge of commerce and as political executants; but they were politically superintended by a new Board of Control, the president of which, in the person of Henry Dundas, soon became the virtual minister for India. The directors dealt with the board through a secret committee of three, but their dispatches to India could be altered, vetoed, and dictated by the board. The governor general could be recalled by the crown. In India, the governor's council was reduced to three, including the commander in chief, and by an amending act he acquired the veto, which Warren Hastings had missed so much. Finally, there was to be a parliamentary inquiry before each 20-year renewal of the company's charter.

Pitt's India Act of 1784

Pitt's India Act proved to be a landmark because it gave the British government control of policy without patronage. The cumbrous dual system developed into a seesaw arrangement of give and take, becoming ever stronger on the government side as greater ability, influence, and power had their effect. The inquiry provision produced a national inquest on Indian affairs every 20 years, the conclusions of which marked successive stages in the diminution of the company's political power. On the first such inquiry, in 1793, the company repelled an attempt to compel it to support Christian missionary work; this incident led to the foundation of the Church Missionary Society in 1799. In 1813 the company was obliged by Parliament to admit missionaries and was deprived of its monopoly on trade with India. By the Act of 1833 it lost its trade altogether and was thenceforth a governing corporation under increasing state surveillance. In 1853, with the introduction of competitive examinations, the company lost most of its patronage and also had to admit nominated directors. Policies were increasingly dictated to a sulky or apathetic board. The last case of the recall of a governor general by the company was that of Lord Ellenborough in 1844; this was the real swan song of the company, because it was recognized at the time that such a thing could never happen again. The company had become a managing agency on behalf of the British government.

Relations with the Marāṭhās and Mysore. After Clive's settlement in 1765, the East India Company had no desire for any further acquisitions. Its object was still trade; it regarded the acquisition of Bengal as a political framework for the safe conduct of trade, justified by the danger of near anarchy in its most profitable scene of operations. But such a resolution was easier to make than to keep. Indian states were ever ready to seek European help in achieving their own projects; many of the company's servants looked longingly at territorial revenues that might assist their own enrichment, and the exigencies of Indian politics at times made nonalignment difficult to observe.

The three centres of the company's power

In 1765 the three centres of the company's power were independent of each other, but the post-Mughal Indian pattern was becoming clear. In the north there were the Mughal fragments of Allahābād, Oudh, and Delhi, with the Sikhs resurgent in the Punjab. In the Deccan the Niẓām of Hyderābād maintained uneasily his Mughal regime, sometimes overwhelmed by two vigorous and expansive powers—the Marāṭhās and Mysore.

The Marāṭhās had made their bid for the Mughal succession in the previous decade, and they were now recovering

from a disastrous defeat at Pānīpat (1761). The unified leadership of the peshwa had given way to a confederacy of the peshwa and four military dictatorships developing into monarchies. The Marāṭhās were restless, energetic, and acquisitive; their greatest enemy was their own divisions. In the south the old Hindu state of Mysore had passed into the hands of a Muslim military adventurer of genius, Hyder (Haidar) Ali Khan, in 1762.

When Warren Hastings took overall control of the company's possessions in 1774, Madras had already stumbled into war with Hyder Ali and had submitted to a virtually dictated peace under the walls of Madras in 1769. The Nawab of the Carnatic had become by degrees dependent on the company because he needed its support against the threat of Hyder and the Niẓām. Ingenious and feckless, the Nawab involved Madras in south Indian politics and the company in his affairs by borrowing from its servants.

Hastings' diplomacy

Hastings had a natural gift for *Realpolitik,* but he was tied to a policy of nonaggression. Much of his diplomatic skill was spent repairing the blunders of others. His major work for British India was preserving the company's dominion against a coalition of country powers, virtually unaided from home at a time when Britain was itself hard pressed both in America and by a European coalition. His first work was to safeguard Bengal from the reviving power of the Marāṭhās, who had conducted Shāh ʿĀlām to Delhi in 1771. Hastings intervened and handed Allahābād and Kora to Shujāʿ-ud-Dawlah of Oudh in return for a subsidy and a treaty. The following year he found himself assisting the Nawab of Oudh to crush the Afghan Ruhellas between the Jumna (modern Yamuna) and the Ganges (this stroke was the first item in the indictment at his impeachment, but its effect was to stabilize the north Indian situation for the next ten years).

In western India Hastings was the victim of Bombay brashness and of directorial blunders. A succession struggle in Poona for the peshwaship led Bombay to support Raghunāth Rāo in the hope of securing the islands of Salsette and Bassein. When this was countermanded by Calcutta, London intervened to renew the venture. In 1779 a British army was surrounded on its way to Poona, one month before a force sent by Hastings completed a brilliant march across India at Surat. In 1782 the British made peace with the Peshwa, abandoning Raghunāth and having only Salsette to show for seven years of war. This first round with the Marāṭhās was a draw.

While this war was in progress, Hastings was confronted with a far greater menace. In 1780 the ineptitude of Madras provoked a coalition of the Niẓām, Hyder, and the Marāṭhās, which defeated the company's armies and swept over the Carnatic. Though without hope of succour from Britain, itself hard pressed, Hastings set about sustaining the Madras forces and dividing his foes. In 1781 the military balance was restored, and the next year the Marāṭhās made peace. Hyder died (1782), French help arrived too late to affect the issue, and in 1784 the Treaty of Mangalore with Hyder's son Tippu (Tipu) restored the status quo. Hastings had thus little to show in the way of empire building. His feat of defense without external aid was nevertheless remarkable. He preserved the British dominion in India, and by so doing he made it possible for others to extend it. The company had become one of the recognized great powers of India.

Pitt's Act of 1784 reiterated the company's own intentions by forbidding aggressive wars and annexations. Lord Cornwallis and his successor Sir John Shore (governor general 1793–98) were eager to comply, but Cornwallis nevertheless found himself involved in the third Mysore war (1789–92) with Tippu Sultan, who possessed his father's ability without his judgment. The cause was a combination of Tippu's intransigence with conflicting obligations undertaken by the Madras government. It took three campaigns before Cornwallis could bring Tippu to bay. Half his dominions were annexed, more as a precaution than as an exercise in imperialism. But Tippu remained formidable and, not unnaturally, more hostile than ever.

The ascent to paramountcy. At this point in time a radical change occured in British policy. Two causes were principally responsible. There was a growing body of opinion within the company that only British control of India could end the constant wars and provide really satisfactory conditions for trade; full dominion would be economical as well as salutary. The more compelling immediate cause was the transformation of European politics by the French Revolution. A new French threat to India emerged, this time overland, with Napoleon's Egyptian expedition of 1798–99. It was certain that a French army under such a leader would find many friends in India to welcome it, and not least Tippu Sultan.

Moves toward imperialism

The government of Lord Wellesley. The next governor general, Lord Mornington (later title, Wellesley), combined the convictions of the imperialist group with a mandate to deal with the French. Wellesley was thus able to use this fear of the French as a cover for his imperialism until he was near to complete success. His term of office (1798–1805) was therefore a decisive period in the rise of the British dominion.

Wellesley decided first to strike at Mysore, still a formidable military power and avowedly hostile. He had little difficulty getting the Niẓām for an ally and securing the neutrality of the Peshwa. The Niẓām, hard pressed by the Marāṭhās, was persuaded to disband his contingent of French-trained troops in return for a promise of protection. This was the first of Wellesley's subsidiary treaties (see below). Tippu Sultan had entertained French republican envoys and had planted a tree of liberty at Seringapatam; but when the British stormed Seringapatam in May 1799, he was isolated and at bay, and he found too late that concessions, in the Indian tradition, would not save him. Tippu died fighting in the breach. The British rejoicings over this unexpectedly complete victory emphasized the fear that the Mysore monarchy had inspired. Wellesley tempered his imperialism with diplomacy by restoring the child head of the old Hindu reigning family as the ruler of half of Tippu's dominions; the other half was divided between the Niẓām and the company. This was the real beginning of the Madras presidency as a substantial territorial unit.

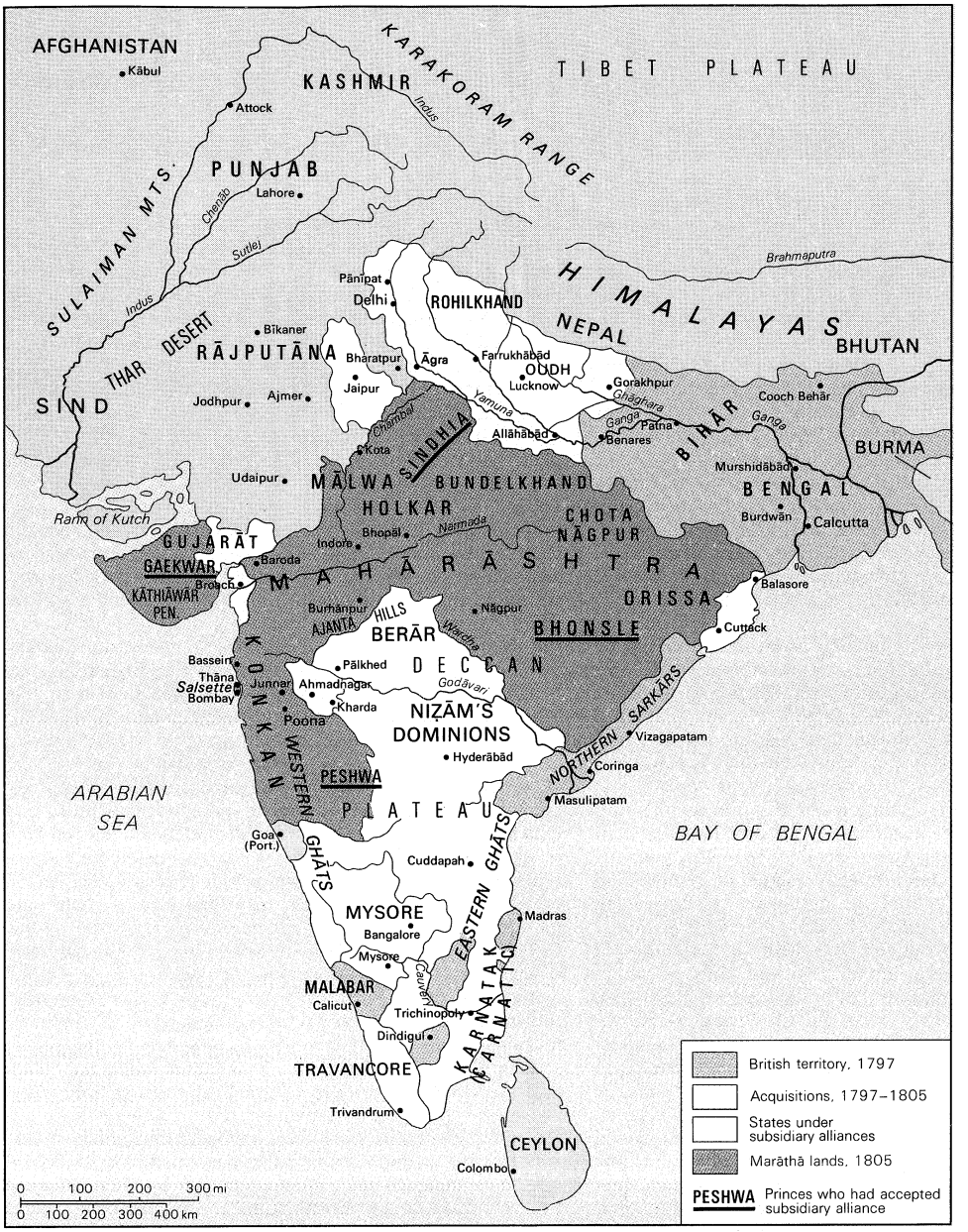
Wellesley's use of *Realpolitik*

For the next three years Wellesley was occupied with certain exercises in *Realpolitik* and with developing his device of the subsidiary treaty. The *Realpolitik* was evidenced in four directions. On the death (1801) of the reigning Carnatic nawab Wellesley took over his territories, pensioning the new nawab with one-fifth of the revenue. The same fate befell the small but highly cultivated state of Tanjore, to the south of the Carnatic (1799), and the city–port state of Surat on a disputed succession.

The biggest of these exercises concerned the Mughal successor state of Oudh in northern India, which had been in treaty relationship with the company since 1765. This rich state had fallen into disorder under the listless though cultured rule of Āsaf-ud-Dawlah; on his death in 1797 a succession dispute and an Afghan invasion of the Punjab gave Wellesley a welcome opportunity for interference. He pressed the Nawab to disband his troops and increase his payment to the company for his subsidiary force. When the Nawab offered to abdicate, it was accepted immediately; but, on finding that abdication would mean annexation and not his son's succession, he withdrew it, and Wellesley treated him as contumacious. In 1801 Wellesley annexed half the state, running along the northern and eastern banks of the Ganges, including Rohilkhand. Whatever the verdict on the means employed, this move had important consequences. Oudh was isolated, and a jumping-off place was secured for an attack on the northern Marāṭhās. The company was no longer looking for buffer states as shields against attack but for territory that would serve as springboards for offensive action.

Development of the subsidiary system

This change of attitude applies to Wellesley's development of the subsidiary system. In the hands of Clive and Warren Hastings it was a defensive instrument to safeguard the company's possessions; in the hands of Wellesley it was an offensive device with which to subject independent states to British control. The essence of the system was that the company undertook to protect a state from external attack in return for the control of its foreign relations. For this purpose it provided a subsidiary force of company troops, who were commonly stationed in a



India, 1797–1805.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1965); George Philip & Son Ltd., London

cantonment near the state capital. The state paid for this force by means of a subsidy, which was often commuted into a cession of territory. In order to protect itself from an external enemy, the state in question bound itself irrevocably to the British power, providing at its heart, as it were, the means of its own coercion should it ever wish to resume independence.

Wellesley first applied this system in 1798 to Hyderābād, when the aging Niẓām, Niẓām Alī Khān, was in dire fear of the Marāṭhās. In 1800 the subsidy was compounded for the Niẓām's share of the Mysore annexations. The same system was applied to Oudh, when the great annexation of 1801 was said to be on account of the subsidiary force. It was then the turn of the Marāṭhās—the remaining bastion of Indian independence. Had the Marāṭhā chiefs remained united, Wellesley could have accomplished little; the death of the young Peshwa released fresh dissensions, however, heightened by that of the minister Nānā Fadnavis in 1800. The chiefs Holkar and Daulat Rāo Sindhia (Sindia) contended for power over the peshwa Bājī Rāo II. On Holkar's success in 1802, Bājī Rāo fled to Bassein and applied for British aid. Such an opportunity at the centre of Marāṭhā power was not to be missed; in addition there was the justification that Daulat Rāo Sindhia, in the north, had 40,000 French-trained troops under a French commander. The Treaty of Bassein (December 31, 1802) placed, as it were, a time bomb at the heart of the Marāṭhā confederacy; British troops were stationed at Poona, at the price of a cession of territory, and the Peshwa was reduced to dependency on the British.

This action provoked the Second Marāṭhā War—at first against Daulat Rāo Sindhia and Bhonsle and then against Holkar. At first the British won resounding victories. Wellesley's brother Arthur (later the duke of Wellington) defeated the Sindhia–Bhonsle coalition in west central India, while Lord Lake broke up Sindhia's French army, occupied Delhi, and took the aged emperor Shāh ʿĀlam II under protection. Then came a check, however, with the intervention of Holkar using the old Marāṭhā cavalry tactics, forcing the British to retreat, and besieging Delhi. Though Holkar was later defeated, this was the signal for which exasperated directors and a doubting ministry had been waiting. Wellesley was recalled. His race for hegemony had been lost in the last lap. But Wellesley's work, avowedly imperialistic, made ultimate supremacy inevitable. The Marāṭhās were too broken to reunite, and there was no one to take their place.

Inevitability of British supremacy

The government of Lord Minto. The next ten years

were an interlude, not a new era. During this period both Sindhia and Holkar plundered the chiefs of Rājasthān, thus preparing them mentally for future Britishoverlordship. Meanwhile, bands of freebooters, known as Pindaris, raided the Nāgpur and Hyderābād states in widening circles and thence entered British territory. These were dispossessed villagers and discarded soldiers—the human flotsam and jetsam of the frequent wars. They had the elusiveness of guerrillas; and they received the tacit countenance of the Marāṭhā princes but not the goodwill of the population, who were their principal victims.

Lord Minto (governor general 1807–13) was occupied with the revived French danger, once more serious with the Treaty of Tilsit (1807) and Napoleon's resulting alliance with Russia. To guard against a French-sponsored Russian attack, British missions were sent to Afghanistan, to Persia, and to Ranjit Singh, the Sikh ruler of the Punjab. The first two proved fruitless, but the Treaty of Amritsar with Ranjit Singh (1809) defined British and Sikh spheres of influence and settled relations for a generation. Minto's other achievement was the capture of the Île de France and of Java from the French-controlled Dutch; the former island became a colony (now the dominion of Mauritius), and the latter was restored under the peace treaty. One result of this episode was the acquisition of the key point of Singapore by Sir Stamford Raffles in 1819.

The government of Lord Hastings. The end of the Napoleonic Wars in 1815 opened a new era in India by strengthening the commercial and economic arguments for completing supremacy and by removing all fear of the French. The Pindari raids, which grew year by year until they affected both the Bengal and Madras presidencies, added further reasons for action. The final act was directed by Lord Hastings (governor general, 1813–23), who came to India as a consolation for his failure to attain the premiership under his friend the Prince Regent. Lord Hastings, however, first had to deal, in 1814–16, with the Gurkhas of Nepal, who inflicted a series of repulses on a Bengal army unprepared for mountain warfare. Each side earned the respect of the other. The resulting Treaty of Kāthmāndu (1816) gave the British the tract of hill country where Simla, the site of the future summer capital of British India, was situated, and it settled relations with British India for the rest of the British period. Nepal remained independent and isolated, supported by the export of soldiers to strengthen the British military presence in India.

The Gurkhas of Nepal and the Pindaris

Lord Hastings then turned to the Pindaris. By a large-scale and well-planned enveloping movement, he hoped to enclose them in an iron net. But this involved entering Marāṭhā territories and seeking the cooperation of their princes. Sindhia agreed after agonizing indecision, and this really settled the issue. Holkar's state was in disorder and was easily defeated. Both the Raja of Nāgpur and the Peshwa resisted and attacked the British forces stationed under their respective subsidiary treaties. Nāgpur quickly collapsed, but the Peshwa kept up a running fight before surrendering in June 1818. The Pindari bands themselves, chased hither and thither, broke up or surrendered.

The East India Company was thus the undisputed master of India, as far as the Sutlej River in the Punjab. This episode was completed by the acceptance of British suzerainty by the Rājput chiefs of Rājasthān, central India, and Kāthiāwār, as they had formerly accepted the Mughals. Thus the year 1818 marks a watershed, when the British Empire in India became the British Empire of India.

The British Empire of India

The settlement of 1818. The diplomatic settlement of 1818, except for a few annexations before 1857, remained in force until 1947 and is therefore worth some attention. The company, under the influence of its guiding star of economy, was glad to be saved the expense of administering as much of India as possible, especially the less fertile portions. Having controlled the larger states by its subsidiary forces (for which they paid), it was content with tribute from the remainder, with control posts at strategic points. Thus Kāthiāwār was controlled from Baroda; Rājasthān, from Ajmer. There was no thought of integration, as in Mughal days. The states were isolated and excluded from any connection with the British. About half of India remained under Indian rulers, robbed of any power of aggression and deprived of any opportunity of cooperation—in the south were the large units of Mysore, Hyderābād, and Travancore; in the west, the states of Śivājī's family; across the centre to the east, Nāgpur and a number of poor "jungle" states; in the west and west central areas, numerous Rājput and other Hindu chiefs with the surviving Marāṭhā states of Sindhia, Holkar, and the Gaekwar; beyond the Jumna, some Sikh princedoms; and in the Ganges Valley, the still prosperous and disorderly state of Oudh. In all there were more than 360 units; politically they were like the surviving fragments of a broken jigsaw puzzle, with all its complexity but without its unity.

The subjection of a whole subcontinent containing a unique civilization has long been a source of historical wonderment. The one-time explanations of innate superiority and of mere fate are no longer seriously entertained. But analysis goes far to dissipate the mystery. In the first place, the feat was not unique; the Turkish Muslims had twice done much the same—for shorter periods, it is true, but also with fewer resources. All these achievements were made possible by the innate divisiveness of Hindu society, rent by class and caste divisions, which rendered it unusually willing to call in unwelcome outsiders to defeat the still more unwelcome neighbour. The foreigners, asked in the first resort to assist in defeating a rival, were in the last resort accepted as masters in preference to dominance by a rival. Thus Marāṭhās preferred the British to the Mughals or Muslims; and the Niẓām, the British to the Marāṭhās. Long historical memories can be inhibiting as well as inspiring. Against this setting can be set the company's urge toward unity in the interests of trade. Even when its Indian trade was no longer profitable, India gave profits to others, and its opium bought the China tea, which gave the company its overall profits. Given the fact of expansion, Britain enjoyed the advantage of overseas reinforcement through its sea power and of reserves of power, far greater than that of any Indian prince, through its rapidly expanding industrial economy. A lost battle for the British was an incident in a campaign; for the Indian prince, usually the end of the chapter. Then there were the technical advantages of arms and military discipline and the immense general advantage of a disciplined civilian morale. In the later stages this was boosted by the rising self-confidence of Europeans in general, with its belief that the western European civilization was the only truly progressive one that had ever existed. For the Hindu, on the other hand, his world was at its lowest ebb—in the Kali Yuga, or dark age—while the Muslim believed in inscrutable fate. The Hindu's heart was in his religiocultural complex, and political dominion meant little to the ordinary Hindu so long as this remained untouched.

Analysis of the British achievement in India

Organization and policy in British India. The realization of supremacy in 1818 made urgent the problem of the organization of and determination of policy for British India. So far only Bengal had been deliberately organized; the extensive areas annexed after 1799 in the north and the south were still under provisional arrangements. Now the Peshwa's dominions in the west awaited settlement. The administrators of the first 30 years of the 19th century gave British India the form it retained until 1947. Outstanding among them were Sir Thomas Munro in Madras, Mountstuart Elphinstone in western India, and Sir Charles Metcalfe in Delhi; to this trio must be added a fourth—Holt MacKenzie, whose planning determined the lines of settlement from Vārānasi to the Jumna.

The form of British India

Organization. The only areas so far definitely settled were those of Bengal, Bīhār and Orissa. Lord Cornwallis had been charged by Pitt with the reorganization of Bengal under the new act. Besides being a soldier of distinction, Cornwallis was a man of outstanding integrity, a landlord with rural tastes, and an instinctive Whig. Cornwallis first undertook a cleansing of the existing system. Discipline among the company's servants was enforced at the price of dismissal. Private trade was forbidden to all government officers, and the service was divided into administrative and commercial branches. These measures were coupled with a generous salary system, which removed the

temptation to corruption. From this time the company's service began to gain its later reputation for efficiency and integrity. All this could be done because the governor general, with his Council of three and his veto power, was now unassailable to the attacks that had ruined Vansittart and frustrated Warren Hastings.

From this base Cornwallis built up the Bengal system. Its first principle was anglicization. In the belief that Indian officials were corrupt (while British corruption had been cured), all posts worth more than £500 a year were reserved for the company's covenanted servants. Next came the government. The 23 districts each had a British collector with magisterial powers and two assistants, who were responsible for revenue collection. The judicial system was organized with district judges for both civil and criminal cases. In civil cases there were four courts of appeal; and in criminal, four circuit courts. Criminal justice was taken over from the nawab's deputy, thus removing the last shred of Mughal authority. The criminal code was the Islāmic, humanely modified. A new police force replaced the former local constables of the *zamīndār*s (land revenue collectors). This new system, which, with its division of authority, showed its Whig influence, was rounded off by the proclamation of the rule of law, making all governmental acts answerable in the ordinary courts of law. Though hardly noticed at the time by Indians, it was a radical innovation with far-reaching effects. It was a charter of civil as distinct from political liberty.

Cornwallis' permanent settlement

Cornwallis' permanent settlement of the land revenue is the measure that most deeply affected the life and structure of Indian society, three-quarters of the revenue coming from the land. He found a system of hereditary *zamīndār*s, who had acquired police and magisterial powers as well, and who were much shaken by the frequent changes of revenue policy under the British. The "settlement" was the decision in 1793 to stabilize the revenue demand at a fixed annual figure, with a commission to the *zamīndār* for collection, and to regard him as the owner of his *zamīndāri;* he had the disposal of wastelands within his jurisdiction, but these lands were liable to be sold for arrears of payment. Thus the land revenue collector became a landlord, with the Achilles heel of sale for arrears, while the tiers of lesser landholders became his tenants. The *zamīndār* reaped the profit of rising prices and of the cultivation of wasteland, while the classes below him lost their occupancy rights. The intended protection of these tenants proved illusory because their rights were customary, unsupported by documents. The legal cases that ensued clogged the courts to the point of breakdown. Initially, the *zamīndār* often lost his holding because the fixed demand was pitched too high. The net result of this measure was the creation of a landlord class, loyal to the British connection but divorced from touch with the cultivators. The government, receiving the revenue from the *zamīndār*s, knew little of the people and could do little for them.

At first the Bengal system was thought to provide the key to Indian administration, but doubts multiplied with the years. In Madras, Sir Thomas Munro retained the paternal framework of government but introduced a radically differing method of revenue management known as ryotwar, in which the settlement was made directly with the cultivator, each field being separately measured and annually assessed. The system eliminated the middleman but sometimes placed the cultivators at the mercy of lower officials, who often formed cliques of caste groups. Munro considered that innovation and ignorance were the ruling British vices. His system tended to be static and to allow the subordinate tail to wag the directing British dog.

Reconciling the Marāṭhās

In western India, Mountstuart Elphinstone had the problem of reconciling to British control the resentful Marāṭhās of the Peshwa's dominions. With a masterly mixture of tact and firmness he largely succeeded. He retained Indian agency as far as possible, and he allowed the Marāṭhā nobles, or *jāgīrdār*s, to retain most of their land and many of their privileges. He even continued some donations to Hindu temples. He used the ryotwar method of assessing land revenue, collecting through local officials from the village headmen. In Bombay he encouraged Western learning and science, tempting suspicious Brahmins to open their minds to the West. He foresaw the ultimate end of British rule through voluntary westernization, and he took the first steps toward introducing the new world without antagonizing the old.

In the north, Sir Charles Metcalfe discovered the largely autonomous village with its joint ownership and cultivation by caste oligarchies. He believed this to be the original pattern of rural organization throughout India, and it became his passion to preserve it as far as possible in current conditions. Like Munro and Elphinstone, he was suspicious of change and wished to leave the villagers alone as far as possible. In this he was powerfully supported by the work of Holt MacKenzie, the Bengal secretary whose memorandum of 1819 set a course of recognition and record of village rights for the whole of the northwestern provinces (as later revised and codified, this marked the end of the Bengal system of permanent revenue settlement).

Administration of British India

The resulting system of administration of British India was still largely Indian in pattern, though it was now British in direction and superintendence. It was paternalistic and hierarchical; and it suffered, like its immediate predecessors, from a chronic tendency to overassess. The emperor was replaced by the mystical entity, the Company Bahadur; and its representative, the governor general, moved about with almost equal pomp. The higher direction was exclusively European, but the officers acted in a Mughal spirit, and the administration at subdistrict and village level went on much as before. But there were also large changes. The British established on a national scale the idea of property in land, and the resulting buying and selling caused large class changes. Their new security benefitted the commercial classes generally, but the deliberate sacrifice of Indian industry to the claims of the new machine industries of Britain ruined such ancient crafts as cotton and silk weaving. The new legal system, with its network of courts, proved efficient on the criminal side but was heavily overloaded on the civil.

The strain and the scandal of this situation created a demand for increased Indian agency and caused the first breaches in the British monopoly of higher office. Indianization began with the confessed inefficiency of the British legal system. The picture is completed by the company's army, separately organized in the three presidencies and officered, like the civil service, exclusively by the British. It was backed by contingents of the British army. The Bengal army preponderated in numbers and fighting spirit. By European standards it was cumbrous and inefficient; some of its defects were exposed in the early days of the Nepal war. But it was more than a match for anything that could be brought against it. Only the Russians, could they have moved so far in force, might have made short work of it.

The determination of policy. The administration of British India thus established was impressive though ponderous. But it was essentially static; it was a repair of the machinery of government without any decision about its direction. Such a situation in a subcontinent could not be viable for long.

The nature of the government in India

In the early 19th century a great debate went on in Britain about the nature of the government in India. The company wanted India to be regarded as a field for British commercial exploitation, with the company holding the administrative whip with one hand and exploiting with the other. This pleased no one but the company itself. As an extension of this, the new regime could be regarded as a law-and-order or police state, holding the ring while British merchants in general traded profitably. But this was assailed from several quarters. There was the Whig demand, first voiced by Edmund Burke in his campaign against Warren Hastings, that the Indian government must be responsible for the welfare of the governed. This was reinforced by Evangelicals in England, both Anglican and Baptist, who added the rider that as the ruler Britain was responsible for India's spiritual and moral welfare as well. The Evangelicals were a rising force, influential in the British "establishment." Their specific for India, as a preparation for conversion, was English education. They were reinforced in this by the rising group of freethinking Utilitarians—followers of Jeremy Bentham and John Stu-

art Mill—who were influential in the company's service, who wished to use India as a laboratory for their theories, and who thought Indian society could be transformed by legislation. Finally, there were radical rationalists who had borrowed the doctrine of human rights from France and wished to introduce them into India; and on the practical side there was a body of British merchants and manufacturers, who saw in India both a market and a profitable theatre of activity and who chafed at the restraints of the East India Company's monopoly.

Some of these influences seeped into the Tory ascendancy, which lasted until 1830. In 1813 the East India Company lost its monopoly of trade with India and was compelled to allow free entry of missionaries. British India was declared to be British territory, and £10,000 was to be set aside annually for the promotion of both Eastern and Western learning. But the real breakthrough came with the governor generalship of Lord William Bentinck (served 1828–35) and with the Whig government that, from 1830, carried the great Reform Bill.

Reforms of Lord William Bentinck

Bentinck was a radical aristocrat. His administrative reforms were in line with Utilitarian theory but with deference to local conditions and in harmony with his own military sense of command. In Bengal, the collector was made the real head of his district by the addition of civil judgeship to his magistracy; he was also disciplined by the institution of commissioners to superintend him. The judiciary was overhauled with the same eye to a chain of authority.

But it was as a social reformer that Bentinck made an indelible mark on the future of India. He was commissioned by the directors to effect economies in order to show a balanced budget in the approaching charter-renewal discussions. In doing this he incurred much odium; but he was able to take the first steps in Indianizing the higher judicial services. On his arrival Bentinck was confronted with an agitation against suttee, or the burning of Hindu widows on the funeral pyres of their husbands. In suppressing the practice, he had to face the reproaches of both Hindus and Europeans on the grounds of religious interference. But he was also fortified by the support of the Hindu reformer Rammohan Ray. In thus acting and in prohibiting child sacrifice on Saugor Island and discouraging infanticide—a widespread practice among the Rājputs—Bentinck established the principle that the general good did not permit violations of the universal moral law, even if done in the name of religion. The same principle applied to the suppression of *ṭhagī* (thuggee), or ritual murder and robbery by gangs in central India in the name of the goddess Kālī. On a more positive side Bentinck substituted English for Persian as the language of record for government and the higher courts, and he declared that government support would be given primarily to the cultivation of Western learning and science through the medium of English. In this he was supported by T.B. (later Lord) Macaulay, whose speech in Parliament on July 10, 1833, was the most eloquent expression of the new spirit ever made.

This period saw the British in India committed to promoting the positive welfare of India instead of merely holding a ring for trade and exploitation; to introducing Western knowledge, science, and ideas alongside the Indian, with a view to eventual absorption and adoption, and to the promotion of Indian participation in the government with a view to eventual Indian self-government. It was the changeover from the concept ofa Mughal successor state—the Company Bahadur—to that of a westernized self-governing dominion. In the former case, the British were wardens of a stationary society; in the latter, trustees of an evolving one.

Concept of a western-zed self-governing dominion

A word should be added about the Indian states. Their place in British India was also a subject of the great debate on the future of India. On the whole, the argument for subordinate isolation held, and no great change occurred in their status until after the Mutiny of 1857 (see below *The mutiny and great revolt of 1857–59*). Out of the discussions, however, came the de facto principle of British paramountcy, which was increasingly assumed though not openly proclaimed. The only important change before 1840 was the takeover of Mysore in 1831 on the ground of misgovernment; it was not annexed, but it was administered on behalf of the raja for the next 50 years.

The completion of dominion and expansion. After the settlement of 1818, the only parts of India beyond British control were a fringe of Himalayan states to the north and a block of territory in the northwest covering the Indus Valley, the Punjab, and Kashmir. To the south Ceylon was already occupied by the British, but to the east lay the valley and hill tracts of Assam and the Buddhist kingdom of Burma straddling the Irrawaddy River.

The Himalayan states were Nepal of the Gurkhas, Bhutan, and Sikkim. Nepal and Bhutan remained independent throughout the British period; but Sikkim came under British protection, providing the hill station of Darjeeling as Bengal's hill capital. The valley and hill tracts of Assam were taken under protection to save them from Burmese attack. From this time the Indian tea plant was cultivated, after the failure of Chinese imported ones, and commenced the great Indian tea industry.

In the early 19th century the Burmese were in an aggressive mood, having defeated the Thais (1768) and subjected Arakan and hill states on either side of the river valleys. Attacks on British territory in 1824 started the First Anglo-Burmese War (1824–26), which, though mismanaged, led to the British annexation of the coastal strips of Arakan and Tenasserim in 1826. The Second Anglo-Burmese War (1852) was caused by disputes between merchants (trading in rice and teak timber) and the Rangoon governor. The governor general, Lord Dalhousie (served 1848–56), intervened, annexing the maritime province of Pegu with the port of Rangoon in a campaign this time well-managed and economical. Commercial imperialism was the motive for this campaign.

Wars with Burma and Afghanistan

To the northwest British India was bounded by the Sikh kingdom of Ranjit Singh, who added Kashmir to his state in 1819 and Peshāwar in 1834. Beyond was confusion, with the Afghan monarchy in dissolution and its lands parcelled between several chiefs and Sind, controlled by a group of *amīr*s, or chiefs. British indifference changed to action in the 1830s, owing to the advance of Russia in Central Asia and to that nation's diplomatic duel with Lord Palmerston about its influence in Turkey. Afghanistan was seen as a point from which Russia could threaten British India or Britain could embarrass Russia. Lord Auckland (served 1836–42) was sent as governor general, charged with forestalling the Russians; and from this stemmed his Afghan adventure and the First Afghan War (1838–42). The method adopted was to restore Shāh Shojāʿ, the exiled Afghan king, then living in the Punjab, by ousting the ruler of Kābul, Dōst Moḥammad. Ranjit Singh cooperated in the enterprise but cleverly avoided any military commitment, leaving the British to bear the whole burden. The route of invasion lay through Sind, because of Sikh occupation of the Punjab.

The *amīr*s' treaty of 1832 with the British was brushed aside, and Sind was forced to pay arrears of tribute to Shāh Shojāʿ. At first things went well, with victories and the occupation of Kābul in 1839. But then it was discovered that Shāh Shojāʿ was too unpopular to rule the country unaided; the British restoring force thus became a foreign occupying army—anathema to the liberty-loving Afghan—and was regularly engaged in putting down sporadic tribal revolts. After two years, a general revolt in the autumn of 1841 overwhelmed the British garrison. Meanwhile, the Russian menace in eastern Europe had receded. Auckland's successor, Lord Ellenborough (served 1842–44), arranged for the evacuation of the country by means of a converging march on Kābul from Qandahār in the south and Jalālābād in the north and a return through the Khyber Pass. Thus honour was satisfied, and the fact of defeat was glossed over. Shāh Shojāʿ was shortly murdered. The episode demonstrated, at a heavy price in terms of money and human suffering, the ease with which Afghanistan could be overrun by a regular army and the difficulty of holding it. The enterprise, though conceived as an insurance against Russian imperialism, developed into a species of imperialism itself. Economics joined with Afghan spirit to put a limit on British expansion in this direction.

Relations with Sind

After the Afghans came Sind. There was little to be said for the *amīr*s themselves—a group of related chiefs who had come to power in the late 18th century and had kept the country in poverty and stagnation. A treaty in 1832 threw the Indus open to commerce except for the passage of armed vessels or military stores; at the same time the integrity of Sind was recognized. Thus Auckland's march through Sind was a clear violation of a treaty signed only seven years before. Sore feelings at the turn of events in Afghanistan produced a final breach. On a charge of unfriendly feelings by the *amīr*s during the Afghan War, Karāchi, occupied in 1839, was retained. Further demands were then made; the moderate resident James Outram was superseded by the militant general Sir Charles James Napier; and resistance was provoked, to be crushed at Miāni in 1843. Sind was then annexed to the Bombay Presidency; after four years of rough-and-ready rule by Napier, it was put on the road to prosperity by Sir Bartle Frere.

There remained the great Sikh state of the Punjab, the single-handed creation of Ranjit Singh. Succeeding to a local chiefship in 1792 at the age of 12, he occupied Lahore in 1799 under a grant from Zamān Shāh, the Afghan king. He could thus pose as a legitimate ruler, not only to his own Sikhs but to the majority of Muslims of the Punjab. From this start he extended his dominions northwestward as far as the Afghan hills and including Kashmir and Jammu and southwestward toward Sind to Multān and near Sheikhpura. The Treaty of Amritsar with the British in 1809 barred his expansion southeastward; besides directing Ranjit's expansionism northwestward, it produced an admiration for the disciplined company's troops, who coolly repelled the wild Sikh Akalis when they attacked the British at Amritsar. From that time dated the formation of the formidable Sikh army with its 40,000 disciplined infantry, 12,000 cavalry, and powerful artillery. It was generally agreed to compare favourably for efficiency with the company's forces.

Relations with the Sikhs

Ranjit Singh employed Hindus and Muslims besides Sikhs, but his regime was in fact a Sikh domination based on tacit Hindu support and Muslim acquiescence. It used most of the revenue to support the army, which made it apparently powerful but retarded development. It was a highly personal system, centred on Ranjit himself. It was thus one that the company would not lightly attack but that had inner weaknesses behind its impressive facade. These weaknesses began to be exposed on the morrow of Ranjit's death in 1839; within six years the state was on the verge of dissolution. Army disbandment or foreign adventure seemed the only way for the Sikhs to deal with this crisis. The former being impossible, at length the Rani (Hindī *rānī*) regent (for the boy prince Dalip Singh), the chief minister, and the commander in chief agreed on a move against the British. The frontier was crossed in December 1845, and a sharp and bloody war ended in a British victory at Sobrāon in February 1846. The British feared to annex outright a region full of former soldiers and wished to retain a buffer state against possible attack from the northwest. By the Treaty of Lahore they took Kashmir and its dependencies, with the fertile Jalandhar (now Jullundur) area, reduced the regular army to 20,000 infantry and 12,000 cavalry, and exacted an indemnity of £500,000. They then sold Kashmir to the Hindu chief Gulab Singh of Jammu, who had changed sides at precisely the right moment. Thus a chronic political problem for the subcontinent was created.

The balance thus created between annexation and autonomy, between Lahore and Jammu and Lahore and the Afghans, did not last. There was no subsidiary force, for fear of antagonizing the Sikhs. Two years later a rising at Multān became a national Sikh revolt, and the Sikh court was helpless. Another brief and still more bloody war, with the Sikhs this time fighting resolutely, ended with their surrender in March 1849 and the British annexation of the state.

British annexation of the Sikh state

Annexation this time proved viable because of the underlying tension between Sikhs and Muslims. The Sikhs preferred the British to a Muslim raj (Hindī *rāj*).

The British depressed the sirdars, or Sikh leaders, but left the rest of the community and its religion untouched. The Sikhs sided with the British during the 1857 mutiny; the Muslims, however, could not forget their loss of power to the Sikhs. There was little commercial exploitation of the state, and the Sikhs found employment in the army. Lord Dalhousie closely supervised the administration through a like-minded agent, Sir John Lawrence. The pair produced a new model administration, establishing what was known as the Punjab school. It was noted for strong personal leadership, on-the-spot decisions, strong-arm methods, impartiality between the communities, and material development, including irrigation. A canal, a road, or a bridge was the Punjabi officials' delight. The cultivator was preferred to the sirdar (landlord); the countryman, to the townsman. The Punjab system was strong and efficient, creating prosperity; but it never reconciled the communities or welded them into unity. The Punjab feeling thus remained a territorial sentiment, without ever becoming a national spirit. This failure was a basis for later partition.

Lord Dalhousie's reign (served 1848–56) is often regarded as an exercise in imperialism; in fact it was more an exercise in westernism. Dalhousie was a man of great drive and strong conviction. His convictions were westernizing and mainly motivated his actions. In general he considered Western civilization to be far superior to that of the Indian, and the more of it that could be introduced the better. Along these lines he pushed Western education, introducing a grant-in-aid system, which later proliferated Indian private colleges, and planned three universities. Socially, he allowed Christian converts to inherit the property of their Hindu families. Materially, he extended irrigation and the telegraph and introduced the railway.

Politically, British administration was preferable to Indian and it was to be imposed where possible. Externally, this led to annexation as in the Punjab and Burma, rather than to the control of foreign relations or to a British-superintended native regime. Internally, it led to the annexation of Indian states on the ground of misgovernment or the doctrine of lapse. The leading case of misgovernment was the disorderly but prosperous Muslim state of Oudh—the oldest ally of the British. The doctrine of lapse concerned Hindu states where rulers had no direct natural heirs. Hindu law allowed adoption to meet these cases, but Dalhousie declared that such must be approved by the supreme government; otherwise there was "lapse" to the paramount power, which meant the imposition of the usual British administration. The three principal cases were Sātāra in 1848 (the descendants of the Marāṭhā king Śivājī), Jhānsi (1853), and the large Marāṭhā state of Nāgpur (1854). Finally, Dalhousie abolished the titular sovereignties of the Carnatic and Tanjore, declined to continue the former peshwa's pension to his adopted son, and gave notice of the termination of the Mughal imperial title itself.

The first century of British influence. The onset of British influence in India differed both in manner and in kind from that of other historical invasions. The British came neither as migrating hordes seeking new homes nor as armies seeking plunder or empire. They had no missionary zeal. Yet eventually they did more to transform India than did any previous ruling power. This apparent paradox requires some explanation.

Political effects. At first the British were only one group of foreign traders among several, fortunate to find in the Mughals a firm government ready to foster trade. Their entry into politics was gradual, first as allies of country powers, then as their virtual directors, and only finally as masters. At each step they were assisted by local powers who preferred British influence to that of their neighbours. It was only in the 20 years 1798–1818 that they were consciously imperialistic and only thereafter that they treated India as a conquered rather than an acquired country. The effect of this was to replace the defunct Mughal regime and the abortive Marāṭhā successor empire with a veiled but very real hegemony.

Indians were used to the idea of political unity and overlordship. They admired the British for being more successful than themselves, while reprobating many of the British habits and doctrines. But the old ruling classes showed little sign of adopting British institutions; after

1818 they withdrew within themselves, nursing their memories rather than feeding their hopes. The Indian regimes of 1857 all assumed a traditional form. The one department in which Western influence was effective was the military. From the time of Mīr Qāsim in Bengal (1760–63), Indian princes began to train troops in the European manner and to form parks of artillery. Some of these bodies, culminating in Ranjit Singh's Sikh army, attained a high degree of efficiency. Their problem was maintenance, for most princes lacked the necessary resources to pay their men and officers regularly and maintain their arms. Indian opinion in general saw the British as the latest holders of the traditional paramount power. There was no novelty in the foreign personnel of the government, for this had been a Mughal practice, too. What was new was the artificial division between British India and Indian-governed India, with virtually no contact between the two. The Mughals practiced partnership for a century; the Turks and Afghans, subordinate cooperation; the British, it seemed, wished to forget the Indian leaders altogether.

Effect of the company's operations

Economic effects. Things were quite different, however, in the economic field. Up to 1750 the effect of the East India Company's operations was marginal. Production of cotton and silk goods, of indigo, saltpetre, and, later, opium was stimulated in particular areas like Bengal, Gujarāt, and Mālwa, with some gain to the middlemen but no sign of any general rise in living standards. India then, as now, was mainly agricultural; and its industries, though significant, were marginal to its whole economy. The latter changed, however, with the acquisition of Bengal. The bias in favour of British merchants diverted trade from their Indian counterparts, though some of the profit went back to the British merchants' Indian agents. The extravagant present giving, a large abuse of a traditional system, diverted much money to Britain. Still more, the pressure on the *zamīndār*s for more revenue, and theirs in turn on the cultivators, further diminished the Bengali income. To this must be added the operation of monopolies, public and private. When the Bengal famine of 1770 occurred, reckoned to have swept away one-third of the population, little attempt at relief was made, though this, with Bengal's network of waterways, was practicable. The cruel severity with which the revenue was still collected at this time delayed recovery for many years. Economic recovery was further delayed by Warren Hastings' makeshift revenue arrangements; and much dislocation was caused in the social structure, with its own effect on economic life. Cornwallis' permanent settlement (1793), after an initial period of dislocation, gave relief and security to the *zamīndār*s, who benefitted by the rise in prices and the cultivation of wastelands; the cultivators themselves, now the *zamīndār*s' tenants at will, remained as poor as before. Apart from the *zamīndār*s, the only class to benefit from the British was that of the entrepreneurs of Calcutta, who acted as agents and bankers to the British. Thus both Clive's and Hastings' business managers became wealthy landowners. In Madras little could be done until the incubus of the Carnatic nawab's debts was removed and the country was settled after the Cornwallis–Wellesley annexations (1792–99). There, economic settlement turned on the working of the ryotwar revenue system; regularity of collection was offset by severity of assessment, and the same may be said of both western and northern India.

After 1800 there was a new factor, in the appearance of machine-made cotton goods from Britain. These steadily undermined the Indian handicraft industries, until all but the highest and coarsest grades of cloth were squeezed out. The district of Dacca (now Dhākā) in the 1820s was illustrative of this process. From about 1830 the tea industry of Assam began with coffee in the south. Coal mining was begun; but its growth, with that of the jute and cotton-machine industries, had to wait for the second half of the century. The average Indian was far more secure than before, except for famine; but he was not much more prosperous, except in patches. India drifted toward the status of a colonial economy, a supplier of raw materials, a market for manufactured articles, to the profit of the foreigner.

Social effects. The social effects of this period were considerable. They took mainly the form of the displacement of classes. As already noted, there was a general disturbance in Bengal caused by the permanent settlement, whereby the lesser landholders were reduced to the condition of tenants at will. But there was also disturbance among the *zamīndār*s. The first upset followed the famine of 1770, when the cultivators were often too few for the revenue demand to be met, and "farming" the revenue for some time took the place of a revenue settlement. The second upset came with the permanent settlement of 1793, when the revenue figure fixed was in many cases too high for the existing cultivation. By 1820 it was calculated that more than one-third of the estates had changed hands through sale for arrears of land tax. The purchasers were in the main the Calcutta entrepreneurs, newly enriched by their contacts with the British. Many were absentees. The social link between landholder and cultivator had been broken, cash nexus replacing traditional rights.

Displacement of classes

In Calcutta itself, these same rentiers formed a fashionable and intellectual society, from which came the first significant cultural contacts with the West. It was composed of the prosperous section of the three upper Bengali castes, with such others as gained acceptance by their wealth or education. Collectively this literate class was known as the *bhadralok*.

In the north there was less dislocation, though the landholders, many of whom had no title but the sword, tended to be depressed. There was a general recognition of rights and broadly of their protection. The chief sufferers were ruling families, who lost power, and the official aristocracy, who lost office. In the south, chiefs whom Munro dispossessed were largely in the class of robber barons.

In western India, a balance between aristocratic and cultivating rights was perhaps better maintained than elsewhere and relations were harmonious. Of significance there was the rapid development of Bombay from the time it came to possess a large hinterland in 1818. With it came the rise of the enterprising Parsi community.

In general, apart from Bengal, there was some depression of the old aristocracy, a regulation and preservation of lesser landholders' rights, and an encouragement of the commercial classes. Communities did not break up, but their fortunes rose and fell with their ability to adjust to these conditions.

Cultural effects of British influence

Cultural effects. The cultural effects of British influence during the century 1757–1857, though less spectacular, were in the long run farther reaching. At first there was little enough. But as the Europeans grew in political importance, Indians became interested in the causes of the growth, so that the first examples of cultural influence were in the military field. Some Europeans, in their turn, early interested themselves in Indian culture, marked by the foundation of the Asiatic Society of Bengal in 1784, by Sir William Jones, and by the translation of such Sanskrit works as the *Bhagavadgītā* and *Abhijñānaśakuntalā* and of Persian works like the *Ā'īn-e Akbarī*.

As the British completed their supremacy, four Indian attitudes could be discerned. There were Indians who rejected all things Western, retiring to their houses and estates to dream of the past. There were those who were clients and employees of the British, as they had been of the Mughals and the Turks before them, without any intention of giving up their traditional culture. But there were also those who, while remaining good Hindus or Muslims, began to study Western ways and thought for careerist purposes. And there was, finally, a small group who sought to study the ideas and spirit of the West, with a view to incorporating in their own society anything that seemed desirable.

The agents of Western influence were government officials, who carried Western ideas, such as utilitarianism, equality before the law, and property into their administration of revenue and the law, and missionaries, who combined hostility to Hinduism and Islām with the presentation of a new ethic—the practice of good works and the promotion of English education as a preliminary for conversion. It was at this point that the Indian careerist and inquirer met the new Western stream of thought. The English language was popular because it opened paths to

employment and influence; orthodox Hindus patronized the English schools and promoted the Hindu College (now Presidency College) in Calcutta (1816). This college, along with Alexander Duff's Scottish Church College, also in Calcutta, became a centre of Western influence and saw the rise of the Young Bengal movement, the westernizing zeal of which denied the Hindu religion itself.

The leadership of Rammohan Ray

But between the complete westernizers and the careerists was a third group, which found a leader of genius in Rammohan Ray. Making a moderate fortune in Calcutta finance, which he invested in *zamīndārīs*, from 1815 Ray advocated reforms in Hindu society and the acceptance of some features of Western thought. He denounced suttee (the burning of widows) and championed the cause of the Indian widow and wife. He advocated English education as a means of bringing Western knowledge to India. He denounced idolatry and preached monotheism. With his *Precepts of Jesus* he both introduced the Christian ethic into Hindu society and drew the sting of missionary attacks. He finally founded a reforming Hindu body, the Brāhmo Samāj (Divine Society), in 1828. Both careerists and Ray's followers cooperated in the spread of English education, but it was the latter who began the movement of borrowing from the West without any feeling of disloyalty to their past.

By the year 1857 the British had established complete political control of the Indian subcontinent, which they ruled directly or through subordinate princes. They had established an authoritarian system of government, making use of Mughal practice and tradition and supported by an efficient civil service and a relatively efficient army. Princely India remained in a stagnant traditionalism; nominal sovereignties were suppressed in the name of progress. In British India land settlements had produced much social dislocation, while purporting to respect traditional rights and to learn from the past; in particular, the Western concept of property in land had led to much social displacement. The westernized legal system was efficient in suppressing crime but dilatory in upholding rights and incomprehensible in its working. Social evils like suttee, thuggee, and infanticide had been suppressed or discouraged, but Hinduism and Islām were still respected. The revolutionary aspect of the British presence was the decision, taken about the time of Bentinck, to introduce Western knowledge and science through the medium of the English language. Western inventions like the telegraph, scientific irrigation, railways, and steamships followed, throwing India open to the industrial mechanistic and democratic world of the developing West. Along with education went the Christian missionary intrusion, with its moral and ideological challenge. This, in its turn, provoked a creative response from Rammohan Ray's circle, who were laying the foundations of a modernized Hinduism, which was later to find political expression in the Indian National Congress.

THE MUTINY AND GREAT REVOLT OF 1857–59

The mutiny of the Bengal army began on May 10, 1857, at Meerut, when Indian soldiers who had been placed in irons for refusing to accept new cartridges were rescued by their comrades. They shot the British officers and made for Delhi, 40 miles (64 kilometres) distant, where there were no British troops. The Indian garrison at Delhi joined them, and by the next nightfall they had secured the city and Mughal fort, proclaiming the aged titular Mughal emperor as their leader. There at a stroke was an army, a cause, and a national leader—the only Muslim who appealed to both Hindus and Muslims.

Nature and causes of the rebellion. But this movement became much more than a military mutiny. There has been much controversy over its nature and causes. The British military commander Sir James Outram thought it was a Muslim conspiracy, exploiting Hindu grievances. Or it might have been an aristocratic plot, set off too soon by the Meerut outbreak. But the only evidence for either of these was the circulation from village to village of chapatis, or cakes of unleavened bread, a practice that, though it also occurred on other occasions, was known to have taken place at any time of unrest. The lack of planning after the outbreak rules out these two explanations, while the degree of popular support argues more than a purely military outbreak.

Nationalist historians have seen in it the first Indian war of independence. In fact, it was rather the last effort of traditional India. It began on a point of caste pollution; its leaders were traditionalists who looked to reviving the past, while the small new westernized class actively supported the British. And the leaders were not united, because they sought to revive former Hindu and Muslim regimes, which in their heyday had bitterly clashed. But something important was required to provoke so many to seize the opportunity of a military uprising to stage a war of independence.

Military cause of the mutiny

The military cause was both particular and general. In particular, it was the greased cartridges supplied for the new breech-loading Enfield rifles. These had to be bitten off before insertion, and the British manufacturers had supplied a fat of mixed beef and pork—anathema to both Hindus and Muslims. This mistake was retrieved as soon as discovered; but the fact that explanations and reissues could not quell the soldiers' suspicions suggests that the troops were already disturbed by other causes. The Bengal army of nearly 130,000 Indian troops contained about 40,000 Brahmins as well as many Rājputs. The British had accentuated caste consciousness by careful regulations, had allowed discipline to grow lax, and had failed to maintain understanding between British officers and their men. In addition, the General Service Enlistment Act of 1856 required recruits to serve overseas if ordered, a challenge to the castes who composed so much of the Bengal army. To these points may be added the fact that the British garrison in India had been reduced at this time to 23,000 men because of troop withdrawals for the Crimean and Persian wars.

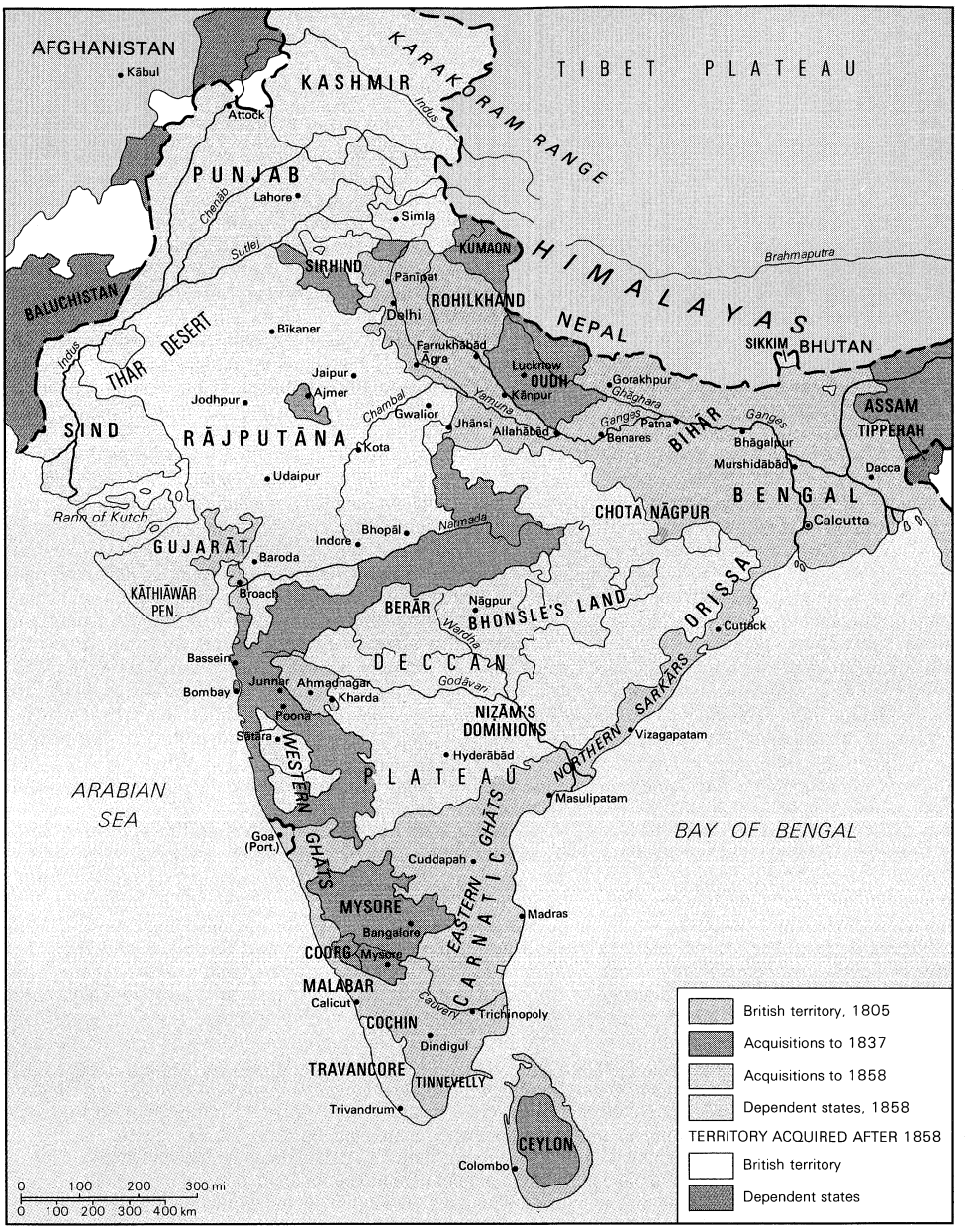
The general factors that turned a military mutiny into a popular revolt can be comprehensively described under the heading of political, economic, social, and cultural westernization. Politically, the princes of India had retired into a sulky seclusion after their final defeat in 1818. But the wars against the Afghans and the Sikhs and then the annexations of Dalhousie alarmed and outraged them. The Muslims had lost the large state of Oudh; the Marāṭhās had lost Nāgpur, Sātāra, and Jhānsi. Further, the British were becoming increasingly hostile toward traditional survivals and contemptuous of anything Indian. There was, therefore, both resentment and unease among the old governing class, fanned in Delhi by the decision to end the Mughal imperial title on Bahādur Shāh's death.

Economically and socially, there had been much dislocation in the landholding class all over northern and western India as a result of British land-revenue settlements, setting group against group. There was thus a suppressed tension in the countryside, ready to break out whenever governmental pressure might be reduced.

Reactions to Western interventions

Then came the Western innovations of the now overconfident British. Their educational policy was a westernizing one, with English instead of Persian as the official language; the practitioners of the old learning felt themselves slighted. Western inventions like the telegraph and railways aroused the conservative prejudice of a conservative race (though Indians crowded the trains when they had them). More disturbing were the interventions, in the name of humanity, in the realm of Hindu custom—*e.g.*, the prohibition of suttee, the campaign against infanticide, the law legalizing remarriage of Hindu widows. Finally, there was the activity of missionaries, by this time widespread. Government was ostentatiously neutral, but an increasing number of its officers were friendly, so that Hindu society was inclined to regard them as eroding Hindu society without openly interfering. In sum, this combination of factors produced, besides the normal tensions endemic in India, an uneasy, fearful, suspicious, resentful frame of mind and a wind of unrest ready to fan the flame of any actual physical outbreak.

The revolt and its aftermath. The dramatic capture of Delhi turned mutiny into full-scale revolt. The whole episode falls into three periods—first came the summer of 1857, when the British, without reinforcements from



British India to 1858.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1965); George Philip & Son Ltd., London

home, fought with their backs to the wall; the second concerned the operations for the relief of Lucknow in the autumn; and the third was the successful campaign of Sir Colin Campbell and Sir Hugh Rose in the first half of 1858. Mopping-up operations followed, lasting until the British capture of Tantia Topi in April 1859.

From Delhi the revolt spread in June to Kānpur and Lucknow. Kānpur surrendered on June 16; but the Lucknow garrison held out in the residency from July 1, in spite of the death of Sir Henry Lawrence on July 4. The campaign then settled down to British attempts to take Delhi and relieve Lucknow. In spite of their apparently desperate situation, the British possessed long-term advantages—they could and did receive reinforcements from Britain; they had, thanks to the resolution of Sir John Lawrence, a firm base in the Punjab, and they had another base in Bengal, where the people were quiet; they had no anxiety in the south and only a little in the west; they had an immense belief in themselves and their civilization, which gave resolution to their initial desperation. The mutineers, on the other hand, lacked good leadership until nearly the end; they had no confidence in themselves and suffered the guilt feelings of rebels without a cause, making them frantic and fearful by turns.

British long-term advantages

In the Punjab were some 10,000 British troops, which made it possible to disarm the Indian regiments; and the recently defeated Sikhs were so hostile to the Muslims that they supported the British against the Mughal restoration in Delhi. A small British army was improvised, which held the ridge before Delhi against greatly superior forces until Sir John Lawrence was able to send a siege train under John Nicholson. With this, and the aid of rebel dissensions, Delhi was stormed and captured by the British on September 20, while the emperor Bahādur Shāh surrendered on promise of his life.

Down-country operations centred on the relief of Lucknow. Setting out from Allahābād on July 7, the British general, Sir Henry Havelock, fought his way through Kānpur to the Lucknow residency on September 25, where he was besieged in his turn. But the back of the rebellion had been broken and time gained for reinforcements to restore British superiority. The next phase saw the relief of the residency (November 16) and the capture of Lucknow by the new commander in chief, Sir Colin Campbell (March 1, 1858). By a campaign in Oudh and Rohilkhand, Campbell cleared the countryside.

The next phase was the central Indian campaign of Sir Hugh Rose. He first defeated the Gwalior contingent and

then, when the rebels Tantia Topi and Lakshmī Bāī, the Rani of Jhānsi, had seized Gwalior, broke up their forces in two more battles. The Rani found a soldier's death, and Tantia became a fugitive. With the British recovery of Gwalior (June 20) the revolt was virtually over.

The restoration of peace was hindered by British cries for vengeance, often leading to indiscriminate reprisals. The treatment of the aged Bahādur Shāh was a disgrace to a civilized nation; also, the whole population of Delhi was driven out into the open, and thousands were killed after perfunctory trials or no trials at all. Order was restored by the firmness of Canning, whose title of "Clemency" was given in derision by angry British merchants in Calcutta, and of Sir John Lawrence in the Punjab. Ferocity led to grave excesses on both sides, distinguishing this war in horror from other wars of the 19th century.

Measures of prevention of such crises in the future naturally began with the army, which was completely reorganized. The ratio of British to Indian troops was fixed at roughly 1:2 instead of 1:5—one British and two Indian battalions were brigaded together so that no sizable station should be without British troops. The effective Indian artillery, except for a few mountain batteries, was abolished, while the Brahmins and Rājputs of Oudh were reduced in favour of other groups. The officers continued to be British, but they were more efficient and more closely linked with their men. The army became an efficient professional body, drawn largely from the northwest and aloof from the national life. (T.G.P.S.)

British imperial power, 1858–1920

CLIMAX OF THE RĀJ

The six decades between the end of the "mutinous" war of 1857–59 and the conclusion of World War I saw both the peak of British imperial power in India and the birth of nationalist agitation against it. The period was haunted by dark memories of the "mutiny" and heartened by dreams of freedom; its first phase was dominated by fear, its later years by hope.

The Government of India Act of 1858. The legacy of the mutiny and great revolt of 1857–59 was of fundamental and far-reaching importance to British India. On August 2, 1858, less than a month after Charles Canning, first viceroy of India, proclaimed the victory of British arms, Parliament passed the Government of India Act, completing at one legislative stroke the transfer of British power over India from the East India Company to the crown. The company's residual powers and responsibilities over Indian affairs were vested in the secretary of state for India, a minister of Great Britain's cabinet, who would be assisted and advised, especially in financial matters, by a Council of India at Whitehall, consisting initially of 15 Englishmen, seven of whom were elected from among the old company's court of directors and eight of whom were appointed by the crown. Though some of Britain's most powerful political leaders became secretaries of state for India in the latter half of the 19th century, including eminent men like Lord Salisbury (1866–67 and 1874–78) and Lord Kimberley (1882–85 and 1886), actual control over the government of India remained secure in the hands of British viceroys in Calcutta and Simla and their "steel-frame" of approximately 1,500 Indian Civil Service (ICS) officials posted "on the spot" throughout British India. Even after the Liberal philosopher-statesman John Morley (later Lord Morley) took charge of the India Office as secretary of state from 1906 to 1910 and fully asserted the singular constitutional powers bequeathed to that post in 1858, the cumbersome machinery of British India's bureaucracy continued to prove itself virtually intractable.

Social policy. On November 1, 1858, Lord Canning announced Queen Victoria's proclamation to "The Princes, Chiefs and Peoples of India," unveiling a new British policy of perpetual support for "native princes" and nonintervention in matters of religious belief or worship within British India, while reiterating the company's long-standing promise of "impartiality" with respect to "race or creed" in the recruitment of officers for the ICS.

Preservation of maharaja autocracy

The first of these new crown policies put an end to Lord Dalhousie's vigorous prewar policy of political unification through princely state annexation. The Dalhousie doctrine of lapse was abandoned, and princes were left free to adopt any heirs they desired, so long as they all swore undying allegiance to the British crown. In 1876, at Prime Minister Benjamin Disraeli's prompting, Queen Victoria added the title empress of India to her regality. British fears of another "mutiny" and consequent determination to bolster Indian states as "natural breakwaters" against any future tidal wave of revolt thus left more than 560 enclaves of autocratic maharaja rule to survive, interspersed throughout British India, for the entire nine decades of crown rule. The new policy of religious nonintervention was born equally out of fear of recurring mutiny, which many Englishmen believed to have been triggered by orthodox Hinduistic and Islāmic reaction against the secularizing inroads of utilitarian positivism and the proselytizing of Christian missionaries. Liberal socioreligious reform, which, since the abolition of the custom of suttee in 1829 had received official British support, came to a halt for more than three decades after the crown took direct control over India. Thirty-five years of official laissez-faire in social reform elapsed between the company's Hindu Widow's Remarriage Act of 1856 and the crown's Age of Consent Act of 1891; yet the latter only timidly raised the age of statutory rape for "consenting" Indian brides from ten years to 12. Not before 1928 would the government of India next find courage to legislate in the sacred realm of marriage, passing at that time its act restraining marriage to a "child bride" under 14. Fearless Indian social reformers such as Mahadev Govind Ranade, Mahārishī (Great Seer) Dhondo Keshav Karve, and Behramji Malabari were obliged to lead the struggle for female emancipation through widow remarriage, female education, and agitation against infant marriage, without official support. But many Britishers in India, official as well as nonofficial, privately encouraged these Hindu and Parsi reformers.The Indian National Social Conference, founded by Ranade in 1897, became the leading unofficial Indian agency for coordinating and advancing the vital socioreligious reformation of Hinduism, which after 1920, thanks to the charisma and reforming zeal of Mahatma (Great Soul) Gandhi, merged with the mainstream of political nationalism of the Indian National Congress.

British racial attitudes

Queen Victoria's promise of racial equality of opportunity in the selection of civil servants for the government of India was never implemented during this era, and the most tragic legacy of 1857–59 was the unbridgeable gulf of racial hatred left by the "mutiny" between British and "native" communities throughout India. Canning, who won the originally opprobrious title Clemency from his Calcutta countrymen for his humane treatment of Indian prisoners of war, was the aristocratic exception to what became a rule of British racial arrogance and often violent abuse of "natives" in the wake of the mutiny, which may have started as a military mutiny but ended as a bitter racial holocaust. The English writer Rudyard Kipling captured the typical attitude of British officials who went to India in this period, as he put it, to "take up the White man's burden." Britishers lived, by and large, throughout the interlude of their Indian service to the crown as super-bureaucrats "Pukka Sāhibs," remaining as aloof as possible from "native contamination" and counting the months till they would be eligible to return home to retire in Great Britain on their ICS pensions. British military cantonments (called camps) were constructed beyond the walls of many of the larger "native" cities in this era. They were built initially for the more effective dispersal of military regiments as part of the general reorganization of British India's armies (Bengal, Bombay, and Madras), which in 1857 had only 45,000 British to 240,000 native troops, but which by 1863 had attained their new "safer mix" of 65,000 English to 140,000 Indian soldiers. British camps soon became suburban residential islands for all Europeans, whose social life was centred around the British Club. After 1869, with the completion of the Suez Canal and the simultaneous introduction of steam transport, reducing the sea passage to India from

almost three months to as many weeks and less, Englishwomen came East with ever greater alacrity, and the British officials they married found it more tempting to return home with their English families during furloughs than to tour India as their predecessors had done. Fewer and fewer Englishmen dared to consort openly with Indian women, unless they were in remote regions where it might be unsafe to take English ladies. Thus, while the intellectual calibre of British recruits to the ICS in this era was, on the average, probably higher than that of servants recruited under the company's earlier "patronage" system, British contacts with Indian society diminished in every respect, and English sympathy for and understanding of Indian life and culture was, for the most part, replaced by suspicion, indifference, and fear. The competitive ICS examinations were theoretically as open to Indian as they were to British subjects of the crown, but since examinations were given only in London to horse-riding male entrants between the ages of 17 and 22 (in 1878 the maximum age was still further reduced to 19), it is hardly surprising to find that by 1869 only one Indian candidate had managed to clear all these hurdles to win admission to the ICS.

Government organization. From 1858 to 1909 the government of India was an increasingly centralized paternal despotism and the world's largest imperial bureaucracy. The Indian Councils Act of 1861 transformed the viceroy's Executive Council into a miniature cabinet run on the portfolio system, with each of the five ordinary members placed in charge of a distinct department of Calcutta's government—home, revenue, military, finance, and law. The military commander in chief sat with this council as an extraordinary member, and when Lord Kitchener took command in 1902 he soon protested against meeting with a military member who was his junior in battle rank but his equal at least on the viceroy's council. Kitchener demanded, therefore, that the military member of council be replaced with a soldier appointed by army headquarters, directly subordinate to the commander in chief. Lord Curzon, who was then viceroy, most vigorously resisted this attempt by Kitchener to subvert Britain's traditional policy of the subordination of military to civil authority in government, and a titanic struggle ensued at Simla, one that resulted in Curzon's resignation and departure from India in 1905. A sixth ordinary member was assigned to the viceroy's Executive Council after 1874, initially to preside over the Department of Public Works, which after 1904 came to be called Commerce and Industry. Though the government of India was by statutory definition the "Governor-General-in-Council" ("governor general" remained the viceroy's alternate title), the viceroy was empowered to overrule his councillors if ever he deemed that necessary. He personally took charge of the Foreign Department, which was mostly occupied with relations with princely states as well as bordering foreign powers. Few viceroys found it necessary to assert their full despotic authority, since the majority of their councillors usually agreed with every position they took, but in 1879 Viceroy Lytton felt obliged to overrule his entire council in order to accommodate demands for the elimination of his government's import duties on British cotton manufactures, despite India's desperate need for revenue in a year of widespread famine and agricultural disorders. From 1854 additional members met with the governor general's Executive Council for legislative purposes, and by the act of 1861 their permissible number was raised to between six and 12, no fewer than half of whom were to be nonofficials. All such legislative councillors were added by appointment of the viceroy, who was empowered to veto any bill passed on to him by this body. But its debates were to be open to a limited public audience, and several of its nonofficial members were Indian nobility and loyal landowners. The legislative council sessions thus served as something of a public-opinion barometer of a very crude sort and the beginnings of an advisory "safety valve" for the government of India, designed to provide the viceroy with early crisis warnings at the minimum possible risk of parliamentary-type opposition. The act of 1892 further expanded this central council's permissible additional membership to 16, ten of whom could be nonofficials, and increased their powers to the extent of allowing them to ask questions of government and formally to criticize the official budget during one day reserved at the very end of each year's legislative session in Calcutta for that purpose. The act, however, was a far cry from what the Indians demanded. They asked for representative institutions and obtained only consultative councils. The Supreme Council still remained, therefore, quite remote from any sort of parliament, but its next giant step in that direction, to be taken in the Indian Councils Act of 1909 (popularly called the Morley-Minto reforms), would bring it much closer to parent British prototypes of responsible, though not as yet representative, government.

Powers of the viceroy

Influence of the viceroys. The viceroyalty of India was one of the more lucrative plums in the crown's patronage orchard, and with the exception of Sir John Lawrence (viceroy from 1864 to 1869), who had been reared in the company's service, went throughout this era to a curious mixture of lesser British nobility, devoid of Indian experience. Lord Canning (governor general and viceroy from 1856 to 1862) was succeeded by James Bruce, 8th earl of Elgin (1862–63), a lackluster viceroy who died after 20 months in India. Lord John Lawrence, whose work in settling the Punjab had established his reputation for firm and efficient administrative leadership, returned to Calcutta as viceroy in 1864, after having served three years on the secretary of state's council in Whitehall. The Punjab and Oudh Tenancy Acts of 1868 were the major achievements of Lawrence's viceroyalty, for, by assuring tenant peasants of their occupancy rights, they helped to restore stability and revive north India's agricultural productivity. Lawrence's tenure, however, was scarred by severe famine in Orissa as well as Rājputāna. Lord Mayo (viceroy from 1869 to 1872) went to India after serving as chief secretary for Ireland and found similarly violent disaffection from British rule, faced, as he was throughout his viceroyalty, with Wahhābī (Muslim puritan movement) Muslim uprisings in Bengal and Kuka Sikh raids in the Punjab. Mayo himself was assassinated by an Afghan prisoner on the Andaman Islands in 1872. Lord Northbrook (viceroy from 1872 to 1876) was sent by Gladstone from the British War Office to succeed Mayo and managed by his policy of inertia to restore calm to India. Disraeli's viceroy, Lord Lytton (1876–80), was a Tory romantic who inaugurated a most costly foreign policy that precipitated the Second Afghan War (1878–80). Lytton's domestic policy was one of laissez-faire for Lancashire cottons and repression of "native" opinion opposed to that policy by the muzzling of India's fourth estate with his unpopular Vernacular Press Act of 1878. The liberal viceroy Lord Ripon (1880–84) was Gladstone's political counterpart in India and soon came to be as widely revered by India's populace as Lytton was reviled. Ripon repealed his predecessor's legislation against the press and sponsored the first official measures designed to introduce to British India representative institutions at the local district and municipal government levels. In 1882 Ripon called for the creation of municipal and district government boards, at least two-thirds of whose members would be elected nonofficials. Though his efforts were sabotaged by a jealous bureaucracy and undermined by Indian lack of enthusiasm and experience in self-government, Lord Ripon deserves an honoured place in Indian history as the first viceroy who laboured to plant the seed of independent rule in British Indian soil. Ripon also struck the first official blow against his government's policy of racism in India by urging his law member of council, Sir Courtney Ilbert, to draft a bill that would allow Indian judges in the service to try cases involving Englishmen as well as Indians. This attempt to remove racial bigotry from British India's legal code by the introduction of the Ilbert Bill in 1883, however, almost ended in a white mutiny in Bengal. The British enclave became so outraged and vituperative in its racist clamour against the proposed measure that the bill could only be enacted in a totally emasculated form, with the added stipulation that any "European British" defendant would have the option of standing trial before a jury at least half of which would be of "European complexion." Though Bengal's

Ripon's attempted reforms

white supremacists won their battle against Ripon and Ilbert, they inadvertently demonstrated to a generation of young Indian nationalists the tactics of political agitation and protest. If a handful of white lawyers, planters, and publicists could by their shrill and persistent demands force the mighty government of India to reverse itself, there was no reason to believe that millions of Indians could not achieve as much. In 1885, just a year after Lord Ripon left India, the Indian National Congress, prior incarnation of the Republic of India's ruling Congress Party, was born.

Economic policy. Economically, this was an era of increased commercial agricultural production, rapidly expanding trade, early industrial development, and severe famine. The total cost of the mutiny of 1857–59, approximately £40,000,000 and equivalent to a normal year's revenue, was charged to India and paid off from increased revenue resources in four years. The major source of government income throughout this period remained the land revenue, which, as a percentage of the agricultural yield of India's soil, continued to be "an annual gamble in monsoon rains." Usually, however, it provided about half of British India's gross annual revenue, or roughly the money needed to support the army. The second most lucrative source of revenue at this time was the government's continued monopoly over the flourishing illicit opium trade to China; the third was the tax on salt, also jealously guarded by the crown as its official monopoly preserve. An individual income tax was introduced for five years to pay off the war deficit, but urban personal income was not added as a regular source of Indian revenue until 1886. Despite continued British adherence to the doctrine of laissez-faire during this period, a 10 percent customs duty was levied in 1860 to help clear the war debt. It was reduced to 7 percent in 1864 and to 5 percent in 1875. In 1879, the cotton import duty was abolished and not reimposed on British imports of piece goods and yarn until 1894, when the value of silver fell so precipitously on the world market that the government of India was forced to take action, even against Lancashire, by adding enough rupees to its revenue to make ends meet.

Early Indian industrialism

Bombay's textile industry had by then developed more than 80 power mills, and the Indian industrialist Jamsetji N. Tata's (1839–1904) huge Empress Mill was in full operation at Nāgpur, competing directly with Lancashire for the vast Indian market. Britain's millowners again demonstrated their power in Calcutta by forcing the government of India to impose an "equalizing" 5 percent excise tax on all India-manufactured cloth, thereby convincing many Indian millowners and capitalists that their best interests would be served by contributing financial support to the Indian National Congress.

Britain's major contribution to India's economic development throughout the era of crown rule was the railroad net that spread so swiftly across the subcontinent after 1858, when there were barely 200 miles of track in all of India. By 1869 more than 5,000 miles of steel track had been completed by British railroad companies, and by 1900 there were 25,000 miles of rail laid. By the start of World War I the total reached 35,000 miles, almost the full growth of India's rail net. Initially, the railroads proved a mixed blessing for most Indians, since by linking India's agricultural, village-based heartland to the British imperial port cities of Bombay, Madras, and Calcutta, they served first of all to accelerate the pace of raw-material extraction from India while speeding up the transition from subsistence food to commercial agricultural production. Middlemen hired by port city Managing Agency Houses rode the trains inland, inducing village headmen to convert large tracts of grain-yielding land to commercial crops. Large sums of silver were offered in payment for raw materials when the British demand was high, as was the case throughout the U.S. Civil War. But when the Civil War ended, restoring raw cotton from the southern United States to Lancashire mills, the Indian market collapsed. Millions of peasants weaned from grain production now found themselves riding the boom-and-bust tiger of a world-market economy. They were unable to convert their commercial agricultural surplus back into food during depression years, and from 1865 through 1900 India experienced the most severe series of protracted famines in its entire history. Though the population of the subcontinent increased dramatically from about 200,000,000 in 1872 (the year of the first census) to over 300,000,000 in 1920, the population actually declined by several millions between 1895 and 1905. The spread of railroads also accelerated the destruction of India's indigenous handicraft industries, for trains filled with cheap competitive manufactured goods shipped from England now rushed to inland towns for distribution to villages, underselling the rougher produce of Indian craftsmen. Entire handicraft villages thus lost their traditional markets of neighbouring agricultural villagers, and craftsmen were forced to abandon their looms and spinning wheels and return to the soil for their livelihood. By the end of the 19th century a larger percentage of India's population (approximately 90 percent) depended directly upon agriculture for support than at the century's start, and the pressure of population upon arable land increased throughout this period. Railroads also provided the military with swift and relatively assured access to all parts of the country in the event of emergency and were eventually used to transport grain for famine relief as well.

Effects of British imports

The rich coalfields of Bihār and Orissa began to be mined in this period to help power the imported British locomotives, and coal production jumped from 500,000 tons in 1868 to over 6,000,000 in 1900 and more than 20,000,000 by 1920. Coal was used for iron smelting in India as early as 1875, but the Tata Iron and Steel Company, which received no government aid, did not start production until 1911, launching India's steel industry in Bihār. Tata grew rapidly after World War I, and by World War II had become the largest single steel complex in the British Commonwealth, its city of location named Jamshedpur, after its founder. The jute textile industry, Bengal's counterpart to Bombay's cotton industry, developed in the wake of the Crimean War (1853–56), which, by cutting off Russia's supply of raw hemp to the jute mills of Scotland, stimulated the export of raw jute from Calcutta to Dundee. In 1863 there were only two jute mills in Bengal, but by 1882 there were 20, employing over 20,000 workers, and by 1908 British India's jute production was greater than Dundee's. The most important plantation industries of this era were tea, indigo, and coffee. British tea plantations were started in North India's Assam Hills in the 1850s and in South India's Nīlgiri Hills some 20 years later. By 1871 there were over 300 tea plantations, covering more than 30,000 cultivated acres (12,000 hectares) and producing over 6,000,000 pounds of tea. By 1900 India's tea crop sufficed to export 137,000,000 pounds to Britain, displacing the tea of China in London. As cheaper Indian tea ousted its Chinese rival from British markets, the Chinese planted opium in land formerly devoted to tea, and during the first decade of the 20th century China's demand for Indian opium dropped so sharply that by the end of World War I, when the League of Nations called upon Britain to stop the opium trade, it had virtually stopped itself. The flourishing indigo industry of Bengal and Bihār was threatened with extinction during the "Blue Mutiny" of 1859–60, but India continued to export indigo to European markets until the end of the 19th century, when synthetic dyes made that natural product obsolete. Coffee plantations flourished in South India from 1860 to 1879, after which disease blighted the crop and sent Indian coffee into a decade of decline. During much of the last decade of the 19th and the first decade of the 20th century, India's economy was ravaged by famine and plague (the bubonic plague entered Bombay in 1896), and it was not until World War I that further strides were made in British India's industrial development and overall economic productivity.

FOREIGN POLICY

The northwest frontier. British India expanded beyond its company borders both to the northwest and the northeast during this initial phase of crown rule. The turbulent tribal frontier to the northwest remained a continuing source of harassment to settled British rule, and Pathan raiders served as a constant lure and justification

Action against the Pathans

to champions of the "forward school" of imperialism in Calcutta, Simla, and Whitehall. Russian expansion into Central Asia in the 1860s provided even greater anxiety and incentive to British proconsuls in India, as well as at the Foreign Office in London, to advance the frontier of the Indian empire beyond the Hindu Kush and, indeed, up to Afghanistan's own northern border along the Oxus. Lord Canning, however, was far too preoccupied with trying to restore tranquillity within India to consider embarking upon anything more ambitious than the northwest-frontier punitive-expedition policy (commonly called "butcher and bolt"), which was generally regarded as the simplest, cheapest method of "pacifying" the Pathans. As viceroy, Lawrence continued the same border-pacification policy and resolutely refused to be pushed or lured into the ever-simmering caldron of Afghan politics. In 1863, when the popular old *amīr*, Dōst Moḥammad Khān, died, Lawrence wisely refrained from attempting to name his successor, leaving the Dōst's 16 sons to fight their own fratricidal battles until 1868, when Shīr ʿAlī finally emerged victorious. Lawrence then recognized and subsidized the new *amīr*, and Mayo met to confer with him at Ambala in 1869, reaffirming Anglo-Afghan friendship but resisting all requests for more permanent and practical support by the *amīr* for his still precarious regime.

The Second Afghan War. Russia's glacial advance into Turkistan and Samarkand sufficiently alarmed Disraeli and Salisbury that by 1874, when they came to power in London, they pressed the government of India to pursue a more vigorous interventionist line with Kābul. Northbrook resisted all such cabinet promptings to reverse Lawrence's noninterventionist policy and return to the militant posture of the First Afghan War era, resigning his office rather than accept orders from ministers whose diplomatic judgment he believed to be disastrously distorted by Russophobia. Lytton, however, who succeeded him, was more than eager to act as his prime minister desired, and, soon after he reached Calcutta, he rushed to notify Shīr ʿAlī that he was sending a "mission" to Kābul. When the *amīr* refused Lytton permission to enter his hermit kingdom, the viceroy bellicosely declaimed that Afghanistan was but "an earthen pipkin between two metal pots." In 1878, Russia's General Stolyetov was admitted to Kābul, while Lytton's envoy, Sir Neville Chamberlain, was turned back at the border by Afghan troops. The Viceroy decided to crush his neighbouring "pipkin" and launched the Second Afghan War on November 21, 1878, with a British invasion over the high passes. Shīr ʿAlī fled his capital and country, dying in exile early in 1879. The British army occupied Kābul, as it had in the first war, and a treaty signed at Gandamak on May 26, 1879, was concluded with the former *amīr*'s son, Yaʿqūb Khān. Yaʿqūb Khān promised to admit a British resident to his Kābul court and to turn over the direction of Afghan foreign relations to him in return for his British promise of money and arms to support and protect the new *amīr* from his many enemies. A telegraph line was to be strung between Kurram and Kābul, and the former border district of Afghanistan was to be occupied by British troops, as would also be Pishīn, Sibi, and the Khyber and Mishmi passes. In July 1879 Sir Louis Cavagnari, the British resident, reached Kābul and on September 3, 1879, was assassinated there. British troops trudged back over the passes to Kābul, removing Yaʿqūb to India; his throne was left vacant until July 1880, when ʿAbdor Raḥmān Khān, nephew of Shīrʿalī, became *amīr*. The Second Afghan War, which thus proved to be a Pyrrhic victory, became the leading issue in the British general elections of 1880, after which Disraeli's Tory cabinet was replaced by Gladstone's Liberal government. Lord Ripon tactfully concluded the negotiations with ʿAbdor Raḥmān, by which Britain managed slowly to extricate itself from the military morass into which Lytton had rushed with all the fervour and lack of foresight of a schoolboy. The Khyber and Kurram were left to the Afghans, as was Kābul, and, though British troops remained in Qandahār until 1881, they were withdrawn thereafter and continued to occupy only the border towns of Sibi and Pishīn. ʿAbdor Raḥmān, one of the shrewdest statesmen in Afghan history, remained secure on his throne until his death in 1901. Though Russia defeated the Turkmen and advanced toward Merv in 1881, British India survived that new threat to its security without another Afghan war. Ripon urged instead that an Afghan diplomatic settlement be reached between Britain and Russia. This was accomplished by the Anglo-Russian agreement on Afghanistan's northern border reached on July 22, 1887. Panjdeh was abandoned to Russia and the northwestern corner of Afghanistan fixed at Zulfiqar.

Settlement with Russia

Creation of the North-West Frontier Province. Lord Lansdowne, viceroy from 1888 to 1894, sought to reassert a more forward policy in Afghanistan, following the advice of his military commander in chief, Lord Roberts, who had served as field commander in the Second Afghan War. Sir Mortimer Durand, the government of India's foreign secretary, was sent by Lansdowne on a mission to Kābul in 1893 to open negotiations to delimit the Indo-Afghan border. Demarcation of this "Durand Line" was completed in 1896, adding the tribal territory of the Afrīdīs, Maḥsūds, Wazīrīs, and Swātīs to the domain of British India, as well as the chieftainships of Chitral and Gilgit. The 9th earl of Elgin (1894–99), Lansdowne's successor, devoted most of his viceregal tenure to sending British Indian armies on punitive expeditions along this new frontier. The frontier rising of all Pathan tribes (1897) took no less than 35,000 troops to put down. The fiercely independent tribal peoples of India's northwest-frontier hill country continued to battle in protracted guerrilla campaigns against British arms. Lord Curzon recognized the impracticality of trying to administer the turbulent frontier region as part of the large Punjab province and in 1901 created a new North-West Frontier Province containing some 40,000 square miles (about 100,000 square kilometres) of trans-Sutlej and tribal borderland territory under a British chief commissioner responsible directly to the viceroy. By instituting a policy of regular payments to frontier tribes, the new province reduced border conflicts, though for the next decade British troops continued to fight against Maḥsūds, Wazīrīs, and Zakka Khel Afrīdīs.

The Anglo-Russian Convention of 1907 was the first comprehensive attempt to define central Asian "spheres of influence" effectively enjoyed by the two great European rivals. The Afghan portion of this convention placed Afghanistan outside Russia's sphere but within Britain's, as far as foreign relations were concerned. Though the new *amīr*, Ḥabībollāh Khān (1901–19), refused to sign the convention, this agreement served to ensure Anglo-Afghan peace until after World War I. Following Ḥabībollāh's assassination in February 1919, however, his son and successor as *amīr*, Amānollāh Khān, issued a call to his people to rise up in holy war (*jihād*) against the British. The Third Afghan War, thus started in May 1919, never posed a serious challenge to British India, however, for after British warplanes were used to bomb Kābul and Qandahār the *amīr* appealed for peace at the end of May, and a preliminary treaty ending all hostilities was signed at Rawalpindi on August 8, 1919. The formal treaty ending the Third Afghan War was concluded on November 22, 1921, establishing Anglo-Afghan relations on the basis of "internal and external independence."

The Third Afghan War

The incorporation of Burma. To the northeast, British India's conquest of Burma was completed during this period. The Second Anglo-Burmese War (1852) had left the kingdom of Ava (Upper Burma) independent of British India, under the rule of King Mindon (1853–78). Mindon built his capital at Mandalay and welcomed British steamers, bringing British residents as well as private traders up the Irrawaddy from Rangoon to his peaceful kingdom. In 1871 Mindon convened the Fifth Buddhist Council in Mandalay, the first such council called in some 1,900 years. Mindon was succeeded by a junior son, Thibaw, who in February 1879 celebrated his ascendancy to the throne of Ava by having 80 siblings massacred. King Thibaw refused to renew his father's treaty agreements with Britain, turning instead to seek commercial relations with the French, who were then advancing toward his kingdom

Burmese overtures to the French

from their base in Indo-China. Thibaw sent envoys to Paris, and in January 1885 the French signed a treaty of trade with the kingdom of Ava, dispatching a French consul to Mandalay. This envoy hoped to establish a French bank in Upper Burma to finance the construction of a railway and the general commercial development of the kingdom, but his plans were not permitted to bear fruit. The viceroy, Lord Dufferin, impatient with Thibaw for delaying a treaty agreement with British India, goaded to action by British traders in Rangoon, and provoked by fears of French intervention in Britain's "sphere," sent an expedition of some 10,000 troops up the Irrawaddy in November 1885. The Third Anglo-Burmese War ended in less than a month with the loss of hardly 20 lives, and on January 1, 1886, Upper Burma, a kingdom larger than Britain and with a population of some 4,000,000, was annexed by proclamation to British India. As in Afghanistan, Britain's military difficulties in Burma began only after its initial conquest, and, though Burmese guerrilla warfare may have been less terrifyingly effective than the Afghan variety, imperial order was only gradually achieved and even by 1890 extended hardly across one-half of the conquered territory. The greater portion of Upper Burma, the Chin Hills and Shan States, was never totally subdued or pacified by the British. From 1886 to 1897 the new province of united Burma was administered by a chief commissioner and after the later date by a lieutenant governor appointed by the government of India.

INDIAN NATIONALISM AND THE BRITISH RESPONSE

Origins of the nationalist movement. The Indian National Congress held its first meeting in December 1885 in Bombay city, while British Indian troops were still fighting in Upper Burma. Just as the British Indian empire attained its outermost limits of expansion, the institutional seed of the larger of its two national successors was thus sown. Provincial roots of Indian nationalism, however, may be traced to the beginning of the era of crown rule in Bombay, Bengal, and Madras. Nationalism emerged in 19th-century British India both in emulation of and as a reaction against the consolidation of British rule and the spread of Western civilization. There were, moreover, two turbulent national mainstreams flowing beneath the deceptively placid official surface of British administration: the larger, headed by the Indian National Congress, which led eventually to the birth of India, and the smaller Muslim one, which acquired its organizational skeleton with the founding of the Muslim League in 1906 and led to the creation of Pakistan.

Many English-educated young Indians of the postmutiny period emulated their British mentors by seeking employment in the ICS, the legal services, journalism, and education. The universities of Bombay, Bengal, and Madras had been founded in 1857 as the capstone of the East India Company's modest policy of selectively fostering the introduction of English education in India. At the beginning of crown rule, the first graduates of these universities, reared on the works and ideas of Jeremy Bentham, John Stuart Mill, and Thomas Macaulay, sought positions that would help them improve themselves and society at the same time, convinced that with the education they had received and the proper apprenticeship of hard work they would eventually inherit the machinery of British Indian government. Few Indians, however, were admitted to the ICS, and among the first handful who were, one of the brightest, Surendranath Banerjea (1848–1925), was dismissed dishonourably at the earliest pretext and turned from loyal participation within the government to active nationalist agitation against it. Banerjea became a Calcutta college teacher and then editor of *The Bengalee* and founder of the Indian Association in Calcutta. In 1883 he convened the first Indian National Conference in Bengal, anticipating by two years the birth of the Congress on the opposite side of India. After the first partition of Bengal in 1905, Banerjea attained nationwide fame as a leader of the *swadeshi* ("of our own country," or India-made goods) and boycott (of British manufactured goods) movements. In Bombay, young leaders of new India were also busily at work in the 1870s, establishing provincial political associations such as the Poona Sārvajanik Sabhā (Poona Public Society) founded by Mahadev G. Ranade, who had graduated first in his class from Bombay University's first bachelor of arts class in 1862. Ranade found employment in the educational department in Bombay, taught at Elphinstone College, edited the *Indu Prakāsh,* helped start the Hindu reformist Prārthanā Samāj (Prayer Society) in Bombay, wrote historical and other essays, and became a barrister, eventually being appointed to the bench of Bombay's high court. Ranade was one of the early leaders of India's emulative school of nationalism, as was his brilliant disciple Gopal Krishna Gokhale (1866–1915). Gokhale, revered by Mahatma Gandhi as his political guru (preceptor), taught at Fergusson College and was an editor and social reformer as well as a nationalist political leader. He was elected president of the Congress in 1905. Moderation and reform were the keynotes of Gokhale's life, and by his use of reasoned argument, patient labour, and unflagging faith in the ultimate equity of British liberalism, he achieved more for India, working within the system of British rule, than did any of his contemporaries. Bal Gangadhar Tilak (1856–1920), Gokhale's Poona colleague at Fergusson College, was the leader of Indian nationalism's revolutionary reaction against British rule. Tilak was Poona's most popular Marāṭhī journalist, whose vernacular newspaper, *Kesari* ("Lion"), became the leading literary thorn in the side of the British. The Lokamānya (Revered by the People), as Tilak came to be called after he was jailed for seditious writings in 1897, looked to orthodox Hinduism and Marāṭhā history as his twin sources of nationalist inspiration. Tilak called upon his compatriots to take keener interest and pride in the religious, cultural, martial, and political glories of pre-British Hindu India, rather than slavishly focussing their attention on foreign learning and emulating the ideas and attitudes of British white Christian oppressors. He helped found and publicize the popular Ganapati and Śivājī festivals in the 1890s in Poona, former capital of the Marāṭhā Hindu glory. Tilak had no faith in British justice, and his life was devoted primarily to agitation aimed at ousting the British from India by any means and restoring *swarāj* ("self-rule" or independence) to India's people. The militant nature of Tilak's agitation helped stimulate youthful terrorists such as the Chapekar brothers of Poona, who assassinated two British officials on the night of Queen Victoria's Diamond Jubilee (June 22, 1897). The orthodox Hindu character of Tilak's revolutionary revival, which mellowed considerably in the latter part of his political career, while bringing many non-English-educated Hindus into the nationalist movement, in a small measure alienated India's Muslim minority and exacerbated communal tensions and conflict.

The viceroyalties of Lytton and Ripon prepared the soil of British India for nationalism, the former by internal measures of repression and the futility of an external policy of aggression, the latter indirectly as a result of the European community's rejection of his liberal humanitarian legislation. One of the key men who helped arrange the first meeting of Congress was a retired British official, Allan Octavian Hume (1829–1912), Ripon's radical confidant. Hume, a mystic reformer and ornithologist, lived in Simla after retiring from the ICS in 1882, studying birds and Theosophy. Hume had joined the Theosophical Society in 1881, as had many young Indians, who found in Theosophy a movement most flattering to Indian civilization. Mme Blavatsky, cofounder of the Theosophical Society, went to India in 1879 to sit at the feet of Swāmi Dayananda Sarasvati (1824–83), the "Luther of Hinduism," whose "back to the Vedas" reformist Hindu society, the Ārya Samāj, was founded in Bombay in 1875. Dayananda called upon Hindus to reject the "corrupting" medieval excrescences of their faith, including idolatry, the caste system, and infant marriage, and to return to the original purity of Vedic life and thought. The Swāmi insisted that post-Vedic changes in Hindu society had led only to weakness and disunity, destroying India's capacity to resist foreign invasion and subjugation. His reformist society was to take root most firmly in the Punjab at the start of the 20th century, and it became that province's leading nationalist organization. Mme Blavatsky soon left

Early figures in the nationalist movement

Influence of Theosophists

Dayananda and established her own "Samāj," whose Indian headquarters were outside Madras city, at Adyar. Annie Besant (1847–1933), the Theosophical Society's most famous leader, succeeded Mme Blavatsky and became the first and only Englishwoman to serve as president of the Indian National Congress (1917).

The early Congress movement. The first Congress session was convened in Bombay city on December 28, 1885. Womesh C. Bonnerjee, a Calcutta barrister, presided over 73 representatives, as well as ten more unofficial delegates, from every province of British India. Fifty-four of the delegates were Hindu, and only two were Muslim, the remainder being mostly Parsi and Jaina. Practically all the Hindu delegates were Brahmins. All of them spoke English. More than half were lawyers, the remainder consisting of journalists, businessmen, landowners, and professors. Such was the first gathering of the new India, an emerging elite of middle class intellectuals devoted to peaceful political action and protest on behalf of their nation in the making.

The first demands of Congress

On its last day Congress passed resolutions that embodied the political and economic demands of its members, serving thereafter as public petitions to government for the redress of grievances, while helping to educate the rest of India's population in the meaning of the new national program. These initial resolutions called for the full restitution to Britain's House of Commons of the authority over Indian administrative affairs that the government of India had autocratically usurped; for the abolition of the reactionary council in Whitehall; for the addition of elected nonofficial representatives to the Supreme and provincial legislative councils; and for real equality of opportunity for Indians to enter the ICS by the immediate introduction of simultaneous examinations in India and England. The Congress' economic demands started with a call for the reduction of "Home charges," which included the entire India Office budget and the pensions of officials living in England in retirement. Dadabhai Naoroji (1825–1917), the "grand old man" of Congress who served three times as its president, was the leading exponent of the popular economic "drain" argument, which offered theoretical economic support to nationalist politics by insisting that India's poverty was the product of British exploitation and annual plunder of gold and silver as well as raw material resources. Other resolutions called for the reduction of military expenditure, condemned the Anglo-Burmese War, demanded retrenchment, and urged reimposition of import duties on British manufactures. Hume, to whom goes the credit of organizing the Indian National Congress, attended the first Congress as the only English delegate, while Sir William Wedderburn (1838–1918), Gokhale's closest English adviser and himself later elected twice to serve as president of Congress, and Professor William Wordsworth, principal of Elphinstone College, both appeared as observers. Most Englishmen in India, however, either ignored Congress and its resolutions as the action and demands of a "microscopic minority" of India's diverse millions or considered them the rantings of disloyal extremists. Despite this combination of official disdain and hostility, however, Congress quickly won substantial Indian support and within two years had grown to number more than 600 delegates. In 1888, when Dufferin, on the eve of his departure from India, dismissed it as "microscopic," it mustered 1,248 delegates at its annual meeting.

Curzon's policies. George Nathaniel Curzon, 1st marquess of Kedleston and viceroy from 1899 to 1905, hoped in 1900 that Congress was "tottering to its fall," confessing that one of his great ambitions was to "assist it to a peaceful demise." By the time his viceroyalty ended, however, Lord Curzon had inadvertently infused Congress with unprecedented popularity and militant vitality. Curzon's untiring capacity for administrative labour, combined with his insensitivity to criticism—especially from "natives"—made him at once the most efficient and least popular viceroy in Indian history. Curzon's reign marks the peak of imperial centralization of authority by the government of India and of indifference to public opinion. In foreign policy, Curzon devoted himself to seeking ways of excluding Russian influence from the Persian Gulf and Seistan, Kābul, and Lhasa. In a personal display of imperial pomp, Curzon toured the Persian Gulf in a British warship in 1903, the same year in which he held his extravagant *darbār* (festival and reception of native princes) in Delhi. In 1904 he launched a mission to Tibet, which soon turned into a military massacre and invasion of the Chumbi Valley and Gyantse. It was followed by British Indian armed occupation of that remote region beyond the Himalayan passes. The Anglo-Russian Convention of 1907 stipulated the withdrawal of British troops from Tibet, and this was carried out in 1908. Curzonian imperiousness and impatience in negotiating with Ḥabībollāh, the new *amīr* of Afghanistan, alarmed Prime Minister Balfour's cabinet more than any threat of Russian influence in Kābul did; for, having just extricated Britain from the Boer War (1899–1902) in South Africa, no British statesmen dared contemplate the prospect of a third Afghan War. In 1905 Sir Louis Dane, who was ordered to remain in Kābul despite Curzon's impatience, successfully completed negotiations of a treaty with Afghanistan, which Curzon considered humiliating. That same year, however, the spectre of Russian power over Asia was diminished considerably by Japan's unexpected victory in the Russo-Japanese War (1904–05).

Curzon's principal domestic concern was to improve the efficiency of Indian administration. To this end he created a new province (the North-West Frontier), a new Department at Simla (Commerce and Industry), reformed the Indian police, reorganized the governing bodies of Calcutta University, and partitioned the province of Bengal. Curzon failed to realize in pursuing the latter two administrative "reforms" that human sympathy and understanding are important components of any government's drive toward achieving efficiency. His trenchant speech to the graduates of Calcutta University in support of the Universities Act of 1904 did more to alienate Bengali intellectuals from his government than any official statement made since the "mutiny," and the actual partition of Bengal in 1905 brought that province to the brink of open rebellion. Bengali Hindus viewed the partition, which created a Muslim-majority province of Eastern Bengal and Assam and left a truncated Bihārī and Oṛiyā-speaking Hindu-majority province in West Bengal, as clear evidence of Curzon's policy of divide and rule. Everyone conceded that, with some 85,000,000, Bengal was much too large for a single province, but why, Bengalis asked, was it necessary to divide the territory in such a way as to strip Bengali Hindus of a majority in either new province? Surely, their leaders argued, this was proof of the government's vindictive antipathy toward the outspoken *bhadralok* ("respectable people") intellectuals. Partition, they argued, was really designed to remove them from positions of provincial power. Rallying to the defense of the "mother province," mother-goddess-worshipping Bengali Hindus responded with religious zeal to the country-wide agitation and to the mass protest meetings called by Congreess leaders before and in the wake of the partition proclamation of October 16, 1905.

The partition of Bengal (1905)

Reaction to the partition of Bengal. Millions of Indians, hitherto unmoved by the austere language of Congress speeches and resolutions, now joined the nationalist movement, swept along by the tide of emotion that rushed over Bengal. "Bande Mātaram" ("Hail to the Mother") became the Congress national anthem, its words taken from a popular Bengali novel by Bankim Chandra Chatterjee, *Ānandamaṭh* ("Abbey of Bliss"), its music written by Bengal's poet Rabindranath Tagore. When Bengalis realized that all of their appeals, the countless petitions they had signed, and the river of rhetoric they had poured from platform and press protesting against partition were of no avail in dissuading government from its ill-conceived plan, they turned to other tactics. First of all, they launched a boycott of British-made goods, dramatizing their determination to live without foreign cloth by igniting huge bonfires of Lancashire goods throughout Bengal. At the same time, they vowed to use only domestic (*swadeshi*) cottons and other goods made in India. Simple India-made saris soon became far more fashionable in Calcutta *bhadralok* society than the finest Lancashire silks. The boycott and *swadeshi* movement spread across India to Poona, Bom-

bay, and Madras and became the major economic plank in Congress' platform. The first partition of Bengal thus served to stimulate indigenous Indian industry, from cloth to matches, from glass products to iron and steel.

Demands for Indian education

Increased demands for national education also swiftly followed partition, and Curzon's outspoken contempt for Indian intellectuals convinced many of them to rely upon themselves rather than depend on government. Pioneers in the development of totally Indian institutions of higher learning in western India were the so-called Indian Jesuits of the Deccan in the 1880s, Vishnu Krishnashastri Chiplunkar (1850–82), Gopal Ganesh Agarkar (1856–95), Tilak, and Gokhale. In Bengal the indigenous education movement spread as swiftly as *swadeshi.* Nor were Bengal and Bombay the only centres of the movement, for in 1910 Pandit Madan Mohan Malaviya founded the private Banaras Hindu University in Vārānāsi, the revered centre of Hindu culture in the United Provinces. *Swarāj* ("self-rule") was the last major demand to be added to Congress' platform as a result of Bengal's partition. The sacred *mantra* ("holy utterance") of *swarāj* was voiced for the first time at a Congress session in 1906, in the presidential address of Dadabhai Naoroji at Calcutta. The demand for self-rule swiftly swept the country, and Tilak soon claimed it as his "birthright."

Nationalism in the Muslim community. While Congress was calling for *swarāj* in Calcutta, the Muslim League held its first meeting in Dacca (now Dhākā), capital of the newly created Muslim-majority province of Eastern Bengal and Assam. Though the Muslim quarter of India's population lagged behind the Hindu majority in uniting to articulate nationalist political demands, Islām had, since the founding of the Delhi sultanate in 1206, provided Indian Muslims with sufficient doctrinal mortar to unite them as a separate religious community. The era of effective Mughal rule (*c.* 1556–1707), moreover, gave India's Muslims a sense of martial and administrative superiority to, as well as separation from, the Hindu majority. In 1857 the last of the Mughal emperors had served as a rallying symbol for many "mutineers," and in the wake of the "mutiny" most Britishers placed the burden of blame for its inception upon the Muslim community. Sir Sayyid Ahmad Khan (1817–98), India's greatest 19th-century Muslim leader, succeeded, in his "Causes of the Indian Revolt" (1873), in convincing many British officials that Hindus, rather than Muslims, were primarily to blame for the "mutiny." Sir Sayyid had entered the company's service in 1838 and was the leader of Muslim India's emulative mainstream of political reform. He visited Oxford in 1874 and returned to found the Anglo-Muhammadan Oriental College (now Alīgarh Muslim University) at Alīgarh in 1875. It was India's first centre of Islāmic and Western higher education, with instruction given in English and modelled after Oxford. Alīgarh became the intellectual cradle of the Muslim League and Pakistan. Sayyid Mahdi Ali, popularly known by his title Mohsin al-Mulk (1837–1907), had succeeded Sayyid Ahmad as leader and convened the deputation of some 36 Muslim leaders, headed by the Aga Khan III, that called upon Lord Minto (viceroy from 1905 to 1910) to articulate the special national interests of India's Muslim community (1906). Minto promised that any reforms his government would enact (the Indian Councils Act of 1909 was then being fashioned) would surely safeguard the separate interests of the Muslim community. Separate Muslim electorates, formally inaugurated by the act of 1909, were thus vouchsafed by viceregal fiat in 1906. This encouraged the Aga Khan's deputation, so that it issued an expanded call to the first meeting of the Muslim League, which convened in December 1906 at Dacca. The All-India Muslim League was formed as a "Political Association . . . to protect and advance the political rights and interests of Mussalmans of India." Other resolutions moved at its first meeting expressed Muslim "loyalty to the British government," as well as approval of and support for the Bengal partition and condemnation of the boycott movement.

Muslim demands

Reforms of the British Liberals. In Great Britain the Liberal Party's electoral victory of 1906 marked the dawn of a new era of reforms for British India. Hampered though he was by the viceroy, Lord Minto, the new head of the India Office, John Morley, was able to introduce several important innovations into the legislative as well as the administrative machinery of the British India government. First of all he acted to implement Queen Victoria's promise of racial equality of opportunity, which since 1858 had served only to assure Indian nationalists of British hypocrisy. He appointed two Indian members to his council at Whitehall: one a Muslim, Sayyid Husain Bilgrami, who had taken an active role in the founding of the Muslim League; the other a Hindu, Krishna G. Gupta, the senior Indian in the ICS. Morley also persuaded Minto to appoint the first Indian member to the viceroy's Executive Council, Satyendra P. Sinha (1864–1928), in 1909. Sinha (later Lord Sinha) had been admitted to the Bar at Lincoln's Inn in 1886 and was advocate general of Bengal before his appointment as the viceroy's law member, a position he felt obliged to resign in 1910. He was elected president of Congress in 1915 and became parliamentary undersecretary of state for India in 1919 and governor of Bihār and Orissa in 1920. The constitutional precedent established by Morley in adding Indians to the highest councils of Whitehall and Calcutta was never thereafter reversed, and in 1910 Sinha was succeeded by Sayyid Ali Imam, a Muslim barrister who served on the Supreme Legislative Council before moving to the Executive Council. Morley's major reform scheme, the Indian Councils Act of 1909, directly introduced the elective principle to Indian legislative council membership, and, though the initial electorate was a miniscule minority of Indians enfranchised by property ownership and education, in 1910 some 135 elected Indian representatives took their seats as members of legislative councils throughout British India. The act of 1909 increased the maximum additional membership of the Supreme Council from 16 (to which it had been raised by the Councils Act of 1892) to 60. In the provinces of Bombay, Bengal, and Madras, legislative councils had been created in 1861. Their permissible total membership had been raised to 20 by Lansdowne's Act of 1892, and this was increased to 50 in 1909, a majority of whom were to be nonofficials. The act of 1909 also permitted the addition of 50 members to the legislative councils of the United Provinces and Eastern Bengal and Assam and of up to 30 in all other provinces. Abolishing the official majorities of provincial legislatures, Morley followed the advice of Gokhale and such other liberal Congress leaders as Romesh Chunder Dutt (1848–1909), overriding the bitter opposition of the ICS as well as his own viceroy and council. True Liberal that he was, Morley believed that the only justification for British rule over India was to bequeath to the government of India England's greatest political institution, parliamentary government. Minto and his officials in Calcutta and Simla did succeed in watering down the reforms as much as possible by writing stringent regulations for their implementation and insisting upon the retention by the executive of veto power over all legislation. Elected members of the new councils were empowered, nevertheless, to engage in spontaneous supplementary questioning as well as formal debate with the executive concerning the annual budget, debate on which would no longer be limited to the last day of each year's legislative session but would extend over many days. Members were also permitted to introduce legislative proposals of their own. Gokhale took immediate advantage of these vital new parliamentary procedures by introducing his Elementary Education Bill into the Supreme Legislative Council in 1910. That measure called for the inauguration of a system of free and compulsory elementary education throughout British India. Although defeated, it was brought back again and again by Gokhale, who used the platform of the government's highest council of state as a sounding board for nationalist demands. Before the act of 1909, as Gokhale told his fellow congressmen in Madras that year, Indian nationalists had been engaged in agitation "from outside," but "from now we shall be engaged in what might be called responsible association with the administration." This most important forward step in the evolution of Indian self-government was taken as a result

Beginnings of parliamentary practice

of the full cooperation of liberal Englishmen and Indian nationalists of the moderate emulative wing of Congress, which by this time had split apart.

Moderate and militant nationalism. The Congress meeting of 1907 never came to order long enough to hear the presidential address of Rash Behari Ghose, the candidate of conservative leaders. The division of Congress reflected broad tactical as well as presidential-candidate differences between the liberal evolutionary and militant revolutionary wings of the national organization. Tilak's New Party militants wanted to extend the boycott movement to the British government, and the cult of the bomb as well as the gun had become popular throughout Bengal and Mahārāshtra. Moderate Congress leaders, who cautioned compatriots against "extreme" speeches that might encourage violence, were themselves attacked as traitors to the motherland. The split left Congress divided for nine years into the influential moderate Conventionists of Gokhale's party and the revolutionary revivalists of Tilak's, who had captured the allegiance of India's youth. Tilak was, however, deported to Mandalay prison for sedition in 1908 and kept there until 1914.

From 1908 to 1910 Bengali terrorism reached its peak of frenzy, as did official repression on the part of Minto's government. At the end of 1910 Lord Hardinge of Penshurst succeeded Minto as viceroy (1910–16) and recommended the reunification of Bengal and the creation of a separate province of Bihār and Orissa. On December 12, 1911, King George V announced the revocation of Curzon's partition of Bengal at his Coronation Darbār in Delhi, which was to be the new capital of British India. By shifting the capital from Calcutta to Delhi, former capital of Mughal glory, the British hoped to placate Bengal's Muslim minority, even as the reunification of their motherland was designed to mollify Bengali Hindus. Eastern Bengal and Assam was dissolved, and Muslims cried bitterly "no bombs, no boons." The following year, in 1912, as he was about to enter Delhi in his regal state procession atop an elephant, Lord Hardinge was wounded by a bomb thrown by a would-be assassin, who escaped.

Reunification of Bengal

The reunification of Bengal had come too late to restore tranquillity to British India. Edwin Montagu, Morley's protégé at the India Office, who served as parliamentary under secretary of state for India from 1910 to 1914 and as secretary of state from 1917 to 1922, announced in 1912 that the goal of British policy toward India would have to be to meet the just demands of Indians for a greater share in government. Britain was finally awakening to the urgency and intensity of India's political demands on the eve of World War I, when more compelling problems of European diplomacy pre-empted Whitehall's attention.

WORLD WAR I AND ITS AFTERMATH

India's contribution to the war effort. The impact of World War I on India was as pervasive as that of the mutiny of 1857–59 had been. India's contributions to the war were extensive and significant, yet the war's contributions to change within British India were far greater. India's initial response to Lord Hardinge's declaration of his government's entry into World War I in August 1914 was, for the most part, enthusiastically supportive. The native princes volunteered their men, money, and personal service, while Congress leaders, from Tilak, who had just been released from Mandalay and had wired the King-Emperor vowing his patriotic support, to Gandhi, who toured Indian villages urging peasants to join the British army, were allied in backing the war effort. Only India's Muslims, whose doctrinal allegiance to the Ottoman caliph had to be weighed against their temporal devotion to British rule, seemed hopelessly ambivalent from the war's inception. India's Congress support was primarily offered on the assumption that England would repay such loyal assistance with substantial political concessions—if not immediate independence or at least dominion status following the war, then surely its promise soon after the Allies won victory. No Indian imagined the war could possibly drag on for four years. The government of India's immediate military support was of vital importance in bolstering the western front, and an expeditionary force including two fully manned infantry divisions, the Meerut and Lahore, and one cavalry division left India in late August and early September 1914. They were shipped directly to France and moved up to the battered Belgian line just in time for the First Battle of Ypres. The Indian Corps sustained extraordinarily heavy losses during the winter campaigns of 1914–15 on the western front. The myth of Indian racial inferiority, especially with respect to courage in battle, was thus dissolved in sepoy blood on Flanders fields. In 1917 Indians were at last admitted to the final bastion of British Indian racial discrimination—the ranks of royal commissioned officers. Indian troops were rushed to East Africa and Egypt as well as the western front in the early months of the war, and by the end of 1914 more than 300,000 officers and men of the British Indian Army had been shipped to overseas garrisons and battlefronts. The Indian Army's most ambitious as well as ill-managed campaign was fought in Mesopotamia. In October 1914, before Turkey joined forces with the Central Powers, the government of India launched an army to the mouth of the Shatt al-Arab to further Curzon's policy of control over the Persian Gulf region. Basra was taken easily in December 1914, and by October 1915 the British Indian army had moved as far north as al-Kūt, barely 100 miles (160 kilometres) from Baghdad. The prize of Baghdad seemed within reach of British arms, but less than two weeks after Gen. Sir Charles Townshend's doomed army of 12,000 Indians started north in November 1915, they were stopped at Ctesiphon, then forced to fall back to al-Kūt, which was surrounded by Turks in December and fell in April 1916. This disaster became a national scandal for Britain and was appropriately investigated by a royal commission, whose report, published in June 1917, led to the immediate resignation of India's secretary of state, Austin Chamberlain. Edwin Montagu, Chamberlain's successor at Whitehall's India Office, informed the Commons on August 20, 1917, that the policy of the British government toward India was thereafter to be one of

The Mesopotamian campaign

> increasing association of Indians in every branch of the administration, and the gradual development of self-governing institutions, with a view to the progressive realization of responsible government in India as an integral part of the Empire.

Soon after this stirring promise of political reward for India's wartime support, Montagu embarked upon a personal tour of India. During his tour, Montagu conferred with his new viceroy, Lord Chelmsford (viceroy from 1916 to 1921), and their lengthy deliberations bore fruit in the Montagu-Chelmsford Report of 1918, the theoretical basis for the Government of India Act of 1919.

Anti-British activity. Anti-British terrorist activity started soon after the war began, sparked by the return to India of hundreds of embittered Sikhs who had sought to emigrate from their Punjab homes to Canada but who were denied permission to disembark in that country because of their colour. As British subjects the Sikhs had assumed they would gain entry to underpopulated Canada, but after wretched months in cramped and unsanitary conditions, with inadequate food supplies aboard an old freighter, they returned to India as confirmed revolutionaries. The Ghadr (Mutiny) Party had been started by Punjābī Sikhs in 1913, and its leaders journeyed abroad in search of arms and money to support their revolution, the foremost among them, Har Dayal, going to Berlin to solicit aid from the Central Powers. Mulsim disaffection also grew and acquired revolutionary dimensions as the Mesopotamian campaign dragged on. Many Indian Muslims appealed to Afghanistan for aid, urging the *amīr* to start a holy war against the British and in defense of the caliphate. After the war, the Khilafat movement, an off-spring of growing pan-Islāmic consciousness in India, was started by two fiery orator-journalists, the brothers Shaukat and Muhammad 'Alī. It lured thousands of Muslim peasants to abandon their village homes and trudge over frozen high passes in a disastrous *hijrah* (flight) from India to Afghanistan. In Bengal terrorist bombings continued to harass officials, despite numerous "preventive detention" arrests made by Indian Criminal Intelligence Division (CID) police under the tough

The Muslim Khilafat movement

martial-law edicts promulgated at the war's inception. The deaths of Gokhale and Mehta in 1915 removed the most powerful moderate leadership from Congress and cleared the way for Tilak's return to power in that organization after its reunification in 1916. The sessions at Lucknow in December 1916 brought even further unity to India's nationalist forces in the form of a pact outlining a program of minimum immediate national demands agreed upon by both Congress and the Muslim League. The Lucknow Pact called first of all for the creation of expanded provincial legislative councils, four-fifths of whose members should be elected directly by the people on as broad a franchise as possible. The League's readiness to unite with Congress was attributed to the pact's stipulation that Muslims should receive a far higher proportion of separate electorate seats in all legislative councils than they had enjoyed under the act of 1909. Thanks to such generous concessions of political power by Congress, Muslim leaders, including Mohammad Ali Jinnah (1876–1949), later to become the *qāʾid-e aʿẓam* ("great leader") of Pakistan, agreed to set aside doctrinal differences and work with the Indian National Congress toward the attainment of national freedom from British rule. This Congress–League rapprochement was, however, short-lived, and by 1917 communal tensions and disagreements once again dominated India's faction-ridden political scene. Tilak and Annie Besant each campaigned for different home-rule leagues, while Muslims worried more about pan-Islāmic problems than all-India questions of unity. Montagu's August 1917 announcement, followed by his popular tour of India, had also brought the prospect of rapid British concessions of power to many political leaders in India, who now returned to entrenched party positions rather than adhering to idealistic schemes of implausible unity. In July 1918, after returning home, Montagu published his lengthy "Report on Indian Constitutional Reforms," noting that he and Chelmsford agreed that "nationhood within the empire represents something better than anything India has hitherto attained." The report went on to elaborate upon the principles on which the government planned to revise British India's constitution, the most novel of which was that of dyarchy (dual government), whereby power over external, financial, and law-and-order matters would remain in official hands, but other affairs would come under direct control of representatives chosen by an electorate.

Indian contribution to World War I

By Armistice Day, November 11, 1918, more than 1,000,000 Indian troops had been shipped overseas to fight or serve as noncombatants behind the Allied lines on every major front from France to Gallipoli. More than 100,000 Indian battle casualties, over 36,000 of which proved fatal, were sustained during the war. India's material and financial contributions to the war effort were hardly less valuable, with some £80,000,000 worth of military stores and equipment shipped from India's ports to various fronts, nearly 5,000,000 tons of wheat valued at over £40,000,000 sent to Great Britain, some £137,000,000 worth of raw jute and sacks alone exported, and vastly expanded output and export of cotton goods as well as rough-tanned hides, wolfram, manganese, mica, saltpetre, timbers, silk, rubber, and various oils. The government of India paid for all of its troops overseas, adding about £20,000,000 annually to its wartime military expenditure, and before the war ended the viceroy presented a gift of £100,000,000 (actually an imperial tax) to the British government. Tata's Iron and Steel Company received Indian government support once the war started and by 1916 was producing 100,000 tons of steel a year. An industrial commission was appointed in 1916 to survey the subcontinent's industrial resources and potential, and in 1917 a munitions board was created to expedite the production of war materials. Wartime inflation was immediately followed by one of India's worst depressions, which came in the wake of the devastating influenza epidemic in 1918, an epidemic that took a far heavier toll of Indian life and resources than all the casualties sustained throughout the war.

The Amritsar massacre. Politically, as well as economically, the postwar years proved depressing to India's high expectations. After the war British officials, who in the first flush of patriotism had abandoned their ICS posts to rush to the front, returned to oust the Indian subordinates acting in their stead and carried on their prewar jobs as though nothing had changed in British India. Indian soldiers also returned from battlefronts to find that back at home they were no longer treated as invaluable allies but reverted immediately to the status of "natives." Most of the soldiers recruited during the war had come from the Punjab, which, with only 7 percent of India's population, had supplied over 50 percent of the combatant troops shipped abroad. It is thus hardly surprising that the flashpoint of postwar violence that shook India in the spring of 1919 was Punjab province. The actual issue that served to rally millions of Indians, arousing them to a new level of disaffection from British rule, was the government of India's hasty passage of the Rowlatt Acts early in 1919. These "black acts," as they came to be called, were peacetime extensions of the wartime emergency measures passed in 1915 and had been rammed through the Supreme Legislative Council over the unanimous opposition of its Indian members. Indian leaders viewed the autocratic enactment of such legislation, following the victorious conclusion of a war in which India had so loyally supported Britain, as a confession of British treachery and duplicity and the abandonment of the promised policy of reform in favour of a new wave of repression. Mohandas Karamchand Gandhi, the Gujarati barrister who had returned from South Africa shortly after the war started and was by then recognized throughout India as one of the most promising leaders of Congress, called upon his countrymen to take sacred vows to disobey the Rowlatt Acts, launching a nationwide movement for the repeal of those repressive measures. Gandhi's appeal received the strongest popular response in the Punjab, where the nationalist leaders Kichloo and Satyapal addressed mass protest rallies from the provincial capital of Lahore to Amritsar, sacred capital of the Sikhs. Gandhi himself had taken a train to the Punjab early in April 1919 to address one of those rallies, but he was arrested at the border station and taken back to Bombay by orders of the tyrannical lieutenant governor of the Punjab, Sir Michael O'Dwyer. On April 10, in Amritsar, Kichloo and Satyapal were arrested and deported from the district by Deputy Commissioner Miles Irving, and when their followers tried to march to Irving's bungalow in the camp to demand the release of their leaders they were fired upon by British troops. With several of their number killed and wounded, the enraged mob rioted through Amritsar's old city, burning British banks, murdering several Englishmen, and attacking two Englishwomen. Gen. R.E.H. Dyer was sent with troops from Jullundur to restore order, and, though no further disturbances occurred in Amritsar until April 13, Dyer marched 50 armed soldiers into the Jallianwalla Bāgh (Garden) that afternoon and ordered them to open fire on a protest meeting attended by some 10,000 unarmed men, women, and children without issuing a word of warning. It was a Sunday, and many neighbouring peasants had come to Amritsar to celebrate a Hindu festival, gathering in the Bāgh, which was a place for holding cattle fairs and other festivities. Dyer kept his troops firing for about ten minutes, until they had shot 1,650 rounds of ammunition into the terror-stricken crowd, which had no way of escaping the Bāgh since the soldiers spanned the only exit. About 400 civilians were killed and some 1,200 wounded. They were left without medical attention by Dyer, who hastily removed his troops to the camp. Sir Michael O'Dwyer fully approved of and supported the Jallianwalla Bāgh massacre, and on April 15, 1919, issued a martial-law decree for the entire Punjab province. Lord Chelmsford characterized the general's action as "an error of judgment," but when Montagu finally learned the atrocious details of the slaughter that had occurred in Amritsar he appointed a commission (under Lord Hunter) to hold a full inquiry in India. Dyer defended his action as

British treachery over reform

> the least amount of firing which would produce the necessary moral and widespread effect it was my duty to produce . . . from a military point of view, not only on those who were present, but more specially throughout the Punjab.

Dyer was relieved of his command, but he returned to England as a hero to many British admirers, who presented him with a collected purse of thousands of pounds and a jewelled sword inscribed "Saviour of the Punjab."

Consequences of the massacre

The Jallianwalla Bāgh massacre turned millions of patient and moderate Indians from loyal supporters of the British rāj into national revolutionaries who would never again trust to British "fair play" or cooperate with a government capable of defending such action. The following year, Mahatma Gandhi launched his first Indian *satyāgraha* ("clinging to the truth") campaign, India's response to the massacre in Jallianwalla Bāgh.

The Government of India Act of 1919. Britain's Parliament still hoped to avert the collision course of confrontation with Indian nationalism that had been set by the government of India by rushing through reforms designed to liberalize the autocracy. These were contained in the Government of India Act of December 1919, known familiarly as the Montagu–Chelmsford Reforms. They initiated a healthy devolution of authority from British to Indian hands, and, had they come before April 1919, they would most probably have received a warm welcome from Congress. The viceroy's Executive Council was now to have at least three Indian members, while the Supreme Legislative Council was to be transformed into a bicameral legislature consisting of an imperial Legislative Assembly and a Council of State. The Assembly would have 140 members, 100 of whom would be elected, with no more than 25 of the remaining 40 to be officials. The Council of State would have 60 members, 20 of whom would be appointed officials, the other 40 to be elected by a more carefully restricted electorate than that that would vote for assembly representatives. Enfranchisement continued to be based on property and education, but under the Act of 1919 a total of some 5,000,000 Indians were to become eligible to vote for provincial council representatives in the 1921 elections. Less than 1,000,000 Indians would be permitted to vote for candidates to the Imperial Assembly, and only 17,000 could choose the conservative Council of State members. Dyarchy was to be introduced at the provincial level, at which executive councils would be divided between responsible ministers elected to preside over "transferred" departments (education, public health, public works, and agriculture) and officials appointed by the governor to rule over "reserved" departments of state (land revenue, justice, police, irrigation, and labour). Provincial legislative councils were expanded in membership. At least 70 percent of their members were to be elected and no more than 20 percent reserved for appointed officialdom. The constitution of British India had thus advanced many strides in its gradual metamorphosis from a totally irresponsible autocracy to a fully responsible parliamentary form of government.

Enlargement of the franchise

Gandhi's strategy. But 1920 saw the beginning of the policy of noncooperation by the Congress. Gandhi set August 1 of that year as the day of fasting and prayer that would launch his nationwide *satyāgraha* movement. For Gandhi there was no dichotomy between religion and politics, and his unique political power was in great measure attributable to the spiritual leadership he exerted over India's masses, who viewed him as a *sādhu* ("saint") and worshipped him as a *mahātma*. He chose *satya* ("truth") and *ahiṃsā* ("nonviolence" or love) as the polar stars of his political movement because the former was the ancient Vedic concept of the real, embodying the very essence of existence itself, while the latter, according to Hindu (as well as Jaina) scripture, was the highest religion (*dharma*). With these two weapons, Gandhi assured his followers, unarmed India could bring the mightiest empire known to history to its knees. His mystic faith magnetized millions, and the sacrificial suffering (*tapasyā*) that he took upon himself by the purity of his chaste life and prolonged fasting armed him with great powers. Gandhi's strategy for bringing the giant machine of British rule to a halt was to call upon his countrymen to start a multiple boycott campaign in 1920, boycotting British-made goods, British schools and colleges, British courts of law, British titles and honours, British elections and elective offices, and, should the need arise if all other boycotts failed, British tax collectors as well. The total withdrawal of Indian support would thus stop the machine, and nonviolent noncooperation would achieve the national goal of *swarāj*.

Muslim reaction to Gandhi

The Muslim quarter of India's population could hardly be expected to respond any more enthusiastically to Gandhi's *satyāgraha* call than they had to Tilak's revivalism, but the Mahatma laboured valiantly to achieve Hindu–Muslim unity by embracing the ʿAlī brothers' Khilāfat movement as the "premier plank" of his national program. Launched in response to news of the Treaty of Sèvres's dismemberment of the Ottoman Empire in 1920, the Khilāfat movement coincided with the inception of *satyāgraha,* thus giving the illusion of unity to India's nationalist agitation. Such unity proved, however, as chimerical as the Khilāfat movement's hope of preserving the caliphate itself, and in December 1920 Mohammad Ali Jinnah left the Nāgpur Congress, disgusted by Gandhi's mass following of Hindī-speaking Hindus. The days of the Lucknow pact were over, and by the start of 1921 the forces of Hindu and Muslim agitation destined to lead to the birth of the independent dominions of India and Pakistan in 1947 were clearly set in motion in their separate directions.

(S.A.Wo.)

Prelude to independence

INDIAN NATIONALISM AND THE END OF THE RĀJ

The final period of British rule can best be understood by studying the interaction of two trends. First, there was the transformation of British policy: the period opened with Britain in the twilight of an era of Liberal reform; then followed a long, almost uninterrupted government by Conservatives, with Labour finally taking over after World War II to implement a program of radical change. The initial reforming impulse of the pre-World War I Liberal governments was thus followed by caution and an insistence upon "safeguards" in constitutional evolution, with the movement toward a transfer of power making rapid headway only after the election of the Labour government in 1945. The second major trend was the endeavour of the Indian National Congress to mobilize a broadly based nationalist movement that would compel the British to hand over power. This endeavour was constantly hampered by the tendency of different sections of the movement to pull away in different directions. In the end, the Muslim political leadership opted for a separate form of independence—Pakistan. Divisive tendencies, which were only partially arrested by the struggle against the British, reasserted themselves after independence in both the new countries.

Response to dyarchy. The Amritsar massacre of 1919 provided an unpromising prelude to the elections which were to usher in dyarchy (see above *British imperial power, 1858–1920*). Gandhi called upon the Congress to boycott the elections. With the deaths of Mehta, Gokhale, and Tilak there was something approaching a vacuum in the leadership of the Congress at this time, and into this Gandhi moved, never to be completely dislodged. But there were other leaders, and those who had supported the old moderates in their faith in constitutional advance now declared that they would contest the elections and work the dyarchical system in order to prepare India for self-government. These leaders may be termed the "old" Congress, for their most venerable spokesman was Surendranath Banerjea, a founder-member of Congress and long known as the Lion of Bengal. Not all, however, were ancients: there was Tej Bahadur Sapru, a brilliant lawyer, Srinivasa Sastri—almost as celebrated as Gandhi for his work among overseas Indians—and M.A. Jinnah, called "the apostle of Hindu–Muslim unity." These men now regretfully seceded from Congress, most joining the Indian Liberal Federation. This was the first of many breakaway movements that were to mark the years of Gandhi's leadership. In three major provinces—Bengal, Bombay, and the United Provinces (now Uttar Pradesh)—ministries were formed by distinguished Liberals. Their period of office (1920–21 to 1923–24) was marked by

Regional politics

impressive legislative reforms. Elsewhere, politics took on regional patterns. Punjab politics was marked by an urban–rural division. The towns were dominated by Hindus with thoroughgoing Congress views. The countryside was fairly equally divided between Muslims and Hindus, with the Sikhs as a sturdy minority. Rural life was dominated by the great landowners, especially in West Punjab. Rural interests were now brought under one umbrella by the formation of the Unionist Party. This was mainly Muslim, but it enjoyed the loyalty of the Hindu Jat cultivators, while Sikh landlords gave it conditional support. Sikh militants also developed a "direct action" form of politics through the fighting Akālis. The Unionist Party asserted an ascendancy over Punjab politics, basing its appeal on rural development. Politics also assumed a regional form in Madras (now Tamil Nādu). The Congress had never emerged as a strongly organized force in the South and was associated closely with the Brahmin elite, a group comprising less than 3 percent of the population but with a monopoly of power and position. An anti-Brahmin movement grew up in the 20th century, drawing upon a sense of pride in the nonpriestly culture of the non-Āryan Tamils. This movement assumed a political form as the Justice Party, which secured a clear majority in the first Madras dyarchy legislature. The leader was designated chief minister, and a regulated pattern of parliamentary government developed.

Having turned his back upon dyarchy, Gandhi endorsed the strangest cause of his infinitely varied public life: that of the Khilāfat, or Muslim unity under the sultan (caliph) of Turkey. The Indian Muslims were anguished by the policy of Prime Minister Lloyd George in backing Greek claims against Turkey. Edwin Montagu's anxiety at his chief's apparent unconcern for the consequences within the Islāmic world caused him to engineer a protest that led Lloyd George to demand his resignation. The Indian Muslims were stirred up by militant religious organizations that made overtures to the "new" militant Congress. They preached the end of the infidel British rule. Congress simultaneously launched a civil-disobedience campaign designed to compel the British to concede *swarāj*. *Swarāj* had been translated as "home rule," but it now became amplified into independence. The dual campaign was designed to demonstrate the unity of the peoples of India: Hindu-Muslim ek hai ("Hindu and Muslim are one"). But the Khilāfat agitation, with its open incitement to religious militancy, stimulated a rising among the coastal Arab–Indian community of South India, the Moplāhs, and they fell upon the Hindu population. In the United Provinces, a Congress mob besieged the Chauri Chaura police station and murdered the constables. Once again, Gandhi called off his campaign because the ordinary people could not understand his message. Gandhi was arrested (March 1922), brought to trial, and sentenced to six years' imprisonment. He was actually in jail until February 1924, being released after a serious operation.

Meanwhile, the Congress came under the influence of what may be called the "middle" Congress. A former moderate, Motilal Nehru, had placed himself beside Gandhi. But he found himself increasingly in alliance with C.R. Das of Bengal, when both emerged from imprisonment after the collapse of civil disobedience. Das insisted that the purpose of politics was fundamental change: he inquired,

> How will it profit India if in place of the white bureaucracy that now rules over her, there is substituted an Indian bureaucracy of the middle classes?

He argued that entry into the legislatures could be the means to force the British to accelerate change. He won over Motilal, and at a special session of Congress in September 1923 he secured a declaration that Congress "suspends all propaganda against entering councils." At the elections in September 1923, the Congress (campaigning as the Swarājya Party) made considerable gains, mainly at the expense of the Liberals. In the Central Provinces they commanded a majority, and in Bengal they formed the largest group, C.R. Das being invited to head the ministry. He refused, and the Swarājists were able to bring ministerial government to a stop. In the Punjab, however, the Unionists fought off the attack and moved on, while in Madras (now Tamil Nādu) the Justice Party also survived as the government.

In London the Conservative government and, in particular, the diehard secretary for India, Lord Birkenhead, looked forward to the time when ten years of dyarchy would be completed; that is, to when a constitutional review would be required. In response to Indian wishes the review was advanced by two years, and the Indian Statutory Commission (Simon Commission) was appointed (1927) to look at the working of dyarchy and recommend future policy. As was usual, this was an all-parties Commission, with the Liberal Sir John Simon as chairman and with C.R. Attlee, a Labour member of Parliament, as one of its members. But the commission was entirely drawn from English public life, although even the Montagu–Chelmsford Report had been prepared with the assistance of Indian political leaders. Almost all shades of Indian opinion condemned this racial exclusivism (as it appeared). An almost complete boycott of the commission was mounted throughout its Indian travels. The Liberals and also the Muslim leader Jinnah joined the Congress in their boycott.

The Simon Commission

The position of the Muslims in politics at this time was ambiguous. The Khilāfat campaign had collapsed—the Turkish leader Mustafa Kemal Atatürk had abolished the office of caliph in 1924. Some Khilāfatists nevertheless hung around on the fringe of the Congress. The Muslim League retained a shadowy existence. Its president (1919–30) was Jinnah, but its membership was a few hundred only, and even these were divided. Some Muslims, such as Fazl-i-Husain, the leader of the Punjab Unionist Party, had their own political power base; some were traditionalists, still looking belatedly to the British rulers for patronage. But, across their divisions, most of the Muslims opposed the Simon Commission.

In an effort to reply to British charges that Indians could not agree among themselves, an all-parties committee was set up under Motilal Nehru as chairman, to produce a draft Indian constitution. This emerged as a moderate document. Dominion status—that is, equality with Canada, Australia, New Zealand, and South Africa—was demanded "as the next immediate step." There would be a two-chamber Parliament, but the governor general would remain as the constitutional representative of the British crown. This constitution served to alarm the Muslims and reunite them in their distress. A conference was convened under the Aga Khan, and subsequently Jinnah produced what came to be known as "the 14 points." This was a demand for a weak federation, with autonomy to the provinces and measures to endow the North-West Frontier, Sind, and Baluchistan (the solidly Muslim areas of the northwest) with equal status alongside the other provinces. In the central government, the Muslims wanted at least one-third of the representation, although numerically they formed less than a quarter of the total population.

The Round Table Conferences. In Britain, a Labour government was briefly in office (1929–31). The viceroy, Lord Irwin, was an ultraconservative, but he developed an empathy for the Indian condition that included a personal understanding with Gandhi. Irwin announced that the goal of British policy in India was indeed dominion status. This infuriated the Conservative right wing, led by Lord Salisbury and Winston Churchill. But the Conservative centre (followers of Baldwin), as well as the Liberal and Labour parties, were ready to talk about reform. It was proposed to hold a "Round Table" Conference in London between representatives of the Indian communities and the British political parties. The first session was convened in the autumn of 1930. Meanwhile, the Simon Commission had reported, recommending that dyarchy in the provinces be replaced by full ministerial government. The Commission recoiled from the idea of parliamentary central government and envisaged an enlarged legislature as a forum of debate.

The Congress was suspicious of British tactics and intentions. A new generation of leaders was coming forward. The two new stars riding in the firmament were Jawaharlal Nehru, son of Motilal Nehru, regarded as the disciple of Gandhi (though he was an avowed Socialist, whereas

Gandhi—if he could be given a label—was an anarchist), and Subhas Chandra Bose, formerly lieutenant of C.R. Das and heir to the militant, physical-force element in Bengali politics. It was a symbol of the new mood that, whereas the older Nehru had been satisfied with dominion status, his son was determined to have a Socialist republic. On January 26, 1930, the Congress, meeting at Lahore, pledged itself to a declaration of independence.

Positions at the first conference

When the first Round Table Conference met (November 1930–January 1931), the Congress was absent. Political India was represented by a group of Liberals—distinguished individuals, but, politically speaking, only middlemen—as well as spokesmen for the minorities, including Jinnah and the Aga Khan for the Muslims. There was another distinct group, for the 600 princes and princelings who still retained their quasi-independent status under British suzerainty (or "paramountcy," as it was termed) had sent their representatives. A few of these rulers (notably the Maharaja of Mysore) had steered their states forward into systems of representative government. But the majority still clung to an arbitrary autocracy, and these included the two largest, Hyderābād and Kashmir. If British India were to move forward to dominion status, where would this leave the princely states? British policy groped toward some means for closer association through the political device of federation. Since Canada adopted federation in 1867, as the solution for provincial differences and difficulties, federation had been invoked as the cure-all remedy for political impasses. Federation had worked in Australia; it had failed in Ireland. Now it was proposed for India. To general surprise, the princes departed from their usual position of immobility to announce that they were prepared to enter a federal form of government.

To the extent that the general principle of an all-India federation was accepted by both sides, the conference was a success, and an atmosphere of goodwill was created. There was a prospect of involving the Congress, even though the party had embarked upon another campaign of civil disobedience. In March 1931 there emerged the Gandhi–Irwin pact, by which Gandhi agreed to call off *satyāgraha* in return for the government releasing all political prisoners not guilty of violence and withdrawing pending prosecutions. The objective was for the two sides to meet at a second London conference. This meeting occurred in the autumn of 1931. By then, Britain was just emerging from the trough of the world depression, and the Labour government had given way to a national, Conservative-majority administration. Congress was represented solely by Gandhi, with Sarojini Naidu at his side. He asked for the substance of dominion status; that is, full self-government. The British countered by talk of safeguards and reservations. They were given justification for this strategy, as the representatives of the Muslims and of other minorities, such as the untouchables, stuck rigidly to exactly this demand. The second session closed on a note of anticlimax, and Gandhi returned to India to resume noncooperation.

As a result of this anti-British activity, Gandhi and the other Congress leaders were arrested. But Indian economic conditions had eased somewhat, and the political pressure eased in response. The failure of Gandhian nonviolence to produce short-term results led to pressures within the party. A number of Congress members drifted away to join the Liberals and others in working the legislative system: these were called Responsivists, because they advocated a positive response to dyarchy. In 1934 there was an attempt to swing the Congress in a left, militant direction. A Socialist Party was formed within the Congress, its leaders young intellectuals such as Jaya Prakash Narayan and Asok Mehta. Jawaharlal Nehru was identified with the group, while Gandhi was sympathetic and influenced the Socialists away from centralism.

The Communal Award of 1932

While Indian politics was in the doldrums, the government in Britain was laboriously fashioning a constitution to replace dyarchy. The Round Table Conference had struggled on, after the departure of Gandhi, through a third session; but only the persistent Indian Liberals took it seriously. To cut through the deadlock that had settled over the relations between the dominant-caste Hindus, as the British saw the Congress, and the Muslims and untouchables, the prime minister, Ramsay MacDonald, issued a Communal Award in August 1932. This continued the system of separate electorates and separate constituencies for all who might be considered minorities and was viewed by most Indians as a blatant example of the policy of "divide and rule." The main beneficiaries were the Muslims, but the principle also applied to the Sikhs in Punjab, to Indian Christians, Eurasians, the aboriginal tribes, special interests (women, organized labour, business, landlords, universities)—and to the untouchables. Gandhi resented the last award especially, because it directly contradicted his campaign to integrate those he called Harijans (People of God) into the wider community. Gandhi succeeded in bypassing the award by negotiating an agreement with B.R. Ambedkar, the so-called untouchables' leader, dramatizing his stand by yet another fast.

The Government of India Act of 1935. After the Communal Award, the Conservative secretary for India, Sir Samuel Hoare, proceeded with his officials to construct an elaborate structure for the next constitutional advance; this was to become the longest statute ever to be framed by a British Parliament. The Conservative centre gave its support, though this was based on the calculation that the Muslims and the princes would balance the Congress. The right wing, led by Winston Churchill, staged an all-out revolt that nearly secured the backing of the Conservative Party conference, and that helped to delay the passing of the Government of India Act till 1935. Under the act, government in the provinces was handed over to ministries elected on a much wider franchise (the total Indian electorate now numbered 40,000,000). At the centre, a new federation was envisaged, but its introduction was dependent upon the willing accession of one-half of the princely states. The act created an elaborate federal structure, including a Supreme Court, and embodied safeguards and reservations designed to take care of any possible contingency. All this ensured Congress' condemnation of the act: Nehru described it as "a Charter of Slavery." What escaped everyone's notice at this time (though it was realized later) was that, if the major political organization in the land should decide to move in and operate the new constitution, there was little the British could do to hold up the inevitable transfer of power. Previous reform measures had included a provision for a further constitutional review; the 1935 act did not, for it was planned as the last installment before full self-government—even though, in the 1930s, the majority of British politicians and officials probably envisaged the final era lasting as long as dyarchy (nearly 20 years).

The elections of 1936

The first elections under the new constitution were held at the end of 1936. Congress hesitated between total boycott (advocated by Nehru) and participation in order to wreck the new constitution. The latter policy prevailed. Congress entered the fight with an integrated organization and a simple, effective appeal to the people. The Muslim League, its would-be rival, was compelled to improvise arrangements in the last few weeks before the poll. There were 1,585 seats open to contest. Congress contested 1,161 and gained 716, winning an absolute majority of seats in five provinces (the United Provinces, the Central Provinces, Bihār, Orissa, and Madras), while just short of a majority in Bombay. Because of inchoate organization, the Muslim League was able to contest only about 150 of the 482 constituencies reserved for Muslims. It won 109 seats, but in no province did the League gain a majority.

The extent of the Congress success surprised all opponents and possibly even its own people. What now? There followed an intraparty debate about accepting office. Nehru and the militants objected, and when overruled insisted that the Viceroy and governors must agree not to use their special powers in face of popular government. No such undertakings were given, but Congress nevertheless proceeded to take office. It was a historic moment. Apart from the brief encounter of the Round Table Conference, the Congress and the British rulers had stood apart. Now they were working together in harness. In addition to the six provinces listed above, Congress was also able to form ministries in Assam and in the North-West Frontier, where

the Pashtun Red Shirts (Khudai Khitmatgars) adhered to Congress. Only Punjab, still staunchly Unionist, Sind, under local barons, and Bengal, governed by an unstable coalition, eluded the Congress dominance.

World War II. The experience of actually operating the machinery of government consolidated the middle Congress. The organization men, the men of affairs, were in the ascendant—Gandhi was less relevant. But the situation was reversed by an external factor: the outbreak of World War II. The viceroy, Linlithgow, chose to observe the strict formalities of protocol and declared that India was at war with Nazi Germany without first gaining the support of the national leaders toward committing their country to war. The Working Committee of the All-India National Congress (the High Command, as it was often called) ordered all the Congress provincial ministries to resign in protest. There was considerable reluctance to follow orders, but by October all had resigned. The initiative in Indian politics was no longer in Congress' hands.

Jinnah's demand for Muslim sovereignty

Jinnah called on the Muslimsto observe a day of thanksgiving when the Congress left office and raised the level of Muslim demands. At their annual gathering, held at Lahore in March 1940, the League passed a resolution demanding that geographically contiguous units be:

> Demarcated into regions which shall be so constituted with such territorial readjustments as may be necessary that the areas in which the Muslims are numerically in a majority, as in the north-western and eastern zones of India, should be grouped to constitute 'independent states' in which the constituent units shall be autonomous and sovereign.

There could be different interpretations of the intentions of this resolution: were the Muslims seeking separation or a looser federal structure? And how were the Muslim majority zones to be defined when in both Bengal and Punjab the communities were almost inextricably intermeshed in town and countryside? But the ambiguity was resolved by the press, which labelled this demand as the Pakistan Resolution. Previously, Pakistan (Land of the Pure), as a separate homeland for the Muslims of India, had been dismissed as the dream or illusion of Rahmat Ali, a Muslim exile in Cambridge. But now Pakistan became the talking point in every discussion about India's political future.

Indian constitutional advance was put to one side by the British government while the German armies occupied western Europe, and British Commonwealth forces (including Indian army divisions) chased Mussolini's troops out of North and East Africa. But, when Japan struck down into Southeast Asia at the beginning of 1942, the threat to India galvanized the British war cabinet into reconsidering the political stalemate in India as a matter of urgency. In February 1942, the members of the war cabinet were formed into committees designed to take over responsibility for important areas of policy. The India Committee was presided over by C.R. Attlee, then leader of the Labour Party and also deputy prime minister. His colleagues included Simon, leader of the 1928 commission, several Conservatives, and the Socialist Sir Stafford Cripps. A plan was rapidly submitted by Attlee for a declaration of British policy that would bind the government to setting up an Indian constituent assembly immediately after the war terminated, to work out the terms of Indian independence. It was to be made clear, however, that any province that did not wish to accept the new constitution might continue, for the time being, under existing arrangements. In return for this commitment it was hoped that the political parties—especially the Congress—would join in the prosecution of the war against Japan. It was decided to send Cripps to India to persuade the politicians to cooperate.

The Cripps mission

Cripps arrived in Delhi on March 23, 1942, and plunged into intensive discussion with all the politicians, especially the Congress. Argument centred mainly on the short-term question of Indian national participation in the war effort. But deadlock was soon reached over the Congress claim—stated by their president Abul Kalam Azad—that Cripps had conceded "a National Government which would function as a Cabinet," from which Cripps backed down. The Cripps mission, as the first of a series of bargaining encounters that were to take place between the British and Indian leaders in the years leading up to independence, showed how misunderstanding would constantly degenerate into deadlock, essentially as both sides were unclear about the consequences of their own and their opponents' plans and proposals. Mutual suspicion could rapidly become mutual hostility. Jinnah and the Muslims had played a waiting game during the Cripps–Congress discussions but had seized upon the proviso that a province might not be compelled, against its own expressed majority opinion, to accept a constitution devised by the rest of India.

The Japanese advance continued, and by May 1942 almost the whole of Burma was within their grasp. But the Japanese attack was halted on the jungle hillsides of the Assam–Burma border. For two years there would be no significant change in the fortunes of war in the Burma theatre. But this was not clear in 1942, either to Indians or to British. Both anticipated an invasion, and Gandhi re-emerged to call upon the British to leave Indians to deal with the Japanese by nonviolent means. The slogan "Quit India" now appeared everywhere. As there was no response, the All-India Congress Committee issued a statement in August 1942 demanding an immediate departure: if this was refused, Congress would begin "a mass struggle on nonviolent lines on the widest possible scale." Two days later Gandhi and the Working Committee were arrested. The leadership of the struggle passed to younger men who intended to fight: "Let us Do or Die" was their slogan. Jayaprakash Narayan, already on the run from the police, led the guerrilla campaign. But, as so often happens when violence is unleashed, the campaign passed largely into the hands of bullies and bandits. Within six weeks the revolt (which was largely confined to the eastern districts of the United Provinces and Bihār) had been suppressed, and for two years there was silence in Indian politics.

The Congress rebellion

THE TRANSFER OF POWER

In October 1943, Field Marshal Lord Wavell, the architect of the early British victories in Africa, became viceroy. He was a distinguished leader but found great difficulty in communicating with any warmth to all but his closest acquaintances. He nevertheless attempted to reopen negotiations with the Congress leaders, who from 1944 were gradually released from internment. The Muslim League had made a considerable impression on the political scene, finding personnel for ministries in several provinces. C.R. Rajagopalachari—who was not interned (the 1942 outbreak hardly affected South India)—came forward in April 1944 with a formula whereby the Muslims would be given the opportunity to decide, by referendum, whether they wanted to form a separate state. After his release, Gandhi met Jinnah (September 1944) to discuss a similar formula, but they were unable to agree. Wavell decided that this might be the moment to take the initiative. After consultations in London, he convened an all-parties conference at Simla in June 1945 to discuss the formation of a government representing the main Indian political groups, but Congress and the League could not agree about the basis on which ministers should be chosen.

Policy of the Attlee government. Soon after Simla there was a general election in Britain in which Labour came to power with a large majority. Attlee became prime minister and reconstituted an India–Burma Committee, with Cripps and the new secretary for India, Lord Pethick-Lawrence, as members. The new government was committed to independence for India but took over the 1942 Cripps position: that no unwilling province should be coerced into accepting an independence that did not safeguard minority rights. What would happen if agreement could not be reached? Lord Wavell had come down firmly on the necessity of Indian unity, but he advised the Cabinet that if Jinnah refused to cooperate with Congress he should be offered a Pakistan that included only the actual Muslim-majority districts. This would mean the division of two great provinces—Bengal and Punjab. Wavell still believed (like most British officials) that Jinnah was bluffing, that he was using Pakistan as a stalking horse in order to gain as many concessions for the Muslims as possible.

The elections of 1946

The Labour Cabinet was not yet ready to test Jinnah's bluff. First, elections to choose a central legislature and

legislatures for the provinces were announced. Both the Congress and the League went into the elections with confidence. This time the League had the support of local-level organization, backed by the enthusiasm of active workers, among whom students were numerous. The results simplified Indian politics into a struggle of Congress versus the Muslim League. At the centre, the League won all the seats reserved for Muslims, but Congress won all but five of the nonreserved seats. In the provincial elections—held on a wider franchise, representing 30 percent of the adult population—the League gained 439 of the 494 Muslim seats, while Congress swamped all the other parties and stood to take office in eight provinces. Paradoxically, despite its spectacular success in terms of votes, the League was able to form ministries only in Sind and Bengal, and even then under somewhat unreliable leaders from a party viewpoint. In the Punjab the League was just short of a majority; the Unionists still clung to office.

Against this unpromising background of a confrontation between two monolithic parties, both having an electoral mandate to demand opposing terms, the Labour government decided to send a mission to Delhi to work out constitutional proposals with the Indian leaders. The cabinet mission was led by Lord Pethick-Lawrence, but Cripps was the dominant personality, with A.V. Alexander, minister of defense, as the third member. Because the failure of the 1942 Cripps mission was partly ascribed to estrangement from the then viceroy, the 1946 mission included Wavell as a fourth member for almost all purposes.

The delegates began by listening to all shades of opinion. Attempts to evolve an agreed solution representing the combined views of Congress and the League soon failed, and the mission was compelled to produce its own formula. This was an ingenious plan, devised initially by Cripps, to create three tiers of government, the lowest (provincial) tier being arranged in groups of Hindu- and Muslim-majority provinces. After prolonged discussion and manoeuvring, both the Congress and the Muslim League failed to accept the plan. The League announced that the time had come for Muslim "direct action," and August 16, 1946, was announced as Direct Action Day. It was marked by an impressive display of Muslim solidarity but was overshadowed by riots and atrocities in Calcutta, during which some 4,000 persons were killed in the course of three days.

Muslim Direct Action Day (August 1946)

Immediately after the "Great Calcutta Killing" (as it was soon called), Wavell formed an interim government of Indian politicians. It consisted of eight adherents of Congress and four independents. In October, five Congress nominees resigned to make way for League members, though Jinnah declined to serve alongside Nehru. This did little for interfaith relations, and a wave of killings swept across East Bengal.

Although events were clearly reaching the breaking point, Nehru and Gandhi still resisted any concession affecting the unity of India. But Vallabhbhai Patel and other prominent Congress figures began to realize that the deadlock must be broken somehow or else India would drift into civil war.

Wavell also understood the need for urgency and informed the Labour government that he must have a time limit for the termination of British rule. Attlee had come to feel that Wavell had lost his way, and decided to replace him with Adm. Lord Louis Mountbatten, the former supreme commander of Southeast Asia Command. On February 20, 1947, Attlee told Parliament that it was the "definite intention" of the government to transfer power "into responsible Indian hands by a date not later than June 1948." Because there was no assurance that agreement could be reached on the form of government for the new India, power might have to be transferred to some of the provincial governments. At the end of his statement, Attlee announced that Wavell would be succeeded as viceroy by Mountbatten.

Partition. Very soon after the latter arrived in India in March 1947, he decided that the goal of a united polity (which he was briefed to attain) was no longer feasible. In Punjab, the Unionist ministry had resigned, and the province was ungovernable. No party ministry could contain the rising tide of hatred that was welling up among Muslims, Sikhs, and Hindus. The British governor took over the responsibility for public order, but the riots and killings continued. Mountbatten had to act fast, and he showed a drive and determination that carried others along with him. If Pakistan was to be conceded, then Jinnah had to understand that he must accept the logic of his own doctrine: the non-Muslim districts of Punjab and Bengal would be denied to him. But Jinnah himself saw that there was something more important than territory, for he told Mountbatten: "I do not care how little you give me so long as you give it to me completely."

Behind the scenes, Patel was reconciling his Congress supporters to Pakistan, though he was determined that the Muslims would not receive one scrap more than their due. He was also determined that Indian unity would be preserved by incorporating the 600 princely states into the new union, so that the old patchwork of ambiguities and obsolescences would be replaced by a unified administration. Thus, when Mountbatten set to work to persuade Gandhi and Nehru, he did not have to worry about the rest of the Congress. Gandhi submitted his own plan to win back the Muslims—by voluntarily putting them in control of a united India. Congress would not swallow this experiment in understanding, so Gandhi withdrew himself from the debate. Nehru was persuaded by the argument that communities should not be coerced into joining a political entity against their will. But, while agreeing with Mountbatten, he seems to have still cherished the belief that the Muslim League did not really represent the will of the Muslim masses.

At the end of April, Mountbatten sent a draft plan to London for Cabinet consideration. They made a number of minor modifications but accepted the overall objective: that the peoples of India should themselves choose, through their elected representatives, whether they wanted to partition their land. Mountbatten gave Nehru a preview of the plan and went personally to London to seek Cabinet approval of modifications de-signed to meet Nehru's objections. On return, he called a meeting of the leaders—Nehru, Patel, Jinnah, and others, with Baldev Singh for the Sikhs. The provinces where actual Congress governments were functioning were assumed to wish to merge in a united India. Sind, Baluchistan, and the North-West Frontier were to make a straight choice: India or a new entity (for Pakistan was still not identified by name). The method of choice varied. The Frontier voters were to take part in a popular plebiscite, and in the two great provinces that might be divided (Punjab and Bengal) there was to be a preliminary meeting of all members of the provincial legislatures to vote on whether to adhere to India or adhere to the new entity as a complete province. Then the legislators were to separate into two groups representing the Muslim- and non-Muslim majority districts and to decide whether or not their province should be partitioned: a majority in favour of partition in either section would decide the issue. If Bengal decided on partition, then the Sylhet district of Assam, which included a majority of Muslims, was to hold a plebiscite to decide whether or not to join the Muslim majority part of Bengal. The plan was accepted by the leaders of Congress and the League and was then announced to the world. At the press conference at which the details were discussed, Mountbatten revealed that the date for the transfer of power "could be about the 15th of August" 1947; this allowed less than ten weeks to prepare for possible partition and independence. Events now seemed to take over from people. The ritual of choosing for India or Pakistan was carried out, and there were no surprises. The Bengal and Punjab legislatures met, and a majority (the Muslims) voted that the undivided provinces should leave India. Then they divided into two parts, and the non-Muslim group voted for partition: thus Bengal and Punjab were scheduled for dismemberment. The other Muslim-majority areas opted for the still-unnamed Pakistan. On the Frontier a majority of the total electorate voted for Pakistan.

The Mountbatten plan

There were four great questions to be solved in the ten weeks before partition. The assets of undivided India—financial and material—had to be equitably apportioned

between the two new countries; the public services had to be dissolved and reconstituted in two portions; and a frontier had to be demarcated within the mixed Muslim–Hindu–Sikh populations of Punjab and Bengal. The fourth issue was the most elusive. The Mountbatten plan had produced a formula whereby British India could be split into two new polities, but nothing whatever had been decided about the future of the 600 princely states, which, in theory, were in various kinds of feudal relationship with the British crown—and which, in theory, would revert to their pre-British status of sovereignty when paramountcy (the suzerainty of the British king-emperor) lapsed with the end of imperial rule.

The division of the assets was entrusted to a Partition Committee. The bulk of the material assets was in the territory of the future Indian Union, and there it remained. Financial assets were also frozen until, after independence, Gandhi entered into a fast to reconcile India and Pakistan, and then some restitution was made. The allocation of the public services was on a basis of individual choice: Muslim officials who belonged to areas within the future Indian Union could choose to go to Pakistan, and vice versa. The division of the Indian army was more difficult. Few units were altogether Hindu, Sikh, or Muslim. The officers and men had to choose their option and then await posting to the country of their choice. Meanwhile, a spirit of communal hostility intensified in all the services, most deplorably in the police.

The boundary award

Even more crucial was the delineation of a new frontier that would satisfy the expectations of both the new countries. Two judicial tribunals were appointed, with judges adhering to both the new countries, and with a joint chairman—the English jurist Sir Cyril (later Lord) Radcliffe. The claims on both sides were so extravagant and the judges aligned themselves so obviously in political and not judicial terms that Radcliffe alone had to invent the new frontiers. He did his best, but the communities were so intermeshed and the clashes of interests so irreconcilable that he was compelled to adjudicate under grave disadvantages. He delivered his award to Mountbatten two days before the transfer of power (on August 13). Knowing that the Radcliffe award would offend both sides, the Viceroy delayed its promulgation until after independence. Perhaps with some reason, Pakistan felt that it had received bare justice; but this did not make the new India any more pleased with the award.

The transformation of princely India was not so tidy. Once again Mountbatten exploited the fact that he knew where he wanted to go, whereas almost no one else was so firm of purpose. The British Labour government had formulated no overall policy for the princes. Mountbatten coaxed and cajoled the great majority of them into signing instruments of accession to the new India. The Muslim state of Bawahalpur, which bordered the West Punjab, naturally adhered to Pakistan, as did the Khān of Kalāt in Baluchistan and some other chieftains contiguous to the new West Pakistan. Hyderābād, in the heart of South India, had a population of 16,000,000, predominantly Hindu, under the rule of a Muslim dynasty. The Niẓām (popularly supposed to be the richest man in the world) intended to make himself independent and held out against Indian attempts to bring him in. His plan miscarried, largely because a local populist Muslim leader launched a militant movement (the Razakars) that provided a motive for Indian military forces to intervene and restore order, in September 1948. Hyderābād accepted the same conditions as other former princely states and was later partitioned and disappeared from the map.

The Kashmir problem

Kashmir, in the north, presented a similar problem. It had a population of more than 4,000,000, a definite Muslim majority, and a Hindu ruler. He also hoped to become independent and negotiated a "standstill" agreement with both India and Pakistan. Meanwhile, popular movements developed within Kashmir against princely autocracy. The maharaja replied with repression. With the assent of the new Pakistan government, tribal fighters from the hills of the Pakistan frontier poured into Kashmir proclaiming *jihād,* a holy war. The maharaja fled to Delhi and acceded to India in return for military assistance. Mountbatten (who continued as governor general of India—without Pakistan—after independence) accepted the accession, subject to this being confirmed by popular referendum. Indian forces arrived by air and threw back the tribesmen. Pakistan replied by infiltrating regular troops, until the new states were locked in an undeclared war. They were saved from escalation of the conflict by the initiative of the United Nations in bringing about a cease-fire in January 1949. The cease-fire line (which left the Vale of Kashmir and the bulk of the state on the Indian side) became the de facto frontier, though neither India nor Pakistan recognized it as such.

India and Pakistan rapidly adapted to nationhood. In India, the putative Constituent Assembly became the new parliament, with Jawaharlal Nehru as prime minister. In Pakistan, Jinnah chose the more withdrawn office of governor general, leaving his lieutenant, Liaquat Ali Khan, to be prime minister and chief executive. The new India and Pakistan had to face an immediate crisis in the unexpected influx of refugees. Seven to eight million people left their homes in India and fled to Pakistan, and about the same number made the nightmare journey in the other direction. Perhaps 200,000 did not complete the journey and were slain on the way. This was the consequence of the communal conflicts that had preceded independence. At the top level of political negotiation, the transfer of power was conducted with dignity, even nobility; but, at the base level of the man and woman in the street and village in those areas that were partitioned, it was carried out amid suffering, loss, and bitterness.

The Republic of India

NEHRU'S PREMIERSHIP

Nehru was already prime minister of India (effectively) at independence, and he retained his arduous office until his death, in May 1964. The burden of leadership—as head of a mighty party, guide to the nation, and international statesman—was intolerably heavy. During the 1950s Nehru demonstrated an extraordinary vitality and endurance, but the unremitting strain inevitably took its toll, and the last years saw a sad decline in his powers, which was inevitably reflected in the quality of national life. By then, however, Nehru had stamped his imprint upon his country so firmly that India is still, recognizably, Nehru's India. His three major bequests were a democratic, parliamentary system of government, an economy conditioned by Socialism and centralized planning, and a vision of society moving toward secularism and social justice.

The constitution

India's constitution was completed by November 1949. A federal union of states was firmly held together by a strongly centralized national government, the core of which was the prime minister and his Cabinet, responsible to the Lok Sabhā (House of the People), the lower house of the central Parliament. The Rājya Sabhā (Council of States), the upper house of the central Parliament, is composed of appointed and indirectly elected members and is not subject to dissolution; its powers are less than those of the Lok Sabhā. The Supreme Court and the president were established as further safeguards of the constitution and the nation, but they function as guardians against misgovernment, not at all as rivals to the parliamentary executive. The constitution provides a charter of rights that have a positive, directive quality. Article 17, for example, declares: "Untouchability is abolished and its practice in any form is forbidden."

The elections of 1952 and 1957. With a new constitution, India prepared (1952) for its first general election to be held on the basis of universal adult franchise. For the Lok Sabhā, 489 members were to be chosen, while the state assemblies provided places for nearly 3,400 members. The Congress—much more than a political party and usually described as the Freedom Movement—dominated the national and local scene. The leaders had proved their patriotism by demonstration and by incarceration in British jails. They enjoyed the support of the rural masses, and the organization was fortified by funds from businessmen and landlords alike. The minorities looked to the Congress for protection, as previously they had looked to the British rāj. In opposition to the Congress there were only minor

parties and a host of independents. The great opponent of the Congress—the Muslim League—was disbanded (except in the extreme south, where a Muslim pocket kept a branch going), and most Muslims looked to the Congress for protection. The Socialist Party broke away from the Congress in 1948 and presented an original program. The Communists—who had grown greatly in strength during World War II—were well organized only in certain areas, and the Hindu religious organizations were discredited by the assassination of Gandhi by a Hindu fanatic. Thus, when the nation went to the polls in 1952, the Congress won dramatic victories everywhere. In the Lok Sabhā they gained 364 of the 489 seats. The Communists came next (16 seats), followed by the Socialists (12 seats). This commanding position of the Congress was based upon 45 percent of the popular vote. The Communists emerged as the main opposition in four states—those areas where they were later to demonstrate grass-roots activism. Elsewhere, the opposition was formed from regional or local groups, based mainly on traditional loyalties.

Consolidation of Congress

The period between the first elections of 1952 and the second elections of 1957 was one of achievement. The First Five-Year Plan (1950–55) was completed, and development goals were fully achieved, and the second plan (1956–61) was launched with a more ambitious overall target. The second general election in 1957 was an endorsement by the people of the success of Nehru's leadership: there was actually an increase in the government's majority to 371 seats out of a total of 494 in the Lok Sabhā, while the Communists made a modest advance to 27 members. The arrival of the right-wing Hindu Bharatiya Jan Sangh Party, with four members, seemed of very little significance. (H.R.T.)

Reorganization of the states. Although the Congress—and Nehru—were impregnable at Delhi, there were indications that at the state level the pattern of politics might change. In southern India the speakers of the Dravidian languages (a group quite distinct from the Indo-Aryan languages of north and central India) began actively campaigning for a redistribution of the existing administrative structure into states based upon the languages: Tamil, Telugu, Kannaḍa, and Malayālam. The first militant movement arose among the Telugu speakers, who were divided between the states of Madras and Hyderābād, and in 1953 Andhra Pradesh came into being as the Telugu state. Following an inquiry by a States Reorganization Commission (1956), the whole political map of India was redrawn, the previous mixture of British-Indian provinces and princely states being reduced to a new pattern of 14 states and six union territories representing linguistic regions. The transformation was most radical in the south, the northern tier remaining largely unchanged. A question mark remained over two states—Bombay, where two languages, Marāṭhī and Gujarātī, coexisted, and Punjab, where Hindī and Punjābī were jointly spoken. Bombay was to witness militant protest by the Marāṭhī speakers until at last the state was partitioned in 1960 into the states of Mahārāshtra and Gujarāt. In Punjab the Sikhs launched a mass campaign to establish a state in which they would be the majority community. Nehru refused to concede this demand, despite extreme political pressure, but the state was later partitioned in 1966 into Punjab, Haryāna, and the union territory of Chandīgarh. In 1985 the national government agreed to adjust the regional boundaries so that Sikhs would have a decisive majority in the state of Punjab, with the city of Chandīgarh as their exclusive capital.

Demands of linguistic minorities

As the states were redefined on the basis of local and regional consciousness, so political parties at this level increasingly adopted postures that were founded in regional issues and that had little relation to overall conceptions of nationalism and Socialism. In Madras (now Tamil Nādu) an avowedly Tamil separatist party, the Dravida Munnetra Kazagham (DMK), speaking for the lowly castes against the Brahmins, emerged to challenge the Congress. In nearby Kerala there was a major breakthrough by the Communists at the 1957 election, and the first Communist state government took office under E.M.S. Namboodiripad, a former Gandhian.

The elections of 1962. Nehru's last general election in 1962 took place against a background of gathering uneasiness. Foreign policy, based upon nonalignment, appeared to be under strain along the Himalayan frontier with China, and the invasion of Goa in December 1961, though a military success, was a sign of the abandonment of nonviolence. The economic program, though still going forward, had lost its momentum. Most important of all, Nehru was tired and old and relied too much upon his own personal "court." At the national level, nevertheless, the Congress retained a remarkable degree of support. The Lok Sabhā then consisted of 494 seats. On a popular vote of 48 percent (the highest so far) Congress obtained 361 seats. The opposition, previously strongest on the left, now split almost equally between left and right. The Bharatiya Jan Sangh increased its membership to 14, while a new secular conservative party, Swatantra (Freedom), gained 18 seats. The Communists remained the largest opposition group, with 29 members.

Foreign policy. Indian foreign policy during the Nehru period may be described as one of nonalignment with either of the two world power blocs. The dispute with Pakistan rumbled on, however, causing a major part of the Indian army to be permanently deployed in Jammu and Kashmir and in the Punjab. In Ladākh, a remote northern province, the Aksai Chin had been vaguely demarcated on maps, though not upon the actual terrain, as part of the Kashmir domain and therefore now appeared as part of the Indian Union. But from 1950 onward the medieval Tibetan state was gradually taken over by the People's Republic of China. Communication between Lhasa and the rest of China had to depend upon a route that ran across the Aksai Chin salient, and the Chinese began to build a modern road. There was another border where conflict might occur, in the hills to the north of Assam. A border had been jointly agreed at a conference in Simla in 1914 between British and Tibetan officials and became known as the McMahon Line, from the name of the British negotiator. Only after World War II was administrative control actually extended to this area.

In 1954 India and China signed a treaty designed to symbolize the friendship of two newly liberated countries within the "Five Principles of Peaceful Co-Existence" known as Panchshul. Relations between the two countries were complicated when in 1959 the Dalai Lama, the spiritual leader of Tibet, fled with thousands of his supporters into India. China indicated through diplomatic initiatives that a joint review of the frontier ought to proceed in order to demarcate a line representing present agreements rather than past imperial "aggression." India remained deaf to these overtures and insisted upon the legal validity of the McMahon Line. Meanwhile, in 1959 and 1960, China concluded agreements with Pakistan and Burma, respectively, to settle the petty disputes that had prevailed since the British period over certain sectors of their frontiers. India was incensed because the agreement with Pakistan covered sections of the former Kashmir boundary, occupied by Pakistan in 1947 but still claimed as de jure Indian territory.

Friendship with China

In the autumn of 1962 orders were issued to consolidate Indian border forces in forward positions. Indian troops moved into what was a disputed area, even according to the 1914 agreements. The Chinese retaliated, both in Ladākh and in NEFA, launching a major offensive that caught the Indians off balance (October 20, 1962). Tawang was occupied, and the Assam plains lay open to attack, whereas in Ladākh a number of army posts were captured. In this emergency, Nehru urgently called for British, Commonwealth, and U.S. assistance. British guns and U.S. bombers were rushed in, but, as unexpectedly as the war began, it ended. On November 7 the Chinese unilaterally announced that they were withdrawing their forces in the northeast back to the lines of actual control before the conflict, intimating that India must observe a neutralized buffer zone between the frontier forces. There was no withdrawal in Ladākh, and China was left in undisputed control of Aksai Chin and the border road. It was a sad finale to the policy of nonalignment.

The Himalaya border conflict of 1962

The war of 1965. After the frontier war with China, morale in India was low. The economic development pro-

gram was levelling off. Nehru died in May 1964 and was succeeded by a meek but by no means undistinguished Gandhian, Lal Bahadur Shastri. In contrast, Pakistan was conscious of doing very well. For the first time the inferiority complex that had afflicted Pakistan vis-à-vis India was lifted. In this atmosphere Zulfikar Ali Bhutto, then foreign minister of Pakistan, conceived the dangerous plan of forcing India to reconsider the closed subject of Kashmir. Early in 1965 there were border skirmishes between police and military forces in the Rann of Kutch, where there was no clearly demarcated frontier. Pakistan appeared to emerge with a military advantage, and, when the British prime minister, Harold Wilson, suggested that the matter be referred to international arbitration, both countries agreed. Could not this be repeated in Kashmir?

Bhutto persuaded Pres. Ayub Khan to agree to a plan whereby irregular forces would be infiltrated into Kashmir to promote a guerrilla war, which would expand to the point where the United Nations would intervene. The guerrilla campaign failed, however, and India retaliated with an action on the 1949 cease-fire line. Pakistan moved an armoured force to cut off Indian supply lines in Kashmir, and on September 6, 1965, the Indian Army sought to relieve the pressure by crossing the Punjab frontier in a three-pronged attack on Lahore. The big powers used their persuasion to end the combat. Superior Indian strength would probably have become decisive had not China created threatening incidents on the Himalayan border in Sikkim. On September 22 a cease-fire came into effect.

The position was debated at the UN, and Indians were startled to discover that many of the Muslim countries of the Middle East, Asia, and Africa saw them as the aggressors. At last the Soviet Union persuaded Ayub to meet Shastri at Tashkent in January 1966. After enormous effort, Soviet Premier Aleksey Kosygin induced the two sides to sign a peace agreement based on a complete return to the status quo and a vague undertaking to negotiate differences. Shastri unexpectedly died at Tashkent.

The Tashkent Conference (1966)

Whereas for years before Nehru died the question of his successor had been under discussion, Shastri's death caught everyone unprepared. One candidate for the premiership was Morarji Desai, a veteran Gandhian and a believer in financial orthodoxy. Support, however, polarized around Indira Gandhi, Nehru's daughter, who secured the backing of three-quarters of the members of the Congress parliamentary party (the elective body).

INDIRA GANDHI'S PREMIERSHIPS

The first three terms. The election of 1967, held in February, was preceded by signs that the economy was in a critical condition. A food shortage of unprecedented proportions, with semi-famine conditions in several states, forced India to appeal to the United States and other grain-growing countries to meet a deficit of up to 15,000,000 tons, and inflation and poor industrial performance had led to devaluation of the rupee in June 1966.

In the Lok Sabhā, Congress gained only 283 of the 520 seats. The largest opposition groups emerged on the right: Swatantra (44 seats) and Bharatiya Jan Sangh (35 seats). The Communists were split into the so-called "Moscow," or right, Communists (23) and the "Peking" Communists (19). The two Socialist parties jointly numbered 36, and the Madras state returned a solid block of 25 DMK members. Congress retained control of half the states but decisively lost in three and in the Delhi union territory.

Following the 1967 election, political turmoil both at the state level and in the Lok Sabhā resulted in a major realignment in the Congress Party. Gandhi's opponents in the party attempted to end her premiership, and Congress was split into opposing and supporting groups. Gandhi declared that "her" Congress (hereafter designated "ruling") would move toward radical policies such as land reform and the nationalization of banking. Deciding to seek electoral support, she scheduled a general election for March 1971.

The result was her personal triumph. The ruling Congress won 352 of the 440 seats contested by the party, giving it a two-thirds majority in the Lok Sabhā and enough to carry through desired constitutional amendments. The Old Congress dwindled to 16 seats; Swatantra crept up to 8; and the Samyukta Socialists fell to only 3. Bharatiya Jan Sangh, with 22 members, was the only right-wing party to maintain some continuity, though in Delhi and North India generally its hold collapsed, leaving some strength in Madhya Pradesh, its original home. The DMK returned with 23 members (previously 25); the Communist Party of India (CPI) held firm with 23; and the Moscow-aligned Communists made a distinct impact, winning 25.

Indira Gandhi's triumph

On August 9, 1971, a treaty of peace, friendship, and cooperation was signed with the Soviet Union, but much of 1971 was dominated by a controversy with Pakistan over the status of its eastern province. Both India and Pakistan declared a state of emergency in November, fighting spread to Kashmir, and open warfare existed on both fronts by December 3. The war ended when Pakistani troops in Dhākā surrendered to Indian and Bangladesh (East Pakistan) forces on December 16. All Indian troops were withdrawn by March 12, 1972, from Bangladesh, which India supplied with generous aid. In June 1972 Gandhi and Bhutto met at Simla, India. The resulting agreement envisaged the bilateral solution of existing problems, and the cease-fire line of December 17, 1971, was accepted.

Thereafter, Gandhi's popularity began to wane. In 1974 the veteran socialist Jaya Prakash Narayan led a campaign against her alleged despotic government that attracted mass support, and in June 1975 her election to the Lok Sabhā of 1971 was declared void. Gandhi persuaded the president to declare a state of emergency throughout India on June 26. Hundreds of politicians, including Narayan and Morarji Desai, an opposition leader in the Congress Party, were jailed. The press was harshly censored, and strikes were declared illegal. The sweeping powers Gandhi assumed were employed to stop inflation and to improve the administration of public services. She amended the constitution to strengthen her position. Many of the emergency measures were popular with the public, but the increasing power of Gandhi's son Sanjay was questioned.

Indira Gandhi's fall

In January 1977 Gandhi suddenly announced that Lok Sabhā elections would be held in six weeks. The emergency was relaxed and her political opponents were freed. Many surmised that her intention was to obtain an electoral mandate to invest Sanjay Gandhi with formal powers. It was expected that the election results would be manipulated, but no attempt was made. The main opposition groups formed a coalition known as the Janata (People's) Party, with Morarji Desai as the leading contender and Jaya Prakash Narayan as the voice of the nation's conscience. One of Gandhi's closest colleagues, Jagjivan Ram, leader of the Harijans (the former untouchables), formed his own opposition party, the Congress for Democracy.

The Janata Party interlude. Once again, Indira Gandhi flung herself unsparingly into the fight, but this time she lost, even in her own constituency. The Janata candidates won 295 seats, the Congress for Democracy 29, and the Congress 153 (of which 92 came from the four southern states where Gandhi's popularity remained undimmed).

Immediately after the results were declared, Gandhi resigned, and it appeared that her political career had ended. Congress members chose another party president. Morarji Desai became prime minister on March 24, and Neelam Sanjiva Reddy was elected president in July.

The new government established a commission to investigate alleged excesses during the state of emergency. Attempts to prosecute Gandhi failed, however, and appeared only to enhance her popularity. On February 25, 1978, state elections held in southern India were contested by a rump Congress Party that acknowledged Gandhi as its leader and called itself the Congress (I), "I" standing for Indira. Her candidates gained a majority in the states of Karnātaka and Andhra Pradesh. She was once again the most talked-about politician in India, and in November she was reelected to the Lok Sabhā. She was expelled the next month, however, and was imprisoned for one week.

The Janata leadership was openly divided over Gandhi's continuing challenge. Tension within the party increased steadily, and Desai found himself more and more alienated from his ministers. Many of the issues he attacked earned him new enemies. In July 1979 he was forced

to resign and was succeeded by Charan Singh, who had been one of his chief opponents. On August 20, however, Charan Singh resigned and asked for new elections.

Indira Gandhi's return. The end of 1979 was a time of political confusion at both the union and state levels. In January 1980, however, the Congress (I) Party won 351 of the 525 contested seats in the Lok Sabhā elections, and no opposition party won more than 10 percent of the seats. Gandhi was again sworn in as prime minister on January 14. Her party continued to win in the midterm state elections held in May, the biennial elections for the Rājyā Sabhā (upper house of Parliament) in July, and state by-elections in November. Legal cases that had been brought against Indira Gandhi, Sanjay Gandhi, and associates were withdrawn. The Prime Minister's triumph was marred, however, by the death of Sanjay Gandhi in an airplane crash on June 23.

Indira Gandhi again dominated India's national life. Her administration faced militant conflict in the 1980s when Sikh extremists agitated for the creation of an autonomous Sikh state in Punjab. In June 1984 the army attacked the Sikhs' holiest shrine, the Golden Temple, in Amritsar, where arms had been stockpiled. More than 400 Sikhs and 100 soldiers were killed. In October Gandhi was assassinated by two Sikh members of her personal guard, and rioting directed against Sikhs broke out across India. Gandhi's elder son, Rajiv Gandhi, was immediately appointed prime minister and swiftly quelled the rioting.

Rajiv Gandhi commanded unprecedented support in the Lok Sabhā after elections in December gave Congress (I) 75 percent of the seats. While Gandhi pledged to follow his mother's policies, he attempted a conciliatory approach toward domestic conflict. Through the late 1980s, however, India continued to be plagued by unrest in many states, including Punjab, and among Muslims fighting for independence in Kashmir and Jammu.

Gandhi was increasingly challenged with allegations of widespread government corruption. Especially serious were charges that Indian officials had received kickbacks from a 1986 arms deal with the Bofors Co. of Sweden. While there was no evidence that Gandhi was involved, his defense minister, Vishwanath Pratap Singh, resigned in 1987 after charging that the government was not supporting his investigation of the scandal. By 1989 Singh had become leader of the centrist Janata Dal (People's Party), which included the Janata Party and formed the largest faction of the Rashtriya Morcha (National Front), a coalition of seven centrist and regional parties that agreed to establish a unified opposition to Congress (I). In Lok Sabhā elections in November, Congress (I), hurt by the charges of corruption and recent price increases of essential goods, gained only 192 seats and lost its majority. The Janata Dal, with 141 seats, became the second largest party, and Singh, by forming a minority government supported by both right-wing and Communist parties, became prime minister.

For later developments in the history of India, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (H.R.T./Ed.)

NORTH INDIA

Chandīgarh

The union territory of Chandīgarh, constituted on November 1, 1966, contains the city of Chandīgarh and the immediate surrounding region. The total area is 44 square miles (114 square kilometres). The territory is a few miles south of the Siwalik Range in the Indo-Gangetic Plain between the Sukhna and Patiāli streams, 151 miles (243 kilometres) north-northwest of Delhi by rail.

The city of Chandīgarh is the capital not only of the union territory of Chandīgarh but also of the states of Punjab and Haryāna; it is also the seat of the High Court of Punjab and Haryāna and the site of Punjab University serving both states. The city is to become the capital of just Punjab, and a separate capital is to be built for Haryāna.

Design of Chandīgarh

The city, with a projected population of 500,000, was planned by the French architect Le Corbusier in collaboration with Indian architects to replace the former capital of Lahore, lost to Pakistan when Punjab was divided upon the partition of India in 1947. Construction began in 1950, and the broad layout of the city was completed by the early 1960s. The city of Chandīgarh is divided into 36 rectangular sectors, 30 of which are residential. Each sector has an area of about 250 acres (100 hectares) and a housing capacity of 15,000 and is self-contained in civic amenities. The grid of the roads is planned so that the fast-moving traffic is segregated from the internal traffic of the sectors. The High Court, nine-storied Secretariat, the Legislative Assembly chamber, and the Rāj Bhavan (governor's house) are all in the northern portion of the city. To the northeast is the Sukhna, a large artificial lake. In the southeast is the sector allotted for industrial establishments, separated from the city by a 60-acre green belt.

More than half the territory is irrigated and produces such major crops as wheat, corn (maize), and potatoes. Forests cover more than a quarter of the total area. Industries in Chandīgarh produce a wide range of products from textiles and furniture to pharmaceuticals and electronic gear.

Delhi

The union territory of Delhi, which covers an area of 573 square miles (1,485 square kilometres), consists of the capital city of Delhi, a few towns, and about 250 villages. Occupying a central position in the Indian subcontinent as well as the entire Eastern Hemisphere, Delhi is equidistant from the great commercial cities of Calcutta and Bombay and a pivotal point in the network of national railways, highways, and airways.

PHYSICAL AND HUMAN GEOGRAPHY

The site of Delhi is a natural fortress in the shape of a triangle. The western side is formed by the famous Delhi Ridge, which is the northernmost extremity of the Arāvalli Range. The eastern side is formed by the Yamunā River, which meets the Delhi Ridge at Wazīrābād in the north and forms the apex. The base is an eastern offshoot of the ridge, branching off from the main body near Mehrauli.

Delhi's administration

The territory is administered directly by Parliament through the home ministry. The chief administrator is the lieutenant governor, who is assisted by four executive councillors appointed by the president of India. The municipal functions are covered by three local bodies: the Delhi Municipal Corporation, the New Delhi Municipal Committee, and the Delhi Cantonment Board.

The area is famous for its handicrafts, particularly ivory carving, jewelry, gold- and silver-thread embroidery, metalwork in brass and copper, and pottery. The main industries are cotton milling, handloom and power-loom weaving, flour milling, and the manufacture of hosiery, wire netting, rubber and leather goods, paints and varnishes, and vegetable oil. Cottage industries have developed considerably since the partition in 1947, and there are a large number of work centres mostly run and managed by women.

Agriculture mainly consists of growing millet, wheat, barley, sugarcane, fruits, and vegetables.

HISTORY

Within the triangular area of the Delhi Territory at least 15 different cities are believed to have emerged at different times. All but the last two, Old and New Delhi, have perished. Each conqueror, having destroyed the city of his predecessor, built a new one on the triangular plain, and so Delhi successively changed its location, its character, and even its name, yet preserved its continuous existence.

In modern times, the district of Delhi was first constituted in 1819 in the Northwestern province of India. In 1858 it was transferred to the Punjab, but Delhi remained the divisional headquarters until 1912, when it became

the capital of British India and a province with a chief commissioner under the direct control of the central government. In 1952 it was made a state; and, on the reorganization of the states four years later, it was constituted as a Union Territory (November 1, 1956). (Ed.)

Haryāna

Haryāna, a state of the northern region of the Indian Union, was constituted November 1, 1966, as a result of the partition of the former state of Punjab into two separate states—Punjābī-speaking Punjab and Hindī-speaking Haryāna. It is bounded on the northwest by Punjab, on the north and northeast by Himāchal Pradesh, on the east by Uttar Pradesh and the union territory of Delhi, and on the south and southwest by Rājasthān.

The reorganization of the former Punjab followed agitation by the Sikhs for a Punjābī Suba (Punjābī-speaking province); but it also met, substantially, the aspirations of the people of the Hindī-speaking region of Punjab for a Vishal Haryāna (Greater Haryāna). The name means "the abode of God," from Hari (the Hindu god Vishnu [Viṣṇu]) and *ayana* (home), although it has also been suggested that the derivation may be from the word *hari* ("green"), denoting the fertility of the countryside.

The area of the state is 17,074 square miles (44,222 square kilometres). The headquarters of the state government are at Chandīgarh, the modern capital of Punjab.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Most of Haryāna lies on the Indo-Gangetic Plain, but in the northeast there is an extension of the low, sub-Himalayan Siwālik Hills. The state, on the whole, is flat, the average height ranging from 700 to 900 feet (210 to 270 metres) above sea level. It is drained by one perennial river, the Yamuna (Jumna), which does not flow in the state but forms its eastern border with Uttar Pradesh. There are numerous seasonal streams, the most important being the Ghaggar, which flows out of the Siwālik Hills and forms its northwestern border with Punjab.

Rainfall is scanty in most areas. Although the state has a system of canal irrigation, there are chronic drought-prone areas and occasional floods from the tributaries of the Yamuna and the Ghaggar. The climate is hot in summer and markedly cold in winter. The maximum temperature in summer (May–June) goes up to 114° F (46° C). The minimum temperature of 28° F (−2° C) occurs in January.

The soils are deep and fertile except in the eroded lands of the northeast and in the southwest where, bordering the Thar Desert in Rājasthān, they are sandy. Out of a total area of 10,930,000 acres, about 10,000,000 acres are classified as arable. Canal and tube well irrigation is available for about 7,000,000 acres, or a little more than two-thirds of the arable land.

The land, as indicated above, is mostly arid or semi-arid, and the natural vegetation, although little remains, reflects that fact, the flora of the plains resembling that of Iran, Saudi Arabia, and North Africa. The largest indigenous trees are the shisham (*Dalbergia sissoo*) and the kikar (*Acacia arabica*).

Dominance of Hindus in the population

The people. Haryāna was constituted as a Hindī-speaking state. Hindus constitute almost 90 percent of the population. Other principal religious groups are Muslims, Sikhs, and Jainas.

The principal language spoken is Hindī. The other important languages are Punjābī and Urdū.

More than three-fourths of the population is rural, living in some 7,000 villages. The smaller urban population is spread over some 60 cities and towns. Ambāla, together with its army cantonment, is the largest town in the state. It is followed in size by Farīdābād, Karnāl, Rohtak, Yamunānagar, Pānīpat, Hissār, and Bhiwāni.

The economy. The economy is agricultural and, despite its pockets of chronic drought and scanty rainfall, Haryāna is one of the few states in which substantial investments in agriculture, irrigation, fertilization, and related programs embodied in India's five-year plans have made a tremendous impact. The spurt in food-grain production (mainly wheat) has led to a significant growth in per capita income. Except in the northeast, most of the land is used for two crops annually. In addition to wheat, the principal winter crops are gram, barley, and mustard. The principal summer crops are millet, rice, corn (maize), sugarcane, and cotton. The state is also known for the quality of its bullocks and dairy cattle.

Agricultural programs

To increase agricultural production, the state has initiated a number of irrigation schemes to tap subsoil water and has undertaken several small and medium-size irrigation projects. The state has been energizing tube wells (driven wells) to augment irrigation facilities.

The state has no heavy industry; nor has it any sizable mineral resources for exploitation. Light industries, however, have been developing, especially since the late 1960s; by the early 1980s Haryāna had become one of the leading states in India in the production of small machinery and spare parts. The increase in agricultural production, especially of sugarcane and cotton, has stimulated related industries, and there were several cotton textile and sugar mills in the state. Other light industries include the manufacture of farm tools, machine tools, electrical and glass goods, cement, paper, and bicycles. Industrial areas with medium- and small-scale factories and other enterprises have developed at Farīdābād and Ballabhgarh, south of Delhi (where Haryāna borders Uttar Pradesh), and at Sonepat, Jagādhri, and Yamunānagar. Industrial areas are also rising up along the Grand Trunk Road to and from Delhi and along the main line of the Northern Railway between Ambāla and Delhi to take advantage of transport facilities.

Administrative and social conditions. *Government.* The governor, appointed by the president of India, is the head of the state. He is aided and advised by a Council of Ministers responsible to the state Legislative Assembly (Vidhān Sabhā). Normally, election to the assembly is for a five-year term. The state has a common High Court with Punjab. The Haryāna Public Service Commission recruits all government employees.

The state is divided into twelve districts—Hissār, Rohtak, Gurgaon, Karnāl, Ambāla, Jīnd, Farīdābād, Sirsa, Kurukshetra, Sonepat, Bhiwāni, and Mahendargarh. The system of village self-government (*pañcāyat raj*) has been extended to all of the villages. Administration, as is the case in the rest of the country, is oriented toward development in all sectors.

Education and welfare. With education getting a high priority in the development program, there has been a large quantitative expansion in school enrollment in Haryāna; however, as in the northern states in general, girls' education has not made much headway. The literacy rate is lower than in the southern states of India. The state has more than 100 teaching colleges; the Kurukshetra University (founded 1956) at Kurukshetra also has a regional engineering college. Maharshī Dayānand University at Rohtak has operated since 1976. Hissār is the seat of the Haryana Agricultural University.

Haryāna has a network of district and subdivisional hospitals and primary health centres. The Family Welfare Programme is making headway. In more than half the villages, drinking water is acutely scarce. A large minority belonging to scheduled castes (former "untouchables") and other backward classes are covered by schemes of assistance in education and technical guidance and for housing.

Road development has been accorded high priority in the state's development plans. Haryāna Roadways, which is a nationalized transport undertaking, operates hundreds of vehicles on local and interstate routes.

The chief lines of communication are those leading to and from Delhi, particularly the historic Grand Trunk Road and the mainline of the Northern Railway between Ambāla and Delhi, which is the centre of the nation's rail, road, and airway networks.

Cultural life. Harȳana is a land of legends, temples, historical places, and archaeological sites. Besides the Hindu pilgrim centre of Kurukṣetra, site of a bathing fair at each solar eclipse, the state has several sacred and historic sites including Karnāl, site of many events in the epic of the *Mahābhārata;* Thānesar, an early (9th–12th century) Hindu capital; and Gurgaon, capital of the Gurjara

dynasty before the 10th century. Rohtak was the centre of the Yaudheya warriors in the 5th and 6th centuries. Religious centres with sacred tanks for bathing and other purposes include Agroha, Pandara, Pehowa, and Ramrah. Surajkund, an ancient centre of sun worship, has a sacred tank built in the shape of the rising sun. Pānīpat, mentioned in the *Mahābhārata,* was the scene of decisive dynastic battles in the 16th and 18th centuries.

The Indus Valley civilization was crystallized at Khokhra Kot (Rohtak), Naurangābād, Meham, and Agroha in Haryāna. The Gurukul, an educational and archaeological institute, at Jhajjar has a rich collection of ancient coins, molds, seals, inscriptions, utensils, and statues, which throw light on the life of people inhabiting the region at different periods.

The people of Haryāna are simple and hardworking and have preserved their old religious and cultural traditions and customs. They hold fairs and festivals with great enthusiasm, among the most notable being the Janam Ashtami (Lord Krishna's birthday) Fair at Bhiwāni and the Masani Fair at Gurgaon. Folk theatre and folk dancing have also been very popular. *Saang* is the most popular variety of performances with the participants—the female roles played by men—singing and dancing continuously for five to six hours. The themes are drawn from mythology or Sanskrit literature and occasionally include love lores of Punjab.

Essentially a rural state, more than three-fourths of the people live in the villages, where their principal occupation is agriculture and animal husbandry. At the same time, the Indian army, traditionally a favourite career for young men, finds Haryana a major recruiting ground.

HISTORY

The region now known as Haryāna was the birthplace of the Hindu religion and the civilization of the Indus Valley. There, the first hymns of the Aryans were sung, and the most ancient manuscripts were written. On the battlefield at Kurukshetra, Lord Krishna (Kṛṣṇa) delivered to the warrior Arjuna the teachings that are contained in the *Bhagavadgītā,* venerated by Hindus as the highest code of ethics.

The early Aryan settlers

The earliest known settlers of Haryāna were Aryans. Lying athwart the route of overland invasions into India, it underwent successive waves of migration from the time of Alexander the Great (326 BC) and was the scene of many famous battles of Indian history, including the battles of Pānīpat in 1526 (at which the Mughal leader Bābur defeated Ibrāhīm Lodī and established Mughal rule in India), 1556 (at which the Afghan forces were defeated by the army of Mughal emperor Akbar), and 1761 (at which Aḥmad Shāh 'Abdāli decisively defeated the Marāṭhās, paving the way for British control in India) and the Battle of Karnāl in 1739 (at which Nāder Shāh of Persia dealt a blow to the crumbling Mughal Empire).

The area included in the present state was ceded to the British East India Company in 1803. In 1832 it was transferred to the then North-Western Provinces. In 1858 Haryāna became a part of Punjab, remaining as such after the partition of India and Pakistan in 1947. The demand for a Haryāna province, however, was raised even before India's independence in 1947. Lala Deshbandhu Gupta and Asaf Ali, prominent figures in the national movement, advocated a separate state of Haryāna. Sri Ram Sharma, a veteran freedom fighter, had headed a Haryāna Development Committee to focus attention on the concept of an autonomous state. The demand for unilingual states by Sikhs and Hindus gained momentum in the early 1960s, and Haryāna became the 17th state of the Indian Union in accordance with the recommendations of the States Reorganization Commission and the provisions of the Punjab Reorganization Act of 1966.

Himāchal Pradesh

Occupying a region of scenic splendour in the western Himalayas, Himāchal Pradesh is a state of the Indian Union at the extreme north of the Indian subcontinent. It is bounded on the north by Jammu and Kashmir, on the east by China (Tibet), on the southeast by Uttar Pradesh, on the south by Haryāna, and on the west by Punjab. Towering snow-clad mountains are divided by deep valleys with thick woods, green fields, and cascading streams, as well as lakes and waterfalls. Himāchal means Snowy Mountain (*hima,* "snow"; *acal,* "mountain"), the state (*pradesh*) taking its name from the Himalayas.

The physical setting

The area of the state is 21,495 square miles (55,673 square kilometres). The capital is Simla, at an elevation of about 7,100 feet above sea level, the largest and most popular hill resort in India and the summer headquarters of pre-independence British viceroys.

Formerly a union territory, Himāchal Pradesh emerged as a state of the Indian Union on January 25, 1971.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The terrain varies with hills of low and high altitudes, wooded valleys, flowing rivers, and rocky soils. The mountains rise to about 22,000 feet and include the Himalayan ranges of the Pīr Panjāl, Hathi Dhār, and Dhaula Dhār. The mountain valleys are known locally as *dūns* (*e.g.,* the Kiarda Dūn in the Sirmūr district); the hill stations of Kulu, Manāli, and Dharmsāla are in the Kulu and Kāngra valleys. The major rivers are the Chenāb (Chandrabhāga), Rāvi, and Beās in the west and the Sutlej and Yamuna (Jumna) in the east.

The climate ranges from mild to cold. Icy winds sweep the state, and the hills and valleys remain covered with snow for months. In the valleys, flowers and fruits grow in abundance, and their rivers abound in fish. The forests harbour wildlife, including the ibex, goral, bear, and snow leopard.

The people. The population presents a mixture of hill tribes, including Gaddis, Gūjars, Kinners, Lahaulis, and Pangwalas. The vast majority of the people are Hindus, who worship in temples dotted throughout the state, but there are also Buddhists, Sikhs, Muslims, and Christians. More than 60 dialects are spoken, the main languages being Hindī and Pahāṛī. Pahāṛī is derived from Sanskrit and Prākrit. Caste distinctions and outdated social customs are slowly fading under the impact of influences from outside, the spread of education, and the growing consciousness of changes taking place throughout the country.

Himāchal Pradesh is the least urbanized state in India, the urban population accounting for less than 10 percent of the total. There are about 40 towns, of which only the capital, Simla, has a population exceeding 50,000. Other towns are Sundarnagar, Mandi, Nāhan, Chamba, Solon, Dharmsāla, Bilāspur, Kāngra, Dalhousie, Una, Kulu, Palampur, and Kasauli. More than 90 percent of the people depend for their livelihood on farming, horticulture, and livestock raising.

The economy. Himāchal Pradesh, with its perennially snow-fed rivers, has the potential to produce about 20 percent of the nation's total hydroelectric power. There is immense scope for the development of forests and for setting up of forest-based industries, including newsprint. Hydroelectric development began with installations in Mandi district on the Uhl River, a tributary of the Beās, and in Bilāspur district at the Bhākra Dam (1963) across the Sutlej. The Giri Project, the first hydroelectric project initiated by the state, started commercial production during 1978–79.

Economy based on the land

The state's economy is still almost entirely based on the products of the land. The chief crops are wheat, corn (maize), barley, rice, and potatoes. One of the top fruit and vegetable producing states of India, Himāchal Pradesh produces excellent varieties of plums, peaches, apricots, pomegranates, ginger, and mushrooms. The state's high-quality apples are exported to neighbouring states and abroad. Sheep and goat rearing are widespread, and wool and pashm (the fleece of the Tibetan goat) are important exports. The coniferous forests have also been a source of revenue.

There is very little industry. A foundry was established at Nāhan in 1872; taken over by the Indian government in 1952, it manufactures agricultural implements. There are also resin and turpentine factories at Nāhan and Bilāspur, and breweries at Solan and Mechtapur. There is a tele-

vision assembly plant at Solan, and a cement factory at Rajban. Some woollen goods are made on handlooms and small machines, and several small plants process foods, notably fruits.

Concentrated efforts have been made to improve the state's economy through a series of development plans. The outlook for the economy is regarded as hopeful in view of the prospects of exploitation of the abundant power potential, mineral deposits, and forests and the promotion of tourism.

Except for the rail line from Kālka (in Haryāna) to Simla, the capital, and the narrow-gauge track connecting Pathānkot (Punjab) and Jogindarnagar (H.P.) through the Kāngra valley, there are no railways or waterways in the state. Roads are the communications lifeline of Himāchal Pradesh. The nationalized transport system operates more than 140 routes in the state.

Administrative and social conditions. *Government.* The head of the state is the governor, appointed by the president of India. The council of ministers, headed by a chief minister, is responsible to the legislative assembly that is elected directly on the basis of adult suffrage. The state has four representatives in the lower house (Lok Sabhā) of India's Parliament and three in the upper house (Rājya Sabhā). The state is divided into ten districts—Bilāspur, Chamba, Hamīrpur, Kāngra, Kinnur, Kulu, Lahul and Spiti, Mandi, Simla, Solan, Sirmūr, and Una.

Education and welfare. In the second half of the 20th century, there has been expansion of education and public health facilities and an improvement of communications. The people, however, remain at the subsistence level, and the state's vast natural resources have yet to be tapped systematically. Development plans lay emphasis on road building, power generation, and exploitation of forest products for industrialization.

Growth of educational institutions

Since 1948, education has made rapid strides. The state has more than 25 degree-granting colleges connected with Himāchal Pradesh University at Simla and Punjab University as compared with one in 1950. There is a medical college at Simla and an agricultural university at Palampur; research is carried on at the Indian Institute of Advanced Study (Simla) and the Central Research Institute (Kasauli). School enrollment has gone up in the wake of a phenomenal rise in the number of primary, middle, and higher schools. In addition to the few hundred hospital-dispensaries and primary health centres, there are several family-welfare centres engaged in the campaign to check the birth rate. Progress has been made in providing drinking water to the people in the remote areas.

Cultural life. The fairs and festivals of the hill people are occasions of joyful song and dance. Besides folk dances, there are folk songs with romantic themes and folktales woven around mythological figures. Both men and women partake in the dance and music and in the folk dramas.

Kāngra is a centre of paintings. In addition to this 18th-century school of painting, Himāchal Pradesh can boast of beautiful and useful crafts, such as the exquisitely designed shawls of Kinnaur and the embroidered handkerchiefs of Chamba.

The Simla hills, Kulu and Manāli valleys, and Dalhousie are tourist attractions. Skiing, golfing, fishing, trekking, and mountaineering are activities for which Himāchal Pradesh is ideally suited.

Pilgrims journey to the Kulu and Kāngra valleys to worship at shrines of legendary antiquity. From Kāngra, one can proceed to Dharmsāla, a pleasant hill station that is the home of the Dalai Lama, who fled from Tibet in 1959 in the wake of Chinese occupation of Lhasa. The Kulu Valley is known as the Valley of the Gods; its pine and deodar forests, flower-spangled meadows, and fruit orchards provide the setting for the colourful Dussehra fair held every autumn. The temple gods are taken in caparisoned palanquins from the hilltops, and through singing and dancing the hill people give vent to their joy.

HISTORY

The history of the area dates to the Vedic period, the earliest known tribal group being Dāsas, who had a warlike reputation. Later the Aryans assimilated the tribes. The area was exposed to successive invasions through the centuries, ending with the British dominion over the native princes of the hill states.

Himāchal Pradesh was constituted as an administrative unit comprising 30 hill states in 1948. This event was preceded by a movement for the end of feudalism and the introduction of responsible government. One of the princely states, Suket, virtually surrendered to the peaceful demonstrators, hastening the process of change.

Achievement of statehood

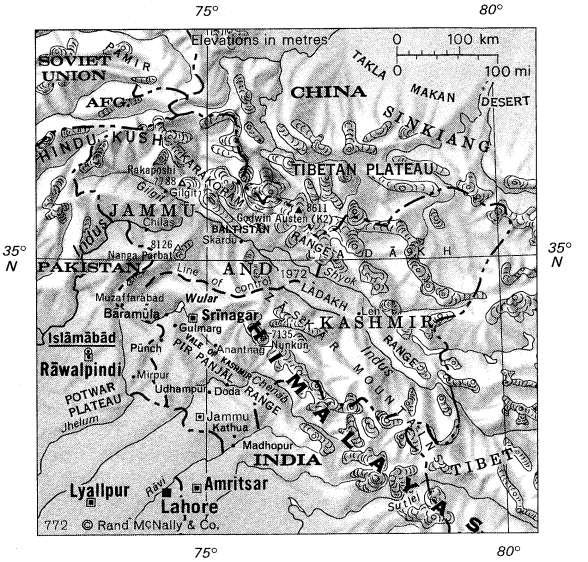
Between 1948 and the achievement of statehood in 1971, Himāchal Pradesh went through various changes in size and administrative setup. First a substate and then a union territory directly administered by the central government, Himāchal Pradesh was enlarged in 1954 by the merger of Bilāspur (a former Indian state and then a chief commissioner's province), and again in 1966 by the merger of Punjab hill areas including Simla, Kāngra, Kulu, and Lahaul and Spiti districts, and parts of Ambāla, Hoshiārpur, and Gurdāspur districts. Y.S. Parmar, who since the 1940s had led the hill people of Himāchal Pradesh in the quest for responsible government, became the first chief minister of the new state. (C.Ra.)

Jammu and Kashmir

The territory of Jammu and Kashmir, situated in the northern part of the Indian subcontinent in the vicinity of the western Himalayan mountain range, has been in dispute between India and Pakistan since the partition of India in 1947. Formerly one of the largest princely states of India, it is bounded on the northeast by the Sinkiang-Uighur Autonomous Region and Tibet (the Tibetan Autonomous Region of China), on the south by the Indian states of Himāchal Pradesh and Punjab, and to the northwest and west by a Pakistani-occupied portion of the territory.

The territory claimed by India has an area of 85,806 square miles (222,236 square kilometres); it is divided by a "line of control," agreed on in 1972, that leaves an area of 46,976 square miles in the north and west to Pakistan and the remainder, amounting to 38,829 square miles, under Indian control. Repeated United Nations attempts to settle the dispute by holding a plebiscite have been unsuccessful; a plebiscite has not been held, and the dispute has been further complicated by Chinese incursions into the territory from the north.

The administrative capital of the Indian sector is Srīnagar in summer and Jammu in winter; the administrative capitals of the three Pakistani-claimed areas are at Muzaffarābād, Gilgit town, and Skārdu, respectively. The Vale of Kashmir, located in the Indian sector, is renowned for its beauty and fertility.



JAMMU AND KASHMIR

PHYSICAL AND HUMAN GEOGRAPHY

The land. Physiographically, the territory comprises a number of zones that are closely associated with the structural components of the western Himalayas. From southwest to northeast these zones consist of the plains, the foothills, the Pīr Panjāl Range, the Vale of Kashmir, the central Himalayan zone, the Upper Indus Valley, and the Karakoram Range.

The seven geographic zones

The plains. The narrow zone of plains country in Jammu and Pūnch is characterized by interlocking sandy alluvial fans (fan-shaped accumulations of debris) that have been deposited by streams discharging from the foothills and by a much-dissected pediment (eroded bedrock surface) covered by loams and loess (wind-deposited silt) of the Pleistocene Epoch (about 10,000 to 2,500,000 years old). Rainfall is low, amounting to about 15 or 20 inches (380 to 500 millimetres) a year, and it occurs mainly in the form of heavy but infrequent showers in the summer (June to September), when the monsoon winds blow. The countryside has been almost entirely denuded of trees, and thorn scrub and coarse grass are the dominant forms of vegetation.

The foothills. The foothills of the Himalayas, rising from 2,000 to 7,000 feet comprise an outer zone and an inner zone. The outer zone consists of sandstones, clays, silts, and conglomerates, influenced by Himalayan folding movements and eroded to form long ridges and valleys (*dūn*s). The inner zone consists of more massive sedimentary rock, including red sandstones of the Miocene Epoch (about 7,000,000 to 26,000,000 years old), that has been folded, fractured, and eroded to form steep spurs and plateau remnants. River valleys are deeply incised and terraced, and faulting has produced a number of alluvium-filled basins, such as those of Udhampur and Pūnch. Rainfall increases with elevation, and the lower scrubland gives way to pine forests at higher altitudes.

The Pīr Panjāl Range. The Pīr Panjāl Range constitutes the first mountain rampart associated with the Himalayas. It has an average crest line of 12,500 feet, with individual peaks rising to 15,000 feet. Consisting of an ancient rock core of granites, gneisses, quartz rocks, and slates, it has been subject to considerable uplift and fracturing and was heavily glaciated during the Pleistocene Epoch. The range receives considerable winter snowfall as well as summer rain and has extensive areas of pasture above the tree line.

The Vale of Kashmir. The Vale of Kashmir is a deep, asymmetrical basin lying between the Pīr Panjāl and the central Himalayan ranges at an average altitude of 5,300 feet. During the Pleistocene Epoch it was occupied at times by a lake, known as Lake Karewa, and is now filled by lacustrine sediments and alluvium deposited by the Upper Jhelum River. Soil and water conditions vary considerably, and the climate is characterized by an annual rainfall of about 30 inches, some of which is derived from the summer monsoon winds and some from winds associated with winter depressions. Temperatures are considerably modified by altitude; average minimum temperatures of about 11° F (−12° C) occur in January and average maximum temperatures of 99° F (37° C) in July. Up to 7,000 feet, woodlands of deodar (Indian cedar), blue pine, walnut, willow, elm, and poplars occur; from 7,000 to 10,500 feet coniferous forests with fir, pine, and spruce are found; from 10,500 to 12,000 feet birch is dominant; and above 12,000 feet there are meadows with rhododendrons and dwarf willows as well as honeysuckle.

Forests and woodlands in the Vale of Kashmir

The central Himalayan zone. Geologically complex and topographically immense, the central Himalayan zone has ranges reaching more than 20,000 feet and deeply entrenched, remote valleys. The region was heavily glaciated in the Pleistocene Epoch, and remnant glaciers and snowfields are still present. The zone receives some rain from the South-west monsoon winds in the summer months, and the lower slopes are forested, but the mountains, nevertheless, constitute a climatic divide, representing a transition from the monsoon climate of the Indian subcontinent to the dry, continental climate of Central Asia.

The Upper Indus Valley. The valley of the Upper Indus is a well-defined feature that follows the geological strike (structural trend) westward from the Tibetan border to the point where it rounds the great mountainous mass of Nānga Parbat to run southward in deep gorges cut across the strike. In its upper reaches it is flanked by gravel terraces; each tributary builds an alluvial fan out into the main valley. The town of Leh in the Indian sector stands on such a fan, 11,500 feet above sea level, with a climate characterized by almost total lack of rainfall, intense insolation (exposure to the Sun's rays), and great diurnal and annual ranges of temperature. Life depends on meltwater from the surrounding mountains, and vegetation is alpine (above the limit of tree growth), growing on thin skeletal soils.

The Karakoram Range. The great granite-gneiss massifs of the Karakoram contain some of the world's highest peaks, including K2, or Mt. Godwin Austen, with an altitude of 28,251 feet (8,611 metres); at least 30 other peaks exceed 24,000 feet. The range, which is still being glaciated, rises starkly from dry, desolate plateaus characterized by extremes of temperature and shattered rock debris. The Karakoram region is aptly named the "roof of the world."

The great mountain peaks

Animal life. Wild animals include ibex, urial (a species of wild sheep with a reddish coat), the rare *hangul* or Kashmir stag found in the Dachhigām game sanctuary, black and brown bears, and many game birds, including vast numbers of migratory duck.

The people. *Rural settlements.* Physiographic diversity is matched by the considerable differences that occur in the forms of human occupation of the territory.

In the plains and foothills, agricultural settlements have resulted from colonization movements from the Punjab extending over a long period; both the people and their culture are similar to those of adjacent parts of the Punjab and other lowlands to the west. Where alluvial soils and the availability of water for irrigation make agriculture possible in the *dūn*s and lower valleys, the population is sustained by crops of wheat and barley, which are gathered in the spring (*rabī'*) harvest, and of maize and millet, gathered in the late summer (*kharīf*) harvest, as well as by the keeping of livestock. The upper sections of the valleys support a sparser population who depend on a mixed economy of corn (maize), cattle, and forestry in which seasonal migration to higher pastures in the spring is necessary for the production of milk and clarified butter, or ghee, for southern lowland markets. In winter the hillmen return to lower areas to work in government-owned forests and timber mills. Agricultural hamlets and nucleated villages predominate, and towns such as Jammu and Udhampur function essentially as market centres and administrative headquarters for the rural populations and estates in the vicinity.

Jammu and Pūnch. Jammu, winter capital of the maharajas and second city of the state, was traditionally the seat of Dogra power and serves a largely Hindu population. Of the total population of Jammu, more than two-thirds are classed as Hindus; of these the greater part live in the southeastern territories. Culturally, ethnically, and linguistically, they are closely related to the Hindus in the Punjab state of India, many speaking the Dogrī language. The majority of the Sikhs also live in Jammu Province. To the northwest, however, the proportion of Muslims increases, with Muslims constituting a dominant majority in western Pūnch. Here cultural and economic ties are closer with Rāwalpindi (the interim capital of Pakistan since 1959) and the Punjab province of Pakistan, and the inhabitants have consistently opposed Hindu–Dogra rule.

Kashmīrīs of the Vale and highlands. The Vale of Kashmir with its surrounding highlands (Kashmir proper) has always retained something of its own individuality. The vast majority of the people are Muslims; culturally and ethnically, links are closest with the northwestern highlands. The Kashmīrī language is influenced by Sanskrit and belongs to the Dardic branch of Indo-Aryan languages also spoken by the various hill tribes of Gilgit in the northwest; physical traits of the population suggest early connections in the same direction.

The Kashmīrī language

Ladākh and Baltistān. Across the main Himalayan divide, the people are as different as the landscape. Ladākh (sometimes called "little Tibet"), to the northeast, is thinly

populated; its inhabitants are of Tibetan racial stock and language (Ladākhī) and hold the Buddhist faith, while Baltistān to the north, with its administrative centre at Skārdu, is inhabited by people of similar stock, the Baltī, who have been converted to Islām.

Agricultural crops

The economy. The majority of the Kashmīrī population is occupied in agriculture of diverse kinds, each adapted to local conditions. Rice, the staple crop, is planted in May and harvested in late September. Corn (maize), millet, pulses (plants such as peas, beans, and lentils), cotton, and tobacco are—with rice—the main summer crops; wheat and barley are the chief spring crops; many temperate fruits and vegetables are grown in areas adjacent to urban markets or in well-watered areas with rich organic soils. In the Vale of Kashmir large orchards produce apples, pears, peaches, walnuts, almonds, and cherries. The Vale of Kashmir (Pampur) is the sole producer of saffron in the Indian subcontinent. Lake margins are particularly favoured, and intensive cultivation of vegetables and flowers is practiced in reclaimed marshland or on artificial floating gardens. The pressure of population on land is everywhere apparent, and all available resources are utilized. The lakes and rivers provide fish, water chestnuts, power (for example, near Bāramūlla, 25 miles east of Srīnagar), and transport and are a tourist attraction. The mountains supply many kinds of timber, pasture for sheep and dairy cattle, and such snow-sports centres as that at Gulmarg, 14 miles south of Bāramūlla.

The region has limited mineral and fossil-fuel resources: small deposits of anthracite coal occur near Jammu; bauxite deposits are in the Udhampur district. Other minerals include limestone, iron ore, zinc, and marble.

All the principal towns, including Leh, and a majority of the villages were electrified, and hydroelectric and thermal generating plants have been constructed to provide power for industrial development based on local raw materials. Major power schemes are at Chineni, Salal, Upper Sind, and Lower Jhelum. Land reforms have been undertaken, grain production was increased, and the quantity of major exports—timber, fruits and dry nuts, and handicraft products—was vastly expanded since 1947. Improvements have also been made in facilities for tourist traffic, although here the potential has largely been untapped. Metalware, precision instruments, sporting goods, furniture, embroidery, matches, and resin and turpentine are the major industrial products of the state.

Srīnagar possesses many specialized agricultural markets, retail shopping centres, and associated industries. The industries are essentially developments from rural crafts—with handloom weaving of local silk, cotton, and wool, carpet weaving, wood carving, leatherwork, and other village industries organized on a factory basis in an urban context. Such industries, together with silverwork and copperwork and jewelry, were stimulated by the presence of the royal court and the tourist trade but also owe something to the important position achieved by Srīnagar in west Himalayan trade. The city acts as an entrepôt for the products of the Punjab on the one hand and of High Asia (the region east of the Karakoram, Pamirs, and Ladākh ranges) on the other; routes run northwestward into Gilgit via the Raj Diangan Pass and northeastward via the Zoji La (pass) to Leh, Ladākh and beyond.

The economy of Ladākh and Baltistān

Cultivation in Ladākh and Baltistān is restricted to such main valleys as those of the Indus and Shyok rivers, where it consists of small irrigated plots of barley, buckwheat, roots, and sown grass situated amid desert surroundings. Pastoralism is a vital feature of the economy since animals are necessary for human survival. Herds of yak and flocks of goatlike sheep provide not only animal products—including cashmere (also called pashmina), which is the downy undercoat of Kashmir goats—but also means of transport between the widely dispersed settlements and monasteries. Grazing rights are subjects of important litigation; small townships are basically caravansaries (inns for the overnight accommodation of caravans) located at points where herders and traders can obtain provisions and grazing prior to undertaking the next stage of their journey through some of the loftiest passes and over some of the most barren terrains on Earth. Political troubles have unfortunately reduced the volume of this transit trade, as well as the benefit that Ladākhī settlements such as Leh once derived from it.

Under successive five-year plans, the Indian union (central) government has made a huge investment in developing highways and communications. As a result of the India–Pakistan dispute over Kashmir, the blocking of the Jhelum Valley route, which runs from Srīnagar to Rāwalpindi, made it necessary to transform the longer and more difficult Banihāl Pass cart road into an all-weather motor road linking Jammu with the Vale. This necessitated the construction of one of the longest tunnels in Asia—the Jawahar Tunnel (1959). Jammu is the terminus of the Northern Railway of India and is also connected with Uri in the north via Srīnagar by highway. Air passenger service exists between Srīnagar and New Delhi.

Administrative and social conditions. Before its division, the territory of the state as a whole was composed of the provinces of Jammu, Kashmir, and the border states—Ladākh, Baltistān, and the Gilgit Agency. Muzaffarābād, Kotli, and part of Mīrpur, and Pūnch districts comprise the state of Azad (meaning "free") Kashmir, which is under the protection of Pakistan, while Baltistān, Astor, and the Gilgit Agency constitute the Northern Areas and are administered directly by Pakistan. The government of Azad Kashmir is assisted by the Pakistan Ministry of Kashmir and Northern Affairs through a chief adviser, who also functions as a resident in the Gilgit and Baltistān agencies. Leh (Ladākh), Anantnāg, Bāramūlla, Srīnagar, Pulwama, Badgom, Kupwara, and Kargil districts of Srīnagar province, and Jammu, Kathua, Udhampur, Rājauri, Doda, and part of Pūnch districts of Jammu province form the Indian-held territory, and are represented in the Indian Parliament as the state of Jammu and Kashmir.

The state of Jammu and Kashmir retains a special status within the union (central) government as defined by Article 370 of the Constitution of India. While the rest of the states follow the Constitution of India, the state of Jammu and Kashmir has its own constitution (adopted in 1956) that affirms the integrity of the state within the Republic of India. The central government has direct legislative powers in matters of defense, external affairs, and communications within the state, and has indirect influence in matters of citizenship, Supreme Court jurisdiction, and emergency powers.

The Indian state. Under the constitution of Jammu and Kashmir, the governor of the state, as in all other Indian states, is appointed by the president of India. Executive power rests in the elected chief minister and the council of ministers. The legislature consists of two houses: the Legislative Assembly of 75 or 76 members, representing constituencies; and the Legislative Council of 36 members. Since the 1967 election, the state directly sends six elected representatives to the Lok Sabhā and two members, nominated by the governor, to the Rājya Sabhā of the Indian Parliament.

The bicameral legislature

The High Court consists of a chief justice and two or more other judges appointed by the president of India.

Education is free at all levels, and there were more than 4,000 educational institutions in the early 1980s. The university at Srīnagar was founded in 1948 and the university at Jammu in 1969. There is an efficient medical service with dispensaries and primary health centres scattered throughout the state. In spite of its high altitude, the Vale of Kashmir is not healthful, and in the past there have been considerable outbreaks of cholera and other infectious diseases, although the incidence of these has been much reduced.

Azad Kashmir. The government, with headquarters at Muzaffarābād, was first set up in 1947. It is under a president, assisted by a State Council. The state has its own High Court but is protected and financed by Pakistan.

Education at its lower stages is free. In addition to primary, middle schools, and high schools, Azad Kashmir has several colleges, at one of which students follow degree courses. There are several hospitals and many more dispensaries and first-aid posts.

The Gilgit Agency and Baltistān. Although both are administered by the government of Pakistan through po-

litical agents, these two areas are governed as separate units because of the difficult mountain terrain and the lack of communications between them. Gilgit town is the headquarters of the Gilgit Agency, and Skārdu is the administrative centre in Baltistān. In the two areas there are a number of primary schools, middle schools, and high schools, as well as several hospitals, dispensaries, and first-aid posts.

The Kashmir problem. After more than a quarter of a century, the "Kashmir problem" has remained intractable. Although there was a clear Muslim majority in the state before the 1947 partition and its economic, cultural, and geographic contiguity with the Muslim-majority area of the Punjab could be convincingly demonstrated, the accidents of history have resulted in a division of territory that has no rational basis. Pakistan has been left with territory that, although basically Muslim in character, was thinly populated, difficult of access, and economically underdeveloped. The largest Muslim group, situated in the Vale of Kashmir and estimated to number more than half the population of the entire state, lies in Indian-administered territory, with its former outlets via the Jhelum Valley route being blocked.

Pakistan has been frustrated in its attempts to use the waters of the Indus, Jhelum, and Chenāb rivers granted to it under the Indus Waters Treaty (1960), since their upper catchments lay in territories outside its control. Indeed, apart from the improvement of communications linking Gilgit and Baltistān with military bases in the Northwest Frontier Province and the extension of administration into these highland areas, not a great deal of development has been possible.

India, on the other hand, has acquired the lion's share of both territory and population and with them substantial linguistic, racial, and religious problems. Only 18 percent of the population in the Indian sector may be classed as literate. There are still movements seeking independence for Jammu and Kashmir from both India and Pakistan. Thus, a strong military presence is necessary in the Indian sector, in part to support the administrative structure of the state, in part to confront the forces of Pakistan along the cease-fire line, and in part to confront Chinese forces on the northeastern frontiers.

HISTORY

Buddhism and Hinduism penetrated Jammu and Kashmir at an early date; from the 9th to the 12th century AD, the region appears to have achieved considerable prominence as a centre of Hindu culture. It was brought under Muslim rule in the 14th century, and it remained so for some five centuries until Sikh and then Dogra (hill Rājput) rulers assumed control in the 19th century. (The Sikhs are a religious community of northern India.)

The state of Jammu and Kashmir, in its contemporary form, dates from 1846, when, by the treaties of Lahore and Amritsar at the conclusion of the First Sikh War, Raja Gulab Singh, the Dogra ruler of Jammu, was created maharaja (ruling prince) of an extensive but somewhat ill-defined Himalayan kingdom "to the eastward of the River Indus and westward of the River Rāvi." For the British the creation of this princely, protected state helped to safeguard their northern flank in their advance to the Indus and beyond during the latter part of the 19th century; the state thus formed part of a complex political buffer zone interposed by the British between their Indian empire and the empires of Russia and China to the north. For Gulab Singh, confirmation of title to these mountain territories marked the culmination of almost a quarter of a century of campaigning and diplomatic negotiation among the petty hill kingdoms along the northern borderlands of the Sikh empire of the Punjab.

Raja Gulab Singh

Boundaries. Some attempts were made in the 19th century to define the boundaries of the state, but precise definition was in many cases defeated by the nature of the country and by the existence of huge tracts lacking permanent human settlement. In the far north, for example, the maharaja's authority certainly extended to the Karakoram Range, but beyond lay a debatable zone on the borders of the Turkistan and Sinkiang regions of Central Asia, and the boundary was never demarcated. There were similar doubts about the alignment of the frontier where this northern zone skirted the region known as Aksai Chin, to the east, and joined the better known and more precisely delineated boundary with Tibet, which had served for centuries as the eastern border of the Ladākh region of what is now Jammu and Kashmir. The pattern of boundaries in the northwest became clearer in the last decade of the 19th century, when Britain delimited boundaries in the Pamir region in negotiations with Afghanistan and Russia. At this time Gilgit, always understood to be part of Kashmir, for strategic reasons was constituted as a special agency in 1889 under a British agent.

Counterclaims of India and Pakistan. As long as the existence of the state was guaranteed by the United Kingdom, the weaknesses in its structure and along its peripheries were not of great consequence; following the withdrawal of the United Kingdom from South Asia in 1947, however, they became apparent. By the terms agreed upon for the partition of the Indian subcontinent between India and Pakistan, the rulers of princely states were given the right to opt for either Pakistan or India or—with certain reservations—to remain independent. The maharaja of Kashmir, Hari Singh, initially believed that by delaying a decision he could maintain the independence of Kashmir, but, caught up in a train of events that included a revolution among his Muslim subjects along the western borders of the state and the intervention of Pashtun tribesmen, he signed an Instrument of Accession to the Indian Union on October 26, 1947. This was the signal for intervention both by Pakistan, which considered that the state was a natural extension of Pakistan, and by India, which intended to confirm the act of accession. Localized war continued during 1948 and was terminated through the agency of the United Nations in a cease-fire, which took effect in January 1949. In July of the same year, India and Pakistan defined a cease-fire line that divided the administration of the territory. This left Jammu, together with part of the districts of Pūnch and Mīrpur, the Vale of Kashmir, and the Ladākh region, under Indian control and western Pūnch (which, together with Muzaffarābād and part of Mīrpur, formed Azad Kashmir), Gilgit, and Baltistān under Pakistan's authority. Regarded merely as a temporary expedient, this partition along the cease-fire line still exists, though warfare between the two contestants was briefly resumed in 1965 and again in 1971, despite the many proposals made to end the dispute.

The 1947 to 1949 hostilities

Chinese interests. The conflicts within the state were heightened by tensions at the periphery when a manifestation of Chinese power created an additional problem. After the Chinese Communist conquest of Tibet and the resurgence of Chinese power in Sinkiang, on the border with the Soviet Union, Chinese forces penetrated into the northeastern parts of Ladākh; by 1956–57 they had completed a military road through Aksai Chin to provide better communication between Sinkiang and western Tibet. The belated discovery of the existence of this road by Indian patrols led to border clashes. The northeastern part of Ladākh has been occupied by China since the Sino-Indian War in October 1962. Indian refusal to negotiate with China on the alignment of the Ladākhi boundary in this area was in contrast with the willingness of Pakistan to reach a settlement with China on the course of the Karakoram boundary that lies within Pakistani-administered Kashmir. (W.Ki.)

Punjab

Punjab, a state in the Indian Union, is situated in northwest India. It is bounded on the north by Jammu and Kashmir, on the east by Himāchal Pradesh, on the south by Haryāna and Rājasthān, and on the west by Pakistan.

The word Punjab is a compound of two Persian words, *panj* ("five") and *āb* ("water"), thus signifying the land of five waters, or rivers. The first known use of it occurs in the writings of the Muslim traveller Ibn Baṭṭūṭah, who visited India in the 14th century. The origin can perhaps be traced to *pañca nada,* Sanskrit for "five rivers," the word used before the advent of Muslims with a knowledge

Land of the five rivers

of Persian to describe the meeting point of the Jhelum, Chenāb, Rāvi, Beās, and Sutlej rivers, before they joined the Indus. The name Punjab subsequently came to be used for the land through which these rivers flow. As applied to the present Indian state of Punjab, it is strictly a misnomer, for, since the partition of India in 1947, only two rivers, the Sutlej and the Beās, lie within its territory.

The state in its present form came into existence on November 1, 1966, as a result of a territorial reorganization carried out on the basis of language. The capital city is Chandīgarh, which is also the headquarters of the government of Haryāna, a new state comprising most of the residuary, non-Punjabi-speaking areas of the older unit. Chandīgarh, planned by the Swiss architect and city planner Le Corbusier, was built as a capital city for what became the Indian Punjab after the partition.

Punjab covers an area of 19,445 square miles (50,362 square kilometres).

PHYSICAL AND HUMAN GEOGRAPHY

The land. More than 95 percent of the total area of Punjab is a flat plain, sloping gently from about 900 feet above sea level in the northeast to about 550 feet in the southwest. Physiographically, it is divisible into three parts: (1) the Siwālik Hills, in the northeast, about 900 to 3,000 feet high (covering a small fraction of the state's area); (2) the narrow, undulating foothill zone dissected by closely spaced seasonal torrents, locally known as *cos*, several of which terminate in the plain below without joining any stream; and (3) the flat tract, with fertile alluvial soils. The low-lying floodplains along the rivers and the slightly elevated flat uplands between them are distinguishable within the plain. Sand dunes occur in the southwest and six to nine miles west of the Sutlej River.

Punjab has an inland subtropical location, and its climate is continental, being semi-arid to subhumid. Summers are very hot, the mean temperature during June being 93° F (34° C), rising above 113° F (45° C) on exceptionally hot days. Winters are fairly cold, the January mean temperature being 55° F (13° C), with night temperatures occasionally touching the freezing point. Annual rainfall is highest in the Siwālik Hills in the northeast, where it is about 49 inches, and decreases gradually to about 14 inches in the southwest. More than 70 percent of the annual rainfall is concentrated in the months of the monsoon winds, July to September. Winter rains from the western cyclones that occur from December to March account for nearly 15 percent of the total rainfall.

With the growth of human settlement over the centuries, Punjab has been largely cleared of its vegetal cover. The existing vegetation is tropical and deciduous. The *shisham* or *sissoo* tree (of which the wood is used for construction and furniture making), the chinaberry, the pipal (which bears a small fig), and jujube tree, and *kikar* (gum arabic) are among the most common trees. Over large parts of the Siwālik Hills, bush vegetation has succeeded tree vegetation as a result of extensive deforestation. New types of trees have been planted on the hillsides and along roads in recent years.

Wildlife in Punjab includes the blue bull (also called the nilgai, a large bluish-gray animal), the wild boar, rabbits, jackals, foxes, and deer of various kinds. The commonest birds are crows, sparrows, doves, pigeons, parrots, partridges and peacock. The koel (cuckoo) is a summer visitor and the mountain crane a winter one. Waterfowl include cranes, herons, and geese. Snakes abound in summer, being especially in evidence in the rainy months; the cobra, viper, and krait (an extremely venomous kind of snake) are poisonous. Land in Punjab is precious, and agriculture impinges heavily on wildlife.

The people. The people of Punjab are mainly descendants of Aryan tribes that entered India from the northwest about 1500 BC. Relics of a highly developed pre-Aryan civilization have been discovered at Ropar, near Chandīgarh. Successive waves of invaders—Greeks, Parthians, Kushans, and Huns—became assimilated with the early Aryans. The Jāṭs, Rājputs, and other allied peoples are the products of this intermixture. Some members of the scheduled castes—formerly called the untouchables but named Harijan (People of God) by Gandhi—and of aboriginal tribes have traces of the physical features of the Dravidian people.

The language pattern

Punjābī is the mother tongue of about two-thirds of the people and Hindī of one-third. Punjab is the only state in India with a majority of Sikhs, who account for about 60 percent of the population. The Sikhs are followers of an Indian religion that had its origin in the teachings of Nānak, a 15th-century guru. Hindus make up about one-third of the population, and the small remainder is composed of Muslims, Christians, Buddhists, Jains, and followers of other religions.

Punjab is comparatively urbanized. A little more than a quarter of its population lives in cities and towns. Its major cities are Amritsar, Ludhiāna, Jullundur, and Patiāla. Though the cities are fast-growing centres of industry and manufacture, more than 70 percent of the state's population is still dependent on agriculture.

The economy. The economy of the state is characterized by a dynamic agriculture; a developing small-scale industry; and a growing per capita income despite a rapid rise in population.

Although Punjab claims less than 5 percent of India's total cultivated area, it accounts for about a quarter of its total output of wheat and a good percentage of its other food grains. The phenomenal progress has been described as the Green Revolution. The principal food-grain crops are wheat, corn (maize), rice, *bājrā* (millet), barley, and pulses (the edible seeds of various leguminous crops, such as peas, beans, and lentils). Other crops include cotton, sugarcane, potatoes, and oilseeds.

The increase in agricultural production is due largely to the successful application of modern agricultural technology. Much of the net area sown has been brought under irrigation. Canals water about half of the irrigated area; and wells, tube wells, and pumping sets account for the remainder. The Bhākra–Nāngal multipurpose river-valley project is the largest irrigation system of its kind in Asia. Other significant changes are the adoption of high-yielding varieties of semidwarf Mexican wheats and hybrids, the use of chemical fertilizers and pesticides, and increased mechanization.

The Bhākra–Nāngal project

The benefits of improved agriculture, augmented by cooperative societies and community-development programs, have tended to give comparatively more help to the bigger farmers, with the result that the disparity in farm-family incomes has increased considerably.

Punjab has practically no mineral or fossil-fuel resources, and only a tiny percentage of its total area is under forest. Consequently, industrial development in the state is associated with agriculture or devoted to consumer goods. The principal industries manufacture cotton, woollen, and silk textiles, machine tools, agricultural implements, bicycles, sewing machines, sports and electrical goods, automobile parts, or process sugar, flour, and milk products.

Increasing agricultural production

Since the introduction of planned development in 1951, Punjab has maintained a growth rate higher than that of any other Indian state in agriculture and has achieved the highest per capita income.

Extension of road transport has taken priority in development plans. A network of roads has been constructed to link the remotest parts of the state with railway stations, marketing centres, and industrial towns, thus opening the countryside to commerce and facilitating the interflow of farm products and consumer and capital goods.

Long-distance transport is confined largely to the railways. Punjab is served by a government-operated railway, which has its headquarters in Delhi. Regular air passenger service is available between Delhi and some of the Punjab towns, such as Amritsar, Ludhiāna, and Patiāla. Delhi and Chandīgarh are similarly connected. Water transport is practically nonexistent. Like the railways, the posts and telegraph service and broadcasting service are owned and controlled by the Union (central) government.

Administrative and social conditions. *Government.* As in other states of the Indian Union, the governor, the constitutional head of the administration, is appointed by the president of India on the advice of the Union cabinet and acts through a council of ministers responsible to a

legislative assembly that is elected every five years on the basis of adult franchise. The legislature was bicameral until January 1970, when the upper chamber (the legislative council) was abolished. The council of ministers is headed by a chief minister, who chooses its members from the legislative assembly.

The executive officials belong to a permanent, nonpolitical secretariat, members of which are selected by union or state public-service commissions. The administration is assisted in policy making and expertise both by the secretariat and by a number of standing boards and committees, such as the Planning Board, the Rural Development Board, and the Electricity Board.

The collector is the chief executive of the district (which forms the basic unit of administration), and the state is divided into 12 such districts, grouped into the two divisions of Patiāla and Jullundur, each headed by a divisional commissioner, whose functions are chiefly supervisory. The 12 districts are Gurdāspur, Amritsar, Bhatinda, Jullundur, Hoshiārpur, Fīrozpur, Ludhiāna, Kapūrthala, Sangrūr, Ropar, Farīdkot, and Patiāla.

The judiciary

The judiciary is independent of the executive. At its head is the high court, which, as constituted at present, is common to Punjab and Haryāna. Justice at the district level is administered by a district and sessions judge, assisted by subjudges and magistrates. Appeals from decisions of the high court are directed to the supreme court of the Indian Union.

Education. Schools are maintained largely by the state. Education is compulsory and free for pupils aged 6 to 11. Secondary education is also free in state schools.

There are four universities in the state: the Pūnjabī University at Patiāla, the Gurū Nānak University at Amritsar, the Punjab University at Chandīgarh, and the Punjab Agricultural University at Ludhiāna. There are also more than 200 colleges and technical institutions. There is a state department to encourage and organize sports, and Punjab youth is well represented in Indian athletic teams.

Health and welfare. Punjab enjoys higher survival prospects and better health conditions than most of the other states. The state has undertaken an extensive family-welfare program, especially in rural areas.

The state has schemes for the uplift of the scheduled castes (former "untouchables")—who form more than a fifth of the population—through scholarships, recruitment to public services, and help in finding employment.

Developments that have had a revolutionary effect on the lives of the people are the spread of electrification in rural areas and the dissemination of vocational and cultural education through broadcasting.

Numerous social services are provided by government and voluntary organizations. These include the care of infirm or handicapped persons, neglected children, widows, and destitute persons. The government also provides pensions for the aged who have no means of subsistence. There is a network of employment exchanges to assist the unemployed.

Cultural life. Folklore, ballads of love and war, fairs and festivals, dancing, music, and Punjābī literature are characteristic expressions of the state's cultural life. Old romances, such as *Heer Rānjhā, Sassī Punnūn,* and *Mirzā Ṣāḥibān,* have been told in verse many times over, and the names of their authors—especially Warris Shah and Hasham, both Muslims—are household names in Punjab. Older than these folk romances is the mystical and religious verse of the 13th-century Muslim Ṣūfī (mystic) Shaikh Farīd and of the 15th–16th-century guru Nānak, founder of the Sikh faith, who for the first time used Punjābī extensively as a medium of poetic expression. These are the origins of Punjābī literature, which, at the beginning of the 20th century, entered its modern phase with the writings of the poet and author Bhai Vir Singh, and the poets Puran Singh and Dhani Ram Chatrik.

The whole populace joins in religious and seasonal festivals, such as Dussehra, Dīwālī, Baisākhi, and in the anniversary celebrations in honour of gurus and saints. Secular dances were introduced by immigrants after partition. The *bhangra, jhumar,* and *sammi* are the popular forms. The *giddha,* a native Punjābī form, is a hilarious song-dance performed by women. In addition to Sikh religious music, semi-classical Mughal forms, such as the *khayāl, ṭhumrī, ghazal,* and *qawwālī,* continue to be popular.

The outstanding architectural monument is the Golden Temple at Amritsar, blending Indian and Saracenic styles. Its chief motifs, such as the dome and the geometrical design, are repeated in most of the Sikh gurdwaras, or shrines. Mural paintings in some of the gurdwaras preserve specimens of Sikh art. The Golden Temple itself is rich in gold filigree work of the most delicate kind and in panels with floral designs and marble facings inlaid with coloured stones. Among other important buildings may be mentioned the Martyr's Memorial at Jalliānwālla Bagh in Amritsar, the Hindu Temple of Durgiāna (also at Amritsar), a Moorish mosque at Kapūrthala, and old forts at Bhatinda and Bahādurgarh.

A Punjābī artist who made a great impact on modern art in the first half of the 20th century was Amrita Sher-Gil (1913–41). Some of the leading present-day Punjābī artists are Satish Gujral, Sobha Singh, S.G. Thakur, and Kirpal Singh.

HISTORY

The foundations of the present Punjab (historical Pañjāb) may be said to have been laid by Bandā Singh Bahādur, a hermit who became a military leader and who, with his fighting band of Sikhs, temporarily liberated the eastern part of the province from Mughal rule in 1709–10. Bandā Singh's defeat and execution in 1716 were followed by a long, drawn-out struggle between the Sikhs on one side and the Mughals and Afghans on the other. In 1764–65 the Sikhs established their sway in the region. Ranjit Singh (1780–1839) built up the Punjab into a powerful kingdom and attached to it the provinces of Multān, Kashmir, and Peshāwar. In 1849 the Punjab fell to the troops of the British East India Company and came under British rule. When independence was achieved in 1947, the Punjab was split between the new sovereign states of India and Pakistan, the smaller, eastern portion becoming part of India.

Since independence, the history of the Indian sector of the Punjab has been dominated by Sikh agitation for a Punjābī-speaking state, first led by Master Tara Singh and later by his political successor, Sant Fateh Singh. In November 1956 the Punjab was enlarged by the merger with it of the states of PEPSU, short for Patiāla and East Punjab States Union, which was formed by the amalgamation of the princely territories of Patiāla, Jīnd, Nābha, Farīdkot, Kapūrthala, Kalsia, Mālerkotla, and Nālagarh. Political and administrative leadership to the new Punjab was provided by Sardar Partap Singh Kairon, Congress chief minister of the state from 1956 to 1964. Demands for the formation of a new state by the consolidation of the Punjābī-speaking areas and the elimination of the Hindī-speaking portions were eventually agreed to by the government of India. On November 1, 1966, the state of Punjab was divided on the basis of language into the two new states of Punjab and Haryāna (the Hindī-speaking area).

The absorption of the princely states

In the early 1980s Punjab was once again the site of Sikh agitation. Demands were made for an autonomous Sikh state, Khālistān. The demands of the militant Sikhs of the Shiromanī Akālī Dal (SAD; Akālī Religious Party) were companied by a series of violent incidents against the Hindus. The state was placed under president's rule in 1983–84 following the resignation of its chief minister.

(H.K.M.S.)

Rājasthān

Rājasthān, a state of the Indian Union, is situated in the northwest of India. It is bounded on the west and northwest by Pakistan, on the north and northeast by the states of Punjab, Haryāna, and Uttar Pradesh, on the east and southeast by Uttar Pradesh and Madhya Pradesh, and on the southwest by Gujarāt. The Tropic of Cancer passes through its southern tip in the Bānswāra district. The state has an area of 132,130 square miles (342,214 square kilometres). The capital city is Jaipur.

Rājasthān, meaning "the abode of the rājās," was for-

"The abode of the rājās"

merly called Rājputana, "the country of the Rājputs" (sons of rājās). Before 1947, when India achieved independence from British rule, it comprised 18 princely states, two chiefships, the small British-administered province of Ajmer-Mārwār, and a few pockets of territory outside the main boundaries. After 1947 the princely states and chiefships were integrated with India in stages, and the state took the name of Rājasthān. It assumed its present form on November 1, 1956, when the act for the reorganization of the Indian states came into force.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The Arāvalli Hills form a line across the state running roughly from Mount Ābu, which is 5,650 feet (1,722 metres) high, in the southwest to Khetri in the northeast. About three-fifths of the state lie northwest of this line, leaving two-fifths in the southeast. These are the two natural divisions of Rājasthān. The northwest tract is sandy and unproductive with little water but improves gradually from desert land in the far west and northwest to comparatively fertile and habitable land toward the east. The area includes the Thar (Great Indian) Desert.

The southeastern area is higher in elevation (330 to 1,150 feet above sea level), more fertile, and very diversified in character. In the south lies the hilly tract of Mewār. A large area of the districts of Kota and Būndī forms a tableland; to the northeast of this there is a rugged region following the line of the Chambal River. Farther north the country levels out; the flat plains of Bharatpur district form part of the alluvial basin of the Yamuna River.

The Arāvallis form the most important watershed, the drainage to the east of the range flowing northeast. The Chambal is the only large and perennial river. Its principal tributary, the Banās, rises in the Arāvallis near Kumbhalgarh and collects all the drainage of the Mewār Plateau. Farther north, the Bāngangā, after rising in Jaipur, flows east to join the Yamuna. In the northwestern sector the Lūni is the only significant river. This rises in the Pushkar Valley of Ajmer and flows 200 miles west–southwest into the Rann of Kutch.

The rivers of Rājasthān

The soils in the vast sandy northwestern plain of the districts of Jaisalmer, Barmer, Jālor, Sirohi, Jodhpur, Bīkaner, Gangānagar, Jhūnjhunu, Sīkar, Pāli, and Nāgaur improve in fertility toward the east and northeast. These soils are predominantly saline or alkaline. Water is scarce but is found at a depth of 100 to 200 feet. The soil and sand are calcareous (chalky). Nitrates in the soil increase its fertility.

The soils in the Ajmer district in central Rājasthān are sandy; clay content varies between three percent and nine percent. The salt content is low. In Jaipur and Alwar districts in the east soils vary from sandy loam to loamy sand. In the Kota, Būndī, and Jhālawār tract, they are in general black and deep and are well drained. In Udaipur, Chittaurgarh, Dūngarpur, Bānswāra, and Bhīlwāra the eastern area has mixed red and black and the western area red and yellow soils.

There is a wide range of climate varying from extremely arid to humid, the humid zone comprising the Sirohi, Bānswāra, and Jhālawār districts in the south. Except in the hills, the heat in summer is great everywhere, with a mean maximum of 108° F (42° C). Hot winds and dust storms occur, especially in the desert tract. Winter temperatures vary from 68° F to 76° F (20° C to 24.5° C). The western desert has little rain (annual average four inches [100 millimetres]), but in the southwest rainfall is copious. The southwest benefits from both the Arabian Sea and Bay of Bengal summer monsoon winds, which bring 90 percent of the rainfall.

The main floral feature is scrub jungle. Toward the west, plants characteristic of the arid zone occur, such as tamarisk and false tamarisk. Trees are scarce, found only sparingly in the Arāvallis and in eastern Rājasthān.

Tigers are found in the Arāvallis and in several districts. Leopards, sloth bears, Indian sambār (dark brown Indian deer), and chital (a kind of deer) occur in the hills and forests, where nilgais (blue bulls) are also to be found in parts; black buck and ravine deer are numerous in the plains. Snipe, quail, partridge, and wild duck occur everywhere except in the desert. The Bīkaner region is well known for several species of sand grouse.

Tribal peoples

The people. Aboriginal tribes in Alwar, Jaipur, Bharatpur, and Dholpur areas include the Mināṣ; the Meos; the Banjārās, who are travelling tradesmen and artisans; and the Gadia Lohārs, another itinerant tribe, who make and repair agricultural and household implements. The Bhīls, one of the oldest tribes in India, inhabit the former princely states of Dūngarpur and Bānswāra and parts of Partāpgarh, Mewār, and Sirohi. Typical Bhīls are small, dark, and broad-nosed. They are famous for their skill in archery. The Grasias and Kathodis live in the Mewār Hills. The Grasias are tall, with good physique; unflinching loyalty to tribe, friends, or employers is a feature of their character. The Kathodis are nomads, Sahariyas are found in the Kota district, and the Rebaris of Mārwār are cattle breeders.

The Rājputs form, though representing only a small percentage of the population, the most important section of the population in Rājasthān. They are proud of their warlike reputation and of their ancestry. The Brahmin class is subdivided into many *gotras,* while the Mahājan (the trading class) is subdivided into a bewildering number of groups. Some of these groups are Jains, some Hindus. In the north and west the Jāṭs and Gujars are among the largest agricultural communities.

The four main dialects

The principal language of the state is Rājasthānī, comprising a group of Indo-Aryan dialects derived from Dingal, a tongue in which bards once sang of the glories of their masters. The four main dialects are Maṛwarī (in western Rājasthān), Jaipurī or Ḍhundhārī (in the east and southeast), Malwī (in the southeast), and, in Alwar, Mewātī that shades off into Braj Bhāsā in Bharatpur district. The use of Rājasthānī is declining with the spread of modern education, its place being taken by Hindī.

Hinduism is the religion of much of the population and takes the form of the worship of Brahmā, Śiva, Śakti, Vishnu (Viṣṇu), and other gods and goddesses. There are also followers of the Arya Samaj, a reforming sect of modern Hinduism, and of other forms of that religion. Jainism is also important; it has not been the religion of the rulers of Rājasthān but has its followers among the trading class and the wealthy section of society. Mahāvīrjī, Ranakpur, Dhulev, and Karera are the chief centres of Jaina pilgrimage. Another important religious sect is formed by the *Dādūpanthīs,* the followers of Dādū (d. 1603) who preached the equality of all men, strict vegetarianism, total abstinence from intoxicating liquor, and life-long celibacy.

Islām expanded in Rājasthān with the conquest of Ajmer by Muslim invaders in the late 12th century. Khwājah Muʿīn-ud-Dīn Chishtī, the Muslim missionary, had his headquarters at Ajmer; and Muslim traders, craftsmen, and soldiers settled there. There are also a few Christians and Sikhs in the state.

Rājasthān is one of the least densely populated states in India. Most villages and towns lie east of the Arāvallis. The urban population has been growing faster than the rural, but even so there are only a few large towns, including Jaipur, Ajmer, Jodhpur, Udaipur, Kota, Bīkaner, and Alwar. Industrial complexes are at Jaipur, Kota, Udaipur, and Bhīlwāra.

Rural houses are huts with mud walls and roofs thatched with straw. They have a single door, but no windows or ventilators. The houses of well-to-do farmers and artisans in larger villages have more than one room. They are roofed with tiles and have a veranda and large courtyard, whose main door will admit a loaded bull cart. The earthen floors are coated with mud and dung.

The economy. Rājasthān is a predominantly agricultural and pastoral state and exports food grains and vegetables. Despite a low and erratic rainfall, nearly all types of crops are grown; in the desert area, *bājrā* (millet); in Kota *jowār* (sorghum); in Udaipur mainly corn (maize). Wheat and barley are fairly well distributed, except in the desert area, as are pulses (the edible seeds of leguminous plants, such as peas, beans, and lentils) and oil seeds. Improved varieties of rice have been introduced, and this crop was expected to be expanded in the areas of the Chambal and Rājasthān Canal projects. Cotton is an important cash

crop in the south and north. The use of improved agricultural implements is increasing.

Animal products

Although more than half its area is arid or semi-arid, Rājasthān has a large livestock population in comparison with the rest of India and is the largest wool-producing state. It has a monopoly in camels and in draught animals of various breeds.

Having much arid land, Rājasthān needs extensive irrigation works, including wells and canals. It receives water from the Punjab rivers and also from the Gurgaon (Haryāna) and Āgra canals (Uttar Pradesh) and from the Narmada River to the south. These resources, properly utilized, could bring one-third of the cultivated land under irrigation. There are thousands of tanks (village ponds or lakes), but they suffer from drought and silt. Rājasthān shares the Bhākra Nangal project with the Punjab and the Chambal Valley project with Madhya Pradesh; both are used to supply water for irrigation as well as for drinking purposes. The Rājasthān Canal project is the largest in the state.

Rājasthān produces India's entire output of lead and zinc concentrates, emeralds, and garnets, and more than 90 percent of its gypsum and silver ore.

The main industries are based on textiles, vegetable oil, wool, minerals, and chemicals, while the handicrafts—such as leather goods, marble work, jewelry, pottery, and embossed brass—formerly patronized by the princely courts, have been earning much foreign exchange. Various industrial concerns received substantial loans and subsidies from the government and from the Rājasthān Finance Corporation, a semi-governmental agency.

Kota, which is the industrial capital of the state, has a nylon factory and a precision-instruments factory, as well as plants for the manufacture of calcium carbide, caustic soda, and rayon tire cord. There is a zinc smelter plant near Udaipur.

Electricity supplies are obtained from neighbouring states as well as from the Chambal Valley project. There is an atomic-energy plant at Rāwatbhāta, near Kota.

Administrative and social conditions. *Government.* Rājasthān state is headed by a governor, appointed by the president of the Indian Union, whom he represents, for a five-year term; he exercises administrative, legislative, financial, and judicial powers. The powers of the state government are controlled by the unicameral Legislative Assembly, the members of which are elected; some of the seats are reserved for representatives of the scheduled castes and tribes. The state is divided into 27 districts: Ajmer, Alwar, Bānswāra, Barmer, Bharatpur, Bhīlwāra, Bīkaner, Būndī, Chittaurgarh, Churu, Dholpur, Dūngarpur, Gangānagar, Jaipur, Jaisalmer, Jālor, Jhālawār, Jhūnjhunu, Jodhpur, Kota, Nāgaur, Pāli, Sawai Mādhopur, Sīkar, Sirohi, Tonk, and Udaipur.

In each district the collector, who is also the district magistrate, is the principal representative of the administration. He functions in close cooperation with the superintendent of police to maintain law and order in the district; he is the principal revenue officer; he coordinates the activities of the other departments in the district; and acts as a link between the state government and the people.

Rule by *pañcāyat*

Rājasthān was the first state to experiment with *pañcāyat rāj* (rule by *pañcāyat,* or village council); the system has elicited a response among the villages.

Education and welfare. There are many educational establishments in Rājasthān, including universities at Jaipur, Udaipur, and Jodhpur and the Birla Institute of Technology and Science at Pilani. There are several state hospitals and dispensaries. There are also many Ayurvedic and Unanī (a medicinal system using prescribed herbs and shrubs) institutions. The state incurs heavy expenditure on education, on maternity and child welfare, on rural and urban water supplies, and on the welfare of the disadvantaged.

Cultural life. Hardly a month passes in Rājasthān without a religious festival. The most remarkable and typical is the festival called Gangor, when clay images of Mahādevī and Pārvatī (representing the benevolent aspects of the Hindu mother goddess) are worshipped by women of all castes for 15 days and are then taken out to be immersed in water. Their procession is joined by priests and officers and is led to the water by trumpeters and drummers. Hindus and Muslims join in each others' festivals; general enthusiasm and gaiety prevail on these occasions.

Another important festival is held at Pushkar near Ajmer, taking the form of a mixed religious festival and livestock fair. It is visited by farmers from all over the state, bringing their camels and cattle, and by pilgrims seeking salvation. The average number of visitors ranges between 100,000 and 300,000.

The typical folk dance of Rājasthān is the *ghoomar,* which is performed on festive occasions only by women. The *geer* dance (performed by men and women), the *panihārī* (a graceful dance for women), and the *kacchi ghori* (in which male dancers ride dummy horses) are also popular. The most famous song is the *Kurja,* which tells the story of a woman who wishes to send a message to her absent husband by the *kurja* (a bird), who is promised a priceless reward for his service. Rājasthān has made its contribution to Indian art, and there is a rich literary tradition, especially of bardic poetry. Chand Bardāī's poem, *Prithvi Raj Raso* or *Chand Rāisā* the earliest manuscript of which dates back to the 12th century, is particularly notable. A popular source of entertainment is the *khyāl,* a dance drama composed in verse with gay, historical, or romantic themes.

Rājasthān abounds in objects of antiquarian interest: early Buddhist rock inscriptions, Jaina temples, forts, splendid princely palaces, and Muslim mosques and tombs.

HISTORY

Archaeological exploration has shown that early man lived along the banks of the Banās and its tributaries some 100,000 years ago. Harappan and post-Harappan culture (3rd–2nd millennium BC) is traceable at Kāli Banga, Ahār, and Gilund. Pottery fragments at Kāli Banga are carbon-dated to 2700 BC. The discovery near Bairāt of two rock inscriptions of the emperor Ásoka (*c.* 250 BC) seems to show that his rule extended westward to this part of the state. Later rulers of the whole or parts of the state were the Bactrian Greeks (2nd century BC), the Scythians (Śakas; 2nd to 4th centuries AD), the Gupta dynasty (4th to 6th centuries), the Huns (6th century), and Harṣavardhana, a Rājput ruler (early 7th century). Between the 7th and 11th centuries several Rājput dynasties arose, including that of the Gurjara-Pratihāras, who kept the Arab invaders of Sind at bay.

Rājput strength reached its zenith at the beginning of the 16th century under Rānā Sangrām Singh (Sāngā) of Mewār, but he was defeated in a fierce battle by the Mughal invader Bābur, and the brief splendour of a united Rājputana waned rapidly.

When the Mughal emperor, Akbar, came to power in midcentury, the chief of the Kachwāhās at Amber entered the imperial service; Jodhpur and Chitor were subdued. It is noteworthy that the emperors Jahāngīr and Shāh Jahān were both born of Rājput mothers. After the death (1707) of the emperor Aurangzeb, the Rājput state of Bharatpur was developed by a Jat conqueror, but by 1803 most of the rest of Rājputana was under the domain of the Marāṭhā people of the central Arabian coast. Later in the 19th century the British subdued the Marāṭhās and with the gradual penetration of British paramountcy the boundaries of the Rājput states were precisely defined. The government of India was represented in Rājputana by a political officer, styled agent to the governor general, who was also chief commissioner of the small British province of Ajmer-Mārwār. Under him were residents and political agents who were accredited to the various states.

The birth of nationalism

During this period the idea of Indian nationalism was born. Maharishi Dayanand wrote at Udaipur his *Satyārath Prakāsh,* intended to restore Hinduism to its pristine purity, which created a ferment in Rājputana. Important movements of thought also occurred among the Jaina *sādhus* ("holy men") and scholars. Ajmer was the centre of political activity, and nationalist leaders included Arjun Lal Sethi, Manik Lal Varma, Gopal Singh, and Jai Narain Vyas. The people of Rājputana joined the people of the rest of India in nationalist aspirations and were no longer

morally dominated by the princes. When the new constitution of India came into force in 1950, Rājasthān became an integral part of India. The Rājput princes surrendered their powers to the central government. (I.P.)

Uttar Pradesh

Uttar Pradesh, one of the Indian states, lies, for the most part, in the upper valley of the Ganges River in northern India. Uttar Pradesh is one of the border states of the country, with both Nepal and the Chinese autonomous region of Tibet adjoining it to the north. This landlocked state is bounded by Himāchal Pradesh in the northwest, by Haryāna and the union territory of Delhi in the west, and by Rājasthān in the southwest. To its south is Madhya Pradesh, and in the east lies the state of Bihār.

In area, Uttar Pradesh ranks fourth among the Indian states. It occupies 113,673 square miles (294,413 square kilometres) of territory, about 9 percent of the total area of the country. In population, however, the state ranks first in the country. Lucknow is its capital. Several government offices, however, are still located in Allahābād, which was the capital of the United Provinces (now Uttar Pradesh) from 1904 to 1949. The state, in its present form, was created on January 26, 1950, when India became a republic. Since that date it has been called by its present name of Uttar Pradesh (literally, Northern Province).

PHYSICAL AND HUMAN GEOGRAPHY

The land. Physiographically, the state can be divided into four distinct regions: (1) The Himalayan region in the north, with several folded mountain ranges, extends from the bare slopes of perpetual snow to the forest-clad Siwālik hills. Geologically, these ranges are made up of Tertiary (dating from 65,000,000 years ago) marine sedimentary, granite, and metamorphic rocks. Several peaks—namely, Badrīnāth, Nanda Devi, Trisūl, and Kāmet—rise to more than 23,000 feet above sea level. (2) The submontane tract has two distinct portions. Immediately below the Siwālik lies the forested piedmont zone called Bhābar. Here, mountainous streams disappear in the dry season beneath masses of gravel and boulders. To the south of this zone is a wider strip of land called the Terai, a damp and marshy tract of thick jungle and tall grass, where streams from the Bhābar reappear on the surface. (3) The Ganges Plain, which makes up the largest part of the state, is composed of alluvial deposits, brought down by the Ganges and its tributaries. Altitude varies from 1,000 feet in the northwest to 190 feet in the extreme east. The entire plain is almost level and featureless and slopes generally from the northwest to the southeast. (4) The hill and plateau region of the south forms part of the central Indian plateau. In some areas it presents an old erosion surface, not more than 1,000 feet high, carved out of granite. The southeastern section consists of hills and valleys carved out of sandstone and limestone. The hills of the Vindhya Range, north of the Son River, rise to more than 2,000 feet.

A land of many rivers, Uttar Pradesh is well drained. In the northwest and northeast flow the Ganges and its tributaries—the Yamuna, the Rāmganga, the Gomti, and the Ghāghara (Sarju). The southwestern part of the state is drained by the Chambal, the Betwa, and the Ken—all tributaries of the Yamuna. In the southeast, the Son River, a tributary of the Ganges, flows from west to east through the Mirzāpur district.

The most important soils are the alluvial ones found in about two-thirds of the total area of the state. They range from sandy to clayey loam. In some areas in the west are found soils that are alkaline and impregnated with salts and that are mostly infertile. In the southern hill and plateau district, soils are mixed red and black.

The entire state, except the northern mountainous region, where some peaks remain covered with permanent snow, falls under the tropical monsoon climate. In the plains, temperatures in January range from 54.5° to 63.5° F (12.5° to 17.5° C) and in May from 81.5° to 90.5° F (27.5° to 32.5° C) with a maximum of 113° F (45° C) or more. In summer, a hot wind blows from the west. Nearly 85 to 90 percent of the annual rainfall is received during the rainy season from the Bay of Bengal branch of the summer monsoon. It varies from 40 to 80 inches in the east to from 24 to 40 inches in the west. Moderately heavy rainfall between 40 and 80 inches occurs in the Himalayan region, though in some places it exceeds 80 inches.

Forests

About one-sixth of the area is under forest, most of which is found in the Himalayan region, the Terai, and the southern hill areas. The Himalayan region contains rich coniferous and broad-leaved forests of spruce, pine, deodar, and sal trees. In the Terai and the Bhābar are found moist tropical deciduous forests and tall grasses. Some areas are covered with a dry tropical deciduous forest. In the southeast, in the Mirzāpur district, an open shrub and thorn forest occurs, while to the south of the Son River an open sal (East Indian hardwood) forest is found. Among the important trees of the plains are mango, pipal (sacred fig tree), and banyan.

The fauna of the state comprises most of the animals and birds common to the Ganges Plain. Tigers are found in the Terai, and leopards abound in the riverine forests. Nilgai (or bluebull, a type of antelope), rabbits, wolves, jackals, and antelopes roam freely in the jungles. Feathered game include teal, wild duck, snipe, jungle fowl, peacock, quail, pigeon, and partridge.

The people. Two ethnographic groups are found in Uttar Pradesh. In the extreme north, near the Tibetan border, people exhibit Mongoloid features. In the hill and plateau region of the south and in the plains in the central part of the state, inhabitants are predominantly Aryo-Dravidian.

Although Uttar Pradesh is the most populous state in the country, the distribution of population is uneven, the influence of the physical environment being obvious in this regard. Because of the mountainous terrain, the Himalayan region is sparsely populated, while the plains support about 70 percent of the total population. The density is highest in the east and decreases westward, rising again in some Upper Doab districts. The southern hill and plateau region is thinly populated.

Uttar Pradesh is one of the Hindī-speaking states of India; Hindī is the mother tongue of most of the population. Urdū is the next most important language.

Hindus (about 85 percent) and Muslims (about 15 percent) are the two important religious communities. Sikhs, Buddhists, Jainas, and Christians together constitute less than 1 percent. The proportion of Hindus in the total population has gone down since 1921, while that of Muslims has been constantly rising, except during 1941–51. Muslims are relatively more numerous in towns and cities than in the countryside, and other minority communities have an even greater urban orientation. Hindus are largely concentrated in the rural areas.

The economy. Agriculture is by far the most important sector of the state's economy. Irrigational facilities have been provided in about one-third of the total cultivated area. Food crops, such as wheat, rice, corn (maize), and pulses (edible seeds), dominate. Among commercial crops, sugarcane is the most important. Oilseeds, potatoes, and cotton come next in terms of acreage. Tobacco and jute are also significant commercially, and Uttar Pradesh is the major opium-growing state in India. The per acre yield of most crops is generally low.

Principal crops

The state is not very rich in forest resources, and commercial timber forms only a small fraction of the total wood production. Lack of adequate road transport hinders the exploitation of the rich Himalayan forests. Afforestation and road development are receiving top priority to meet the demands for raw material by the various forest-based industries.

The state also ranks very low in mineral resources. Limestone, silica, phosphatic shale, and magnesite are the important minerals mined in the state. Uttar Pradesh is a major producer of silica and limestone. Coal has been discovered in the Singrauli coal fields in Mirzāpur district. The mineral-based industries of the state include cement, glass, and ceramic industries.

Uttar Pradesh has vast hydroelectric power potential in the northern and southern hilly areas, and the Rihand, Sarda, Yamuna, and Ganges (besides many minor rivers)

have been harnessed for this purpose. Thermal power, however, remains the most important. The coal requirements of the thermal plants are met by the West Bengal–Bihār fields.

Uttar Pradesh is comparatively less industrially developed than other states of India. Cottage and village industries such as weaving, ceramics, musical instruments, leather and woodwork, and metalware dominate the industrial structure. Among large-scale activities, sugar and textile industries are the most important. Kānpur is the industrial metropolis of Uttar Pradesh, having a large industrial complex consisting of textile, leather, and engineering factories.

For a number of years the state government has been trying to boost industrial activity by establishing cement factories near Mirzāpur, a precision-instrument factory at Lucknow, and a chemical plant at Bareilly. The central government has also been involved in establishing several industrial projects in the public sector, including an antibiotics factory at Rishikesh, a diesel-locomotive works at Vārānasi, an oil refinery at Mathura, and a nitrogenous-fertilizer factory at Gorakhpur.

The transportation system of the state consists of a network of roads, railways, air routes (operated by the Indian Airlines), and the three major inland waterways of the Ganges, the Yamunā, and the Ghāghara. Competition with roads and railways has led to a decline of inland water transport. The major defect of rail transport lies in the fact that there are two different gauges, with seven break-of-gauge transshipment points. A substantial part of the goods traffic is, therefore, diverted from the railways to the roads. Several large towns are linked by air routes with Delhi.

Rail service

Administrative and social conditions. *Government.* The state executive consists of the governor and a council of ministers with a chief minister as its head. The governor, as head of the state, possesses both legislative and executive powers. The government is run by the elected representatives of the people. The legislative assembly consists of 425 members. The legislative council, which has 108 members, is a permanent body, with one-third of the members retiring every second year. In the absence of the elected government, as in other states of India, the governor carries on the administration of the state with the help of an advisory body.

For purposes of administration the entire state is divided into 56 districts, each under a collector (of revenues) and magistrate, and 245 tahsils, each under a tahsildar. The latter is responsible for both revenue collecting and the maintenance of law and order.

The High Court of Judicature is located at Allahābād, with a bench also at Lucknow. The executive has been separated from the judiciary in all the districts of the state except the seven hill districts.

Health and education. The state's lack of development is especially critical in the sphere of public health and education. There are proportionately few doctors and hospitals for the immense population, and in literacy Uttar Pradesh is far behind other states. About a quarter of the population is literate. Although there has been considerable improvement since independence (literacy in 1941 was under 10 percent), the situation is far from satisfactory. The state has 20 universities and more than 350 degree-conferring colleges.

Lag in literacy

Welfare. The social-welfare department of the government is engaged in providing institutional services to the physically handicapped, in suppressing immoral traffic in women and girls, and in providing probation services aimed at reformation and rehabilitation of youthful first offenders and delinquent children. The government takes special interest in the welfare of the Harijans (people formerly belonging to the "untouchable" caste). The state government has passed several acts for the welfare of labourers, including fixing a minimum wage and subsidizing industrial housing, retirement funds, and employees' state insurance schemes.

Cultural life. Uttar Pradesh has contributed much to the composite culture of India. From time immemorial various arts have flourished in towns and cities and in the countryside. Handicrafts developed at various centres, notably in Lucknow, famous for its *chikān* (a type of embroidery), in Morādābād, noted for its metal enamelling, in Vārānasi, known for its brocades and brassware, and in Nagīna, known for its ebony work.

The songs and dances of the countryside are a notable feature of the local culture. They include the *kajari* of Mirzāpur and Vārānasi, the Alha Udal, a folk epic, and various village dances, such as *kaharva* and *lilli ghori* (hobby horse), with its vigorous movements and pungent satire of society.

The Lucknow College of Arts, the Morris College of Music in Lucknow, and the College of Indology of the Banaras Hindu University are doing much to encourage the growth of the fine arts in the state. Various drama groups at Allahābād are also active.

Since independence, Hindī has taken great strides, and its poetry has become rich. Its medieval poets apart, Uttar Pradesh has produced notable modern Hindī poets, as well as several famous Urdū poets.

HISTORY

Uttar Pradesh, the birthplace of Rāma and Krishna of the *Rāmāyaṇa* and the *Mahābhārata* (the two great Indian epic poems), has been the seat of Hindu religion and culture from the earliest to the present times. Buddha preached his first sermon at Sārnāth, and he attained *parinirvāṇa* (the freedom of spirit brought about by release from the body) at Kuśinagara. In the Vedic period (*c.* 1500–600 BC) the area formed part of the ancient country known as Madhyadeśa—the middle country. Of the 16 *mahājanapadas* (great states) of the early Buddhist period, as many as seven were located in this state. Later, great Hindu kings, such as Aśoka, Candragupta Maurya, Samudra Gupta, and Harśa (Harṣavardhana), ruled over the area.

During the Muslim period, this part of the great plain of northern India was known as Hindustān. The state remained under the Mughal emperors, who ruled from Delhi, until the close of the 18th century. One of them, Shāh Jahān, built the famous Tāj Mahal at Āgra.

The arrival of the British in India was followed by a gradual extension of their power west from Bengal during the 19th century. The British Presidency of Āgra was first formed in 1833. Prior to this date, the area had been included in the British administrative region known as the Presidency of Bengal. In 1836, possessions of the East India Company in the present state of Uttar Pradesh and Delhi were combined under the name of North-Western Provinces. In 1856 the Province of Oudh was annexed by the British and kept under separate administration until 1877, when it was united with the North-Western Provinces under the name North-Western Provinces and Oudh. The area was rechristened United Provinces of Āgra and Oudh in 1902.

The Indian Mutiny of 1857 was practically confined to Uttar Pradesh. The first decisive outbreak started at Meerut and spread to Lucknow, Kānpur, Allahābād, Āgra, Jhānsi, and some nearby places outside the state. Nana Sahib of Bithūr, Maharani Lakshmi Bai of Jhānsi, and Tantia Topi were the main personalities in this unsuccessful revolt against foreign rule. After the mutiny, the administration of Uttar Pradesh, together with other parts of India, passed from the East India Company to the British crown.

Role in the independence movement

Uttar Pradesh was in the forefront of the national movement for independence. The fourth session of the Indian National Congress was held at Allahābād in 1888. The state gave to the country political leaders, such as Motilal Nehru (father of Jawaharlal), Madan Mohan Malaviya, and Purushottamdas Tandon. It was due to mob violence at Chauri Chaura, a village near Gorakhpur in Uttar Pradesh, that the civil-disobedience movement was for some time suspended by Mahatma Gandhi. The great revolutionary Chandra Shekhar Azad was shot dead by the British at Allahābād. The state is the birthplace of India's first three prime ministers—Jawaharlal Nehru (1947–64), Lal Bahadur Shastri (1964–66), and Indira Gandhi (1966–77, 1980–84).

(R.L.D.)

NORTHEAST INDIA

Arunāchal Pradesh

Arunāchal Pradesh, meaning "land of the rising sun," is a sparsely populated mountainous area in India's extreme northeast. Formerly the North East Frontier Agency, it became a union territory in 1972 and attained statehood in 1987. It stretches from the Brahmaputra River plain in Assam northward to the main Assam Himalayas and eastward to an irregular line passing through a series of lofty peaks, known as the "hump" during the World War II airlift to China. The territory covers an area of 32,270 square miles (83,578 square kilometres). The capital is Itānagar.

PHYSICAL AND HUMAN GEOGRAPHY

The state comprises nine districts. West Kameng in the west adjoins the sovereign kingdom of Bhutan; the other districts from west to east are East Kameng, Lower and Upper Subansirī, West and East Siang, and Dibāng Valley districts. At the blind eastern end of the Assam valley is Lohit district, and Tirap district lies to the south on the Burmese border. The area is intersected by rivers; its greater part has a heavy rainfall and is thickly forested. It is inhabited by tribes mostly of Mongoloid stock such as the Monba, Mishmi, Abor, Miri, Dafla, and Aka. The tribes occupy roughly a defined tribal area and do not intermarry as a general rule. About 50 distinct languages and dialects are spoken. The peoples are mostly animists and practice shifting cultivation. Crops include rice, corn (maize), wheat, potatoes, mustard, cotton, millet, tapioca, vegetables, and fruits. Communication within the state is a major problem.

Arunāchal Pradesh has great potential, though as yet undeveloped: hydroelectric power, as well as forest, coal, and oil resources. There are no major industries; small industries include rice processing, saw and plywood mills, fruit preservation units, and handlooms. There is one degree college at Pasīghat.

The government is headed by a governor appointed by the president of India, and by an elected chief minister. The chief minister and the council of ministers are responsible to the unicameral Legislative Assembly.

HISTORY

The northern boundary, about 550 miles long, in dispute between India and China, is known as the McMahon Line because Sir Henry McMahon was secretary in the Indian foreign department and represented Great Britain at the conference held in Simla (1913–14) to settle frontier and other matters relating to Tibet. The line was regarded by the British as the natural, ethnic, and administrative boundary. Representatives of Britain, China, and Tibet agreed that the frontier between Tibet and northeastern India should follow the crest of the high Himalayas. Two days later the Chinese republican government disavowed its plenipotentiary and refused to sign a convention. Prior to that, the British–Indian government had made agreements with the indigenous tribes and set up the Balipara frontier tract in the west and the Sadiya frontier tract in the east, together constituting the North East Frontier Agency (1912–13) and including undoubted Assamese territory.

After the independence of India in 1947, China made claims to practically the whole area covered by the districts of East and West Kameng, Lower and Upper Subansirī, East and West Siang, and Lohit, arguing that the McMahon Line had never been accepted by China and was the result of British "aggression." In letters to the Indian prime minister, Jawaharlal Nehru, the Chinese prime minister, Zhou Enlai, quoted a map in the 1929 edition of *Encyclopædia Britannica* showing the disputed territory as Chinese, with the boundary following the alignment of Chinese maps. Some Chinese maps before 1935 showed the North East Frontier Agency (Arunāchal Pradesh) as part of India, and since then as part of Tibet. The *Survey of India* (1883) showed the disputed tribal areas as de facto administered by British India. British and Indian maps since 1914 have usually followed the McMahon Line. If the Chinese claims were allowed, the Indian–Chinese border would follow roughly the margin of the Assam plain, a frontier almost impossible to defend. Following this dispute, Chinese troops crossed the McMahon Line on August 26, 1959, and captured an Indian outpost at Longju, a few miles south of the line. They abandoned this in 1961 but in October 1962 crossed the line, this time in force. After first striking toward the Tangla ridge and Towang near the Bhutan border, the Chinese later extended their attack along the whole frontier. Deep inroads were made at a number of points. Later the Chinese agreed to withdraw approximately to the McMahon Line and in 1963 returned Indian prisoners of war. (Ed.)

Assam

Assam is located in the northeast of India, from which it is almost isolated on the west by Bangladesh. It has an area of 30,318 square miles (78,523 square kilometres). Until an administrative reorganization in 1972, the states of Arunāchal Pradesh, Mizorām, and Meghālaya were part of Assam.

Assam is bounded to the northwest by the kingdom of Bhutan; to the northeast by Arunāchal Pradesh; to the east by Nāgāland and Manipur; to the south by Mizorām and Tripura; to the southwest by Meghālaya and the country of Bangladesh; and to the west by West Bengal. A narrow corridor running through the foothills of the Himalayas in West Bengal connects Assam with India. Geographically, Assam has no natural boundaries. The capital of Assam, formerly located at Shillong (now the capital of Meghālaya), was shifted to Dispur, a suburb of Gauhāti, in 1972; and again in 1984 to Prāgjyotiṣapura, near Gauhāti.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Except for the districts of Korbi Anglong and North Cāchār Hills, Assam is generally composed of plains and river valleys. The state may be divided into three principal physical regions—the Brahmaputra Valley in the north, the Barāk Plain in the south, and the hilly regions that lie between these two valleys.

The Brahmaputra River Valley is the dominant physical feature of Assam. Known as the Ya-lu-ts'ang-pu Chiang in Tibet, it enters Assam near Sadiya at the extreme northeast corner and runs westward for nearly 450 miles, traversing every district across the length of Assam before turning due south to enter the plains of Bangladesh. The river valley, rarely more than 50 miles wide, is studded with numerous low, isolated hills and ridges that rise abruptly from the plain. The valley is surrounded on all sides, except on the west, by mountains and is intersected by many streams and rivulets that flow from the neighbouring hills to empty into the Brahmaputra.

The Barāk River in the southeast separates Assam from Manipur. The Barāk Valley, about 60 miles wide and about 125 miles long, mostly in the Cāchār district, supports the region's agriculture and a relatively dense population.

Geologically, the Brahmaputra and Barāk valleys are alluvial in origin and consist of sand, sandstone pebbles, clay, and sometimes a mixture of sand and clay containing decomposed vegetational matter. The Tertiary deposits (from 2,500,000 to 65,000,000 years old) include a similar variety of rocks, among them hard sandstone, soft and loose sand, conglomerates, coal seams, shales, and sandy clays. The lowest Tertiary deposits include limestone.

The Korbi Anglong and part of the North Cāchār Hills form part of the Meghālaya Plateau, which may have been an extended portion of Gondwanaland (an ancient landmass in the Southern Hemisphere that once grouped together South America, Africa, Australia, and part of the Indian subcontinent). This truncated portion of the main plateau displays a rugged topography, with shales of the Barail Series exposed in valley bottoms. It has, roughly, a northerly slope, with the outer ranges of the Mīkīr Hills

reaching an average elevation of about 1,500 feet. Higher average elevations of about 3,300 feet are reached in the central portion of the Korbi Anglong district east of the Kopili River. (The river has formed plain embayments into the interior of the Korbi Anglong, largely isolating this section from the main plateau.)

The northern ranges, which extend from Dabaka (in Nowgong district) in the southwest to Bokākhāt (in Sibsāgar district) in the northeast, attain an average elevation of 2,000 feet. Major peaks in the north include Basundhari Parbat (2,540 feet), Raisang (2,420 feet), Mehekongthu (2,095 feet), and Kud Parbat (2,055 feet).

The southern ranges, known as the Rengma Hills, have an average elevation of about 3,000 feet. Main peaks in the south are Chenghehishon (4,460 feet) and Khunbaman Parbat (4,300 feet).

Earthquakes

Earthquakes are a common phenomenon in Assam. In modern times the most important Assamese earthquakes have been those that occurred in 1897, with the Shillong Plateau as the epicentre; 1930, with Dhubri as the epicentre; and 1950, with Ch'a-yŭ (Rima) in Tibet at the Arunāchal Pradesh border as the epicentre. The 1950 earthquake is considered to have been one of the most disastrous earthquakes in history. It created heavy landslides that blocked the courses of many hill streams. The floods that followed the bursting of these artificial dams caused more loss of life and property than the earthquake itself.

The average temperature is moderate—about 84° F (29° C) in the hottest month (August). In the cool season, when the northeast monsoon (rain-bearing) wind blows down the Brahmaputra Valley, the average valley temperature in January is 61° F (16° C). At this season the climate in the valley is marked by fogs and a little rain. Assam escapes the normal Indian hot, dry season. Some rain occurs from March onward, but the real force of the monsoon winds is felt from June onward, when they blow at right angles to the Assam hills. Rainfall in Assam is not only the highest in the country but also ranks among the highest in the world; its annual average varies from 70 inches in the west to more than 120 inches in the east. Largely concentrated during the months from June to September, it often results in widespread destructive floods.

The state's forested area covered 18,500 square miles before the creation of Meghālaya and Mizorām and formerly represented nearly 39 percent of the total area of the state, of which reserved forest, which is efficiently managed by the government, constituted only about 6,400 square miles, or about 12 percent of the total area. The creation of Meghālaya and Mizorām, however, substantially reduced Assam's forested area to about 20 percent.

A portion of the reserved forest in Assam is devoted to wildlife sanctuaries and national parks. The Kāziranga National Park, the stronghold of the fast-disappearing great Indian one-horned rhinoceros, is its most famous wildlife refuge. The most important forest products are timber and bamboo, firewood, and lac (the source of shellac). There are about 74 species of timber, of which two-thirds are commercially exploited. The forests are inhabited by wild animals such as elephants, tigers, deer, and wild pigs.

The people. The distribution of population is uneven owing to the hilly terrain, the number of rivers, the forests, the small amount of cultivable land, and the lack of industrialization.

In the 20th century the unusually rapid growth of population in the state has been due mostly to the immigration into Assam of tea-garden labourers, graziers from Nepal, Muslims from West Bengal, and refugees from what is now Bangladesh.

Urbanization

About 90 percent of the population is rural, but there has been an increase in the urban population in recent years. The growth of many industries, increased commercial activity, and the desire of the Bangladesh refugees to live near towns are some of the causes of the increase in urban population. Only Gauhāti has a fairly large urban population.

The people of the plains of the Brahmaputra and Barāk valleys are mainly of Indo-Iranian stock, although many are of Mongoloid stock, particularly in the upper Assam region. By the time of their arrival in the Brahmaputra Valley, it would appear that the original Aryan people of Assam had become intermixed with other peoples. Assamese is the principal language and is regarded as the lingua franca of the whole of northeast India. Although scholars trace the history of the Assamese literature from the beginning of the second millennium AD, yet an unbroken record of literary history is traceable only from the 14th century. The majority of the people of the Cāchār district in the Barāk Valley speak Bengali.

About two-thirds of the Assamese are Hindus; about a quarter of the population is Muslim. The Muslims are mostly recent settlers from Bangladesh, or converts belonging to the lower strata of the peoples of Indo-Iranian origin. A majority of the Hindus accept Vaisnavism (Vaiṣṇavism), which is based on the deity Vishnu.

The tribes in the hills, as well as those in the plains, are of Mongoloid stock. They speak dialects of Tibeto-Burman origin, with the exception of the Khāsi dialect, which is derived from the Austroasiatic branch of languages. Many of the hill tribes have been converted to Christianity by missionaries, but the majority still observe the customs and festivals of their traditional religion, which is based on animism and has a close affinity to the ancient form of Hindu worship. The Mikirs and Kachāris of the Korbi Anglong and North Cāchār Hills are mostly Hindus; although they speak dialects of Tibeto-Burman origin, they have adopted Assamese as their first language.

The economy. Agriculture is of basic importance to Assam. More than half of the total working population is directly engaged in agriculture, with another 10 percent employed on tea plantations, in forestry, or in other occupations related to agriculture.

Rice accounts for about two-thirds of the sown area. Tea and jute are important foreign-exchange earners. Their cultivation in the Brahmaputra Valley is extensive. Assam grows just over half of the country's tea. Other crops include oilseeds, pulses (leguminous plants, such as peas, beans, or lentils), sugarcane, rape (an oil-yielding plant, the leaves of which are used for fodder), mustard, and potatoes.

Fruits include oranges, pineapples, and bananas. The state produces a surplus of cereals, but there is a shortage of oilseeds and pulses. Double-cropping and other improved methods of cultivation are being introduced.

The known minerals of the state are petroleum, coal, natural gas, limestone, fireclay, china clay, and feldspar. Of these, oil, coal, natural gas, and limestone are being commercially exploited. Oil is found at Digboi, Nahorkatia, Hugrijan, Morān, Rudrasagar, and Lakowa—all in Upper Assam; Assam produces about one-half of the country's petroleum and natural gas. Coal is found in the districts of Dibrugarh (Upper Assam) and in Korbi Anglong. Railways, plantations, and steamships are the main consumers of Assam coal. Limestone is quarried in the Mīkīr Hills.

In spite of its raw materials, industrial growth in Assam is in its infancy, for—with the exception of tea and oil—there are few industries of significance. Industrial development is handicapped mainly by Assam's isolation from the rest of India, by a bad transport system, by a small local market, and by the lack of sufficient capital. A number of industrial enterprises have nevertheless been started. These include a fertilizer plant at Nāmrūp, a jute mill at Silghāt, a sugar mill at Dergaon, a paper mill at Jogighopa, a spun-silk mill at Jagiroad, and a cement factory at Bokajān. There are also many sawmills and plywood and match factories that make use of the timber supplies of Assam. A number of food-processing units for rice, oil, and fruit have been established. The oil refinery at Digboi was built as early as 1899; the Nunmāti refinery near Gauhāti started production in 1962. The Bāruni refinery is in Bihār state but is supplied crude oil from Assam through a pipeline.

A poor transport and communication system has been the main hindrance to the economic development of Assam. Though the four recognized forms of transport—road, rail, air, and water—all exist, none satisfactorily serves the needs of the region. The isolated location of Assam, linked as it is with the rest of India by only a

narrow corridor through the Himalayas, has to a great extent prevented the development of quick and efficient road and rail systems in the region. The growth of an inland water-transport system has similarly been checked by the inclusion of the lower part of the Brahmaputra in Bangladesh. There is also a considerable amount of airborne traffic between Assam and Calcutta. The important airports are Borjhār (Gauhāti), Mohanbāri (Dibrugarh), Lilābāri (North Lakhīmpur), Rowroyāh (Jorhāt), Sāleni (Tezpur), and Kumbhirgrām (Silchar). The airports are used by the Indian Airlines Corporation (IAC) and Vayudoot for civil air transport of passengers and freight.

Administrative and social conditions. *Government.* The state government has a unicameral legislature under a governor and a council of ministers. The state of Assam comprises 10 districts—Goālpāra, Kāmrūp, Nowgong, Darrang, Sibsāgar, Dibrugarh, and Lakhīmpur in the Brahmaputra Valley; Cāchār in the Barāk Plain; Korbi Anglong; and the North Cāchār Hills in the Meghālaya Mikīr Plateau. In February 1983, the state elections for the legislative assembly were accompanied by violence in which more than a thousand people lost their lives.

Education and welfare. Compulsory primary education is in force for children between the ages of six and 12 in several towns and a few thousand villages. Education up to the secondary level is free in the state. There are universities in Gauhāti, Jorhāt, and Dibrugarh. Welfare-extension projects, operating through more than 80 centres and covering more than 300,000 people, provide recreational and cultural facilities for women and children.

Cultural life. The cultural life of Assam is interwoven with the activities of a number of cultural institutions and religious centres like *satra* (seat of a religious head, the Satrādhikār) and *namghar* (prayer hall). Several *satras* established at different places in Assam have been looking after the religious and social well-being of the villagers as well as of some urban dwellers for the last 400 years. The Assamese people observe all the pan-Indian religious festivals such as Durgā-pūjā (the festival of the goddess Durgā), Dol-jātrā or Holī (the spring festival), and Janmastami (birthday of Lord Krishna). But the most important social and cultural celebrations are the three Bihu festivals observed with great enthusiasm irrespective of caste, creed, and religious affinity. The Bihus were originally agricultural festivals observed by the villagers at different seasons of the year. The Bohāg Bihu, celebrated in mid-April with the commencement of the new year (first day of the Bohāg or Baiśākh month which falls usually on the 14th of April) during the spring season, is the most important one. It is also known as Rangāli Bihu (*rang* means merry-making and fun). It is observed by dancing and singing in open spaces as well as in the houses of villagers. The mother or other ladies of the family present a hand-woven *gāmochā* (towel) to each and every member of the family on this day.

The second important Bihu, the Māgh Bihu, celebrated in mid-January (in the month of Māgh), is a harvest festival. It is performed with community feasts and bonfires. It is also known as Bhogāli Bihu (*bhog* means enjoyment and feasting). The third Bihu festival, the Kāti Bihu, is observed in mid-October. It is also known as the Kangāli Bihu (*kangāli* means poor) because by this time of year, the house of a common man is without foodgrains as the stock is usually consumed before the next harvest.

Another important aspect of the cultural life of the people of Assam, particularly of the women, is weaving of fine silk and cotton cloths with various floral and other decorative designs. Every Assamese house, irrespective of caste, creed, and social status, has at least one loom and each woman is required to know the art of weaving.

HISTORY

In the earliest times the territory and its environs were known as Kāmarūpa, a state that had its capital at Prāgjyotiṣapura (modern Gauhāti, in Assam on the southern bank of the Brahmaputra). Ancient Kāmarūpa included roughly the Brahmaputra Valley, Bhutan, the Rangpur region (now in Bangladesh), and Cooch Behār, in West Bengal. King Narakāsura and his son Bhagadatta were famous rulers of Kāmarūpa in the Mahābhārata period, at least 1000 BC. A Chinese traveller, Hsüan-tsang, who visited the country when it was ruled by Bhāskaravarman, about AD 640, left a vivid account of the country and its people. After this, information about ancient Assam is very meagre for a few centuries, but the discovery of several copperplates, clay seals, and stone inscriptions that date from the 7th to the mid-12th centuries has thrown some light on the condition of the country at that time. These archaeological records indicate that the inhabitants of the region attained considerable power and a fair degree of civilization. The copperplates further provide clues to the locations of important ancient settlements and also the routes connecting them.

Assam was ruled by various dynasties—the Pālas, Koches, Kachāris, and the Chutiyas—but, because of the constant warfare amongst these princes, there was no stable government in the area until the coming of the Ahoms in the 13th century. The Ahoms crossed the Pātkai Range from Burma and conquered the local chieftains who then ruled the Upper Assam Plain. In the 15th century the Ahoms, who probably gave their name to the country, were the dominant power in Upper Assam. Two centuries later, after defeating the Koches, the Kachāris, and other local rulers, they also became the rulers of Lower Assam up to Goālpāra. The power and prosperity of the Ahoms reached a zenith during the rule of King Rudra Singh (1696–1714).

Rule by the Ahoms

Dissension and jealousy among the princes gradually weakened the central administration, with the result that, in 1786, the ruling prince Gaurinath Singh, in despair, sought the aid of the British in Calcutta. Captain T. Welsh, sent by the British governor general in India, restored peace and was recalled in spite of the protests of the Ahom king. Internal strife then caused one crisis after another until, in 1817, the Burmese entered Assam in response to the appeal of the *bar phukan* (governor) Badan Chandra, a rebel against the king. They swept over the country thrice, bringing destruction and misery.

The British, whose interests elsewhere were threatened by these developments, ultimately drove out the Burmese invaders, and, after the Treaty of Yandabo was concluded with Burma in 1826, Assam became a part of British India.

A British agent, representing the governor general, was appointed to administer the devastated area of Assam. In 1838 Assam was incorporated into British-administered Bengal. By 1842 the whole of Assam Valley had come under British rule. In 1874 a separate province of Assam was created under a chief commissioner, with its capital at Shillong. In 1905, on the initiative of the British viceroy, Lord Curzon, the province was amalgamated with eastern Bengal. The subsequent partition of Bengal was, however, resented, and in 1912 this arrangement was reversed, with Assam once more made a separate province. During World War II, Assam became one of the main supply routes for Allied forces operating in Burma, and within Assam decisive battles were fought in 1944 in several places (including Bishenpur, in Manipur, and Kohīma, in Nāgāland). These battles were instrumental in preventing the Japanese advance into India.

After India achieved independence in 1947 and as a result of the ensuing partition, the district of Sylhet (excluding the Karīmganj subdivision) was ceded to Pakistan. In 1961 and 1962 Chinese armed forces, disputing the McMahon Line as the boundary between India and Tibet, occupied part of the North East Frontier Agency (Arunāchal Pradesh). In December 1962, however, they voluntarily withdrew to Tibet. In 1963 the Nāga Hills district, formerly a part of Assam, became the 16th state of the Indian Union, under the name of Nāgāland. A part of Tuensang, a former territory of the North East Frontier Agency, was also added to Nāgāland. In 1970, after the tribal people of the Gāro district and the United Khāsi and Jaintia Hills district, located in the centre of Assam, had demandeda separate state, an autonomous state named Meghālaya was created with these two districts. In 1972, the North East Frontier Agency was renamed Arunāchal Pradesh and was separated from Assam, and the Union Territory of Mizorām was created in the south. (H.D./Ed.)

Bihār

Bihār (Behar), one of the constituent states of the Republic of India, is bounded on the north by Nepal, on the east by West Bengal, on the south by Orissa, and on the west by Madhya Pradesh and Uttar Pradesh. The state has an area of 67,134 square miles (173,876 square kilometres); it is one of the most populous of the Indian states. Its capital is Patna.

Bihār occupied an important position in the early history of India; for centuries it was the principal seat of imperial powers and the main focus of Indian culture and civilization. Under the British it formed a part of the Bengal Presidency until 1912, when the province of Bihār and Orissa was formed. In 1936 Bihār was separated from Orissa and constituted a separate province, and, on India's independence in 1947, it became a state. In 1948 the small states of Saraikela and Kharsāwān were merged with Bihār. In 1956, when the Indian states were reorganized on a linguistic basis, 3,140 square miles of territory was transferred from Bihār to West Bengal.

The state is naturally divided into two parts—the alluvial plains of the Ganges (Gaṅgā) in the north and the Choṭa Nāgpur Plateau in the south. The plains districts are popularly known as Bihār proper and the plateau districts as Choṭa Nāgpur. Bihār proper is further divided into two parts by the Ganges, and it is customary to call these North Bihār and South Bihār. This designation of the central portion of Bihār as South Bihār is somewhat misleading but has the sanction of usage.

The name Bihār (Vihār) dates back to about AD 1200 and was given by the Muslim invaders who were struck by the large number of Buddhist monasteries (*vihāras*) that they saw, especially near the then-capital of Odantapurī (present-day Bihārsharif).

PHYSICAL AND HUMAN GEOGRAPHY

The land. Bihār, as has been pointed out, consists of two physiographic units—the Gangetic Plains in the north and the Choṭa Nāgpur Plateau in the south. The plains form part of the middle Gangetic Plains and are in turn divided into north and south plains by the Ganges River. Except for the Himalayan foothills in the extreme northwest, the north Ganges Plain forms a flat alluvial country, less than 250 feet above sea level and liable to flooding. The Ghāghara, the Gandak, the Bāghmati, the Kosi, the Mahānanda, and other rivers flow down from the Nepal Himalayas and make their way to the Ganges in frequently changing channels. Depressions and lakes mark the abandoned courses of streams. The soil consists mostly of new alluvium—chalky and light-textured (mostly sandy loam) west of the Burhi (Old) Gandak River and nonchalky and heavy-textured (clay and clay loam) to the east.

The plains and the plateau

The south Ganges Plain is more diversified than the north, and many hills rise from the level alluvium. The rivers, with the exception of the Son, are all small; their water is diverted into irrigation channels. Beyond the low-lying, treeless country immediately to the south of the Ganges levee (embankment), there are no swamps or waterlogged areas. The soil consists mainly of older alluvium, composed of a darkish clay or yellowish loam, with poor, sandy soils predominating toward the south of this region.

The Choṭa Nāgpur Plateau, a series of plateaus, hills, and valleys, covers the southern half of Bihār and consists mostly of crystalline rocks. The main plateaus, Hazāribāgh and Rānchi, are separated by the faulted, sedimentary coal-bearing basin of the Dāmodar River, and they average about 2,000 feet in height. In the west there are more than 300 dissected but flat-topped plateaus over 3,000 feet high, known as *pats*. The highest point in Bihār is formed by the conical granite peak of Parasnāth—4,477 feet (1,365 metres) high—in Hazāribāgh; it is sacred to both the Jaina religion and to the Santāl tribe. In the extreme northwest, beyond the Son Valley, lies the Kaimur Plateau, with horizontal sandstone strata underlain by limestone. In the Dāmodar Valley, the soil is sandy; the typical soil of the plateau is red soil, and there are patches of lateritic (red, leached, iron-bearing) soil on the uplands.

There are three well-defined seasons: the hot-weather season, lasting from March to mid-June; the rainy season, from mid-June to October; and the cold-weather season, from November to February. May is the hottest month, with the mean temperature exceeding 90° F (32° C), except in the extreme north and the plateaus of Rānchi and Hazāribāgh. The normal annual rainfall varies from 40 inches in the west central part to more than 60 inches in the extreme north and in the southwest. The rainfall on the plateau—more than 50 inches—is heavier than on the plains. Nearly all the rain (85–90 percent) falls between June and October, and nearly 50 percent of the annual rain falls in July and August. The cold weather season is the pleasantest part of the year.

The natural vegetation is deciduous forest, but less than one-fifth of the total area is forested; most forests occur in the Himalayan foothills and on the Choṭa Nāgpur Plateau. In the Himalayan foothills, valuable sal (a resin-yielding species) is found, and bamboo, reeds, and grass are widespread. Choṭa Nāgpur forms a rich sal area; other timbers include some that are used for the production of lac (a resinous substance used to make varnishes), while tussah silkworms (*Antherea pernyi*) are fed on the leaves of the asan tree (*Terminalia tomentosa*). Mahua (an East Indian tree) yields sweet, edible flowers, also used in the distillation of liquor. Bamboo and sabai (a valuable Indian fibre grass also known as bhabar) of Choṭa Nāgpur supply raw materials for paper manufacture. Common trees of the plain are the banyan, pipal, and palmyra palm.

Tigers, leopards, elephants, and bears are found only in the more inaccessible forests.

The people. Nearly three-fourths of the population of Bihār is concentrated in the cultivated plains. The average density decreases toward the east because of the ravages of the Kosi River, which is apt to flood or to change its course, but the position has begun to change as a result of the harnessing of the river. On the South Bihār Plain, a highly developed system of irrigation supports a large population. Density declines toward the south. Settlement in Choṭa Nāgpur is confined largely to river valleys, deforested peneplains (areas reduced almost to plains by erosion), and mineral and industrial belts.

The great majority of the people live in villages. Compact or clustered villages are usually found in the plains, while dispersed rural settlements are characteristic of the plateau. Aboriginal tribes are concentrated in Choṭa Nāgpur, especially in the districts of Rānchi, Singhbhūm, and Santāl Parganas. Santāl, Oraon, Muṇḍā, and Ho are the principal tribes and together constitute four-fifths of the total tribal population.

With little more than 10 percent of its population classified as urban, Bihār is one of the most rural states in India. The major cities are Patna, Gayā, Bhāgalpur, Muzaffarpur, Darbhanga, Rānchi, Monghyr, Bihārsharif, and Jamshedpur; and the major town groups are Dhānbād–Jharia–Sindri and Bokāro and Chas.

Three-fifths of the total nonemployed urban population consists of females. Even in the tribal belt of Choṭa Nāgpur, where both men and women participate in economic activities, the high participation of females in work is confined mainly to rural areas.

Hindus constitute about 85 percent and Muslims more than 10 percent of the population. Christianity and animism are confined to Choṭa Nāgpur. Muslims are largely settled in North Bihār, particularly in Purnea in the northeast. Of the tribal population, most are Hindus, a few are Christians, and many adhere to animism. The Ho is the only tribe in which the majority follow animism; Khaṛiā is the only tribe in which the majority are Christians. Christianity is significant among the Muṇḍās and Oraons.

Indo-European languages

Indo-European languages—including Hindī, Urdū, and the dialects of Bhojpurī, Maithilī, and Magahī—are spoken by most of the population. Bhojpurī is spoken in the western districts of Bhojpur, Rohtās, Saran, and East and West Champaran; Maithilī, in Darbhanga and Saharsa; Magahī, in Patna, Gayā, and Monghyr. Austro-Asiatic (Muṇḍarī, Santālī, Ho) and Dravidian (Oraon) languages are confined to the aboriginal tribes.

The economy. *Mining.* The Choṭa Nāgpur Plateau is

the richest mineral belt in India, and Bihār produces more than a third (by value) of all the minerals mined in the country. Bihār produces almost the entire national output of copper, kyanite (an alumina–silica mineral used in the manufacture of heat-resistant porcelain), and phosphate and much of the output of bauxite (a source of aluminum), mica, coal, china clay, fireclay, and iron ore. Coal accounts for most of Bihār's mineral production. The principal coalfields, all in the Dāmodar Valley, supply nearly all the coking coal of India. Singhbhūm, together with districts in adjoining Orissa state, contains one of the world's richest and best hematite-iron-ore deposits. Copper is mined and smelted near Ghātsīla, in the Singhbhūm district. High-grade bauxite is mined in the Lohārdaga region of Rānchi. The Kaimur hills in Bhojpur and Rohtās provide much of Bihār's limestone, mostly used for cement manufacture. A mica belt, extending for over 90 miles from east to west and from 12 to 16 miles from north to south, mainly in Hazāribāgh, produces high-quality muscovite (a light-coloured mica). China clay is abundant in the Singhbhūm, Santāl Parganas, and Bhāgalpur districts, and fireclay is found in the coalfields of the Dāmodar Valley. The kyanite deposits of Kharsāwān in Singhbhūm are among the largest in the world. Singhbhūm is also important for manganese, chromite, apatite (rock phosphate, a source of fertilizer), and uranium. Extensive pyrite (an iron disulfide mineral used in the manufacture of sulfuric acid) deposits have been found near Amjhore, Rohtās.

Mineral resources

Agriculture. About half the area is under cultivation, but pressure of population has pushed cultivation to the furthest limits, and little remains to be developed.

The crops

The transitional nature of the climatic zone is reflected in the cropping pattern, which shows a mixture of wet and dry crops. Rice is everywhere the dominant crop, but corn (maize), wheat, barley, gram, oilseeds, and pulses (leguminous plants, such as peas, beans, and lentils) are important supplementary crops. Jute, a crop of the hot, moist lowlands, is found only in the easternmost plain districts. There are three harvests in a year: *bhadai*—sown from May to June and harvested in Bhado (August to September); *aghani*—sown in mid-June and harvested in the month of Aghan (December); and *rabī*—harvested in spring. Rice, essentially an *aghani* crop, covers 50 percent of the total cropped area. Areas of densest cultivation lie east of the Burhi Gandak, in North Bihār, and in the irrigated area of Patna division. In the plateau region, rice is cultivated in the river valleys.

Corn (maize), a *bhadai* crop, is grown principally in the northwest, west of the Burhi Gandak. In the plateau it is the most important food crop after rice and is grown in almost every village. Wheat, the most valuable *rabī* crop, is grown only in the plains, and barley is grown mainly in the northwestern districts. Pulses are the most important crop after rice in the plains. Sugarcane is grown in a fairly well defined belt in the northwest.

Fruits and vegetables are extensively grown. Muzaffarpur and Darbhanga are particularly noted for mangoes, bananas, and litchi fruits. Vegetables are important in the vicinity of large towns. The potato-growing area near Biharsharif, in Patna district, produces the best variety of seed potato in India. Chilies and tobacco are important cash crops on the banks of the Ganges.

Industry. Despite rich mineral resources, Bihār has been slow to develop industry. Only about one-fifth of the population is employed in industry as against about three-fourths in agriculture. A number of Industrial Area Development authorities have been set up by the state government to boost the pace of development. Most industrial workers are employed in household industries; the rest are employed in steel and other metal-based and food-processing industries. Rānchi, Bokāro, and Jamshedpur rank among the largest industrial complexes in India.

Industry

Regional industrial distribution shows heavy concentration in the two plateau districts of Singhbhūm and Dhānbād. It is, however, possible to recognize several significant zones of economic development. Singhbhūm, the richest mineral-bearing district, is important for heavy industries. Jamshedpur, the seat of ironworks and steelworks, has also attracted a number of satellite engineering industries. Copper is smelted at Moubhandar near Ghātsīla. Chaibāsa manufactures cement from Jamshedpur slag, and there is sheet-glass manufacturing at Kandra.

The Dāmodar Valley has coal-washing plants, coking plants, and large coal-burning thermal-power plants at Bokāro, Chandrapura, and Patratu. Sindri produces fertilizers and cement, and there is another cement factory at Khalāri. Other important manufactures include alumina at Muri (Rānchi) and sheet glass at Bhurkunda (Barka Kāna), in Hazāribāgh. Rānchi, just south of the Dāmodar Valley, with its heavy-machine building, foundry forge, and heavy-machine-tool plants, has grown into an important centre of heavy engineering. There is a large steel plant at Bokāro.

The Bihār mica belt has only a little significant manufacturing industry; processing is largely by skilled hand labour. Factories in this area are located mostly near Giridih, Kodarma, and Jumri Tilaiya.

Dālmianagar is the focus of the Son Valley industrial area and has paper, cement, sugar, chemical and other industries. Japla and Kalyanpur produce cement.

Light engineering and consumer industries have developed in the trading centres situated along the south bank of the Ganges at Patna, Mokameh, Monghyr–Jamālpur, and Bhāgalpur. North of the Ganges lies the petrochemical complex at Bāruni, connected by a pipeline with the Assam oil fields. Monghyr is noted for cigarette manufacture, and Bhāgalpur is noted for silk. The sugar belt of Tirhut manufactures four-fifths of the state's sugar. Rice milling is situated in the extreme north, where paddy (rice) for husking also comes from the adjoining Nepal Terai (a marshy lowland area). In the jute-growing area, there are jute mills at Katihār and Samāstipur.

Traditional small-scale cottage industries are popular in Bihār; they include siriculture (tussur silk), lac and glass work, handloom products, brassware, handicrafts, and pottery. Madhubani paintings, wall hangings with mythological stories on cloth, have become a foreign-exchange item.

The Dāmodar Valley Corporation (DVC) is the most prominent multipurpose scheme of Bihār; four dams forming a series of reservoirs have been constructed under the DVC at Tilaiya, Maithon, Konār, and Panchet Hill. Other schemes include Patratu, Bāruni, and Bokāro thermal power stations, and Subarnarekha and Kosi hydroelectric projects.

Transportation. The waterways, once important, are now of little significance. North Bihār is served by a narrow-gauge railway. Because of the dense population, the railways carry a heavy load of traffic. They run parallel with the rivers because of the difficulty of constructing bridges, and communication between important towns is consequently often long and tedious.

The rail network

South Bihār has better rail facilities. Three lines starting from Calcutta and passing through Sāhibganj, Madhupur and Gomoh converge onto the plain. There are road and rail bridges on the Son at Dehri and at Koilwar. The Rajendra Pul (Bridge), a rail and road bridge over the Ganges at Mokameh, connects South with North Bihār.

In Choṭa Nāgpur Plateau the railways follow the natural terrain. Extensive rail development has taken place in the coal-mining areas of the Dāmodar Valley. One of the railroads connecting Calcutta and Bombay passes through Jamshedpur via the Sanjai–Koel gap. There are, however, large areas only poorly served by rail transport. The mica-mining area in Hazāribāgh is served by the Eastern Railway.

In North Bihār most of the roads are unpaved. Dense population, inadequate rail facilities, and the requirements of the sugar industry have resulted in a dense network having been constructed toward the west, although the eastern half is poorly provided with roads. In South Bihār the important towns are well served with roads, but paved roads are frequently interrupted by unpaved stretches. Road accessibility is best around Patna and Gayā. The Choṭa Nāgpur Plateau has good roads, built principally during World War II. The Grand Trunk Road, a national highway, cuts across the plateau through Dhānbād and Dehri. Another national highway connects Patna with Rānchi and Jamshedpur.

Administrative and social conditions. *Government.* Bihār has a bicameral legislature consisting of the upper house, or Legislative Council, and the lower house, or Legislative Assembly. Appointed by the president of India, the governor is the constitutional head of the state and functions on the advice of the chief minister, who is the head of the council of ministers. The bureaucratic hierarchy located in the Patna secretariat is headed by a chief secretary.

The state is divided into seven administrative divisions and 33 districts. The divisions, with headquarters in parenthesis, are Patna (Patna); Tirhut (Muzaffarpur); Darbhanga (Darbhanga); Kosi (Saharsa); Bhāgalpur (Bhāgalpur); North Choṭa Nāgpur (Hazāribāgh); and South Choṭa Nāgpur (Rānchi). Local administration is the responsibility of a divisional commissioner in each of the divisions, a district magistrate and collector in each of the districts, and a subdivisional officer in each of the 76 subdivisions.

There is a separate administration for development, consisting of a commissioner at the state level, a district magistrate assisted by a district development officer at the district level, and a block (district subdivision) development officer assisted by committee, or *pañcāyat samiti,* at the local level. Below the blocks are *grām pañcāyats,* or elected village councils. The rural administration (*pañcāyati rāj*) has a similar district, block, and village organization.

The police administration is headed by an inspector general, assisted by superintendents at the district level. There is a high court at Patna, with a chief justice and several other judges. Below the high court are district courts, subdivisional courts, *munsifs*' (subordinate judicial officers') courts, and village councils.

Education. About one-fourth of the population is literate, males greatly outnumbering females in literacy. South Bihār is the most literate part, while North Bihār, mostly rural and agricultural, is the least literate. In urban areas the literacy rate is generally higher.

There has been a remarkable expansion of educational facilities since the early 1950s. Enrollment has increased but is still unsatisfactory, especially among the girls and in the higher age groups. Education is free for children up to age 11. Vocational and technical institutions are sponsored by government departments.

The universities

Higher education is the responsibility of the eight universities: Patna University, the oldest and most important, at Patna; Mithila University, at Darbhanga; Bihār University, at Muzaffarpur; Bhāgalpur University at Bhāgalpur; Ranchi University at Rānchi; Magadh University, at Buddh Gayā; and Rajendra Agricultural University, at Pua (Samāshpur). The Kameohwar Singh Darbhanga Sanskrit University, at Darbhanga, provides teaching and research facilities in all branches of Sanskrit learning.

Health. Medical facilities, though improving, are still inadequate outside the towns. Villages are served mainly by allopathic (traditional western) and ancient Indian medical (Āyurvedic) dispensaries. Unanī and homeopathic systems of medicine are also popular. Large and well-equipped hospitals and medical colleges are located at Patna, Darbhanga, Rānchi, Jamshedpur, and Bhāgalpur. Smallpox, respiratory diseases, dysentery, and diarrhea figure prominently among the causes of death. Plague, cholera, and malaria have been largely controlled. A tuberculosis sanatorium, a mental hospital, and a leprosarium are all located near Rānchi.

Cultural life. The cultural regions of Bihār show a close affinity with the linguistic regions. Maithilī is the language of old Mithilā (the area of ancient Videha, now Tirhut), which is dominated by orthodoxy and Brahminical way of life. Maithilī is the only Bihār dialect with a script of its own called Tirhuta, and a strong literary history; one of the earliest and most celebrated writers of Maithilī was Vidyāpati (15th century), noted for his lyrics of love and devotion.

Bhojpurī, a virile dialect spoken by an active and enterprising people, has hardly any literature but does have a considerable oral folk literature. Magahī too has a rich folkloric tradition. The Bihār Plain has also contributed significantly to modern Hindī and Urdū literature.

Tribal customs and traditions

Most tribal villages have a dancing floor, the sacred grove (*sarna*)—where worship is offered by a village priest—and a bachelor's dormitory (*dhumkuria*). All the tribes believe in one supreme being. The *hāt,* or weekly market, plays an important part in tribal economy. Tribal festivals (such as Sarhūl), a spring festival (Sohrai), and a winter festival (Mage Parab) are occasions of great festivity. Tribal culture is fast changing under the impact of external influences, such as Christianity, industrialization, new communication links, tribal welfare schemes, and community development projects.

Places of religious and cultural interest abound in the plains. Nālandā is the seat of the ancient and celebrated Nālandā Buddhist monastic university; nearby Rājgīr, with its ancient and modern temples and shrines, is visited by people of many faiths; Pāvāpurī is the place where Mahāvīra, the founder of Jainism, attained Nirvāna (Enlightenment, or freedom from an endless cycle of reincarnation). Gayā is an important place of Hindu pilgrimage, and nearby Buddh Gayā, where the Buddha attained Enlightenment, is Buddhism's holiest place. The holy town of Deoghar is well-known for its Baidyanāth temple. Hariharkshetra, near Sonepur, north of Patna, is famous for one of the oldest and largest animal fairs in India, which is held every November.

HISTORY

The ancient kingdoms

In the early Vedic period (beginning about 1500 BC) several kingdoms existed in the Bihār plain. North of the Ganges was Videha, one of the kings of which was the father of Princess Sītā, the wife of Lord Rāma and the heroine of the *Rāmāyaṇa,* one of the two great Hindu epic poems of India. During the same epoch, the capital of the ancient kingdom of Magadha was Rājagṛha (modern Rājgīr), about 45 miles southeast of Patna; to the east was the kingdom of Aṅga, with its capital at Campā (near Bhāgalpur). A new kingdom later arose in southern Videha, with its capital at Vaiśālī. By about 700 BC, the kingdoms of Vaiśālī and Videha were replaced by a confederacy of the Vṛjjis—said to be the first republican state known in history. It was in Magadha, in the 6th century BC, that Buddha developed his religion and that Mahāvīra, who was born at Vaiśālī, founded the religion of Jainism.

In about 475 BC the capital of the Magadha Empire was located at Pātaliputra (modern Patna), where it remained under Aśoka (emperor of India from about 273 to 232 BC) and the Guptas (a dynasty of emperors who ruled India in the 4th and 5th centuries AD) until the onslaught of the Hūṇas in the middle and late 5th century. In the 6th–7th century AD, the city was devastated by the migration of the Son River—the Chinese pilgrim Hsüantsang recording that in AD 637 the city had few inhabitants. It regained some of its glory, but it is doubtful that it ever served as the capital of the Pāla Empire (which lasted from about 775 to 1200). During the ensuing Muslim period (about 1200 to 1765), Bihār had little independent history, remaining a provincial unit until 1765, when it came under British rule and—together with Choṭa Nāgpur—was merged with the state of Bengal.

Originally, Choṭa Nāgpur was mostly forest-clad and was ruled by chiefs of various aboriginal tribes. Though British authority was only gradually established in the plains to the north during the second half of the 18th and the beginning of the 19th century, occasional revolts against them took place in Choṭa Nāgpur, the most important being the Ho revolt of 1820 to 1827 and the Muṇḍā uprising of 1831 to 1832.

Later, Bihār was an important centre of the Indian mutiny of 1857 to 1859 against British political authority. Bihār subsequently played an active role in the successive phases of Indian nationalism. Mahatma Gandhi, the nationalist leader who advocated nonviolent resistance, first launched the *satyāgraha* (passive-resistance) movement against the tyranny of the European indigo planters in the Champāran region of northern Bihār. Dr. Rajendra Prasad, who played a leading part in the freedom movement and was elected the first president of independent India, was born in the Saran district, northwest of Patna.

(P.D.)

Manipur

Manipur is a state of India, established in 1972 and located in the northeastern part of the country. It was previously a union territory. It is bordered by the Indian states of Nāgāland on the north and Assam to the west, by the state of Mizorām on the southwest, and by Burma to the south and east. It covers an area of 8,632 square miles (22,356 square kilometres) and is larger than Wales but smaller than Belize (formerly British Honduras). The capital, Imphāl, is the state's major trade centre.

Manipur is largely isolated from the rest of India, and the region has periodically been subject to invasions from the east. Its economy centres on agriculture and forestry, and there is lively activity in trade and cottage industries.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The state is composed of two major physical features, the Manipur River Valley and a large surrounding tract of mountainous country. The valley extends from north to south; it covers an area of about 650 square miles and lies at an altitude of 2,600 feet above sea level. The main physical feature of the valley is the reedy Lake Logtāk, which covers an area of about 40 square miles. The lake is the source of the Manipur River, which flows southward through the valley into Burma, where it joins the Myittha River, a tributary of the Irrawaddy.

Upland areas

The hill ranges have a general north–south trend; connecting spurs and ridges run from west to east between them. The ranges include the Nāga Hills to the north, the Manipur Hills along the eastern Burmese border, and the Mizo and Chin hills to the south. The average elevation lies between 5,000 and 6,000 feet above sea level. The hills are highest in the north, where they attain heights of more than 9,500 feet. In the west the mountains are broken by the valley of the Surma River. Known as the Barāk River in Manipur, it rises at Mount Japvo and flows south and west to join the Meghna River of Bangladesh.

The climate is temperate in the valley and cold in the hills. Rainfall is abundant, with about 65 inches (1,650 millimetres) of precipitation occurring annually.

The hills are densely covered with tree jungle and forests containing stands of bamboo and teak. Other trees include oak, magnolia, and chinquapin. The Luzon pine grows in the Nāga Hills. Plants include rhododendrons, primroses, and blue poppies. Animal life includes the Asiatic elephant, tiger, and leopard. The rhinoceros is hunted for its horn, which, when ground, is valued as an aphrodisiac. The brow-antlered deer is in danger of extinction. Gaurs are the largest wild oxen in the world; the smaller gayals have been domesticated. There is also a fair number of wild buffalo.

The people. About two-thirds of the people are the Meithei, who occupy the Manipur Valley and are largely Hindus. Meithei women conduct most of the trade in the valley and enjoy high social status.

The rest of the population is composed of indigenous hill tribes that are divided into numerous clans and sections. The Nāgas are found mainly in the north and the Kukis in the south. The tribesmen speak languages of the Tibeto-Burman family and practice traditional animist religions, although some of the Nāgas have been converted to Christianity. About 60 percent of the population speak Manipurī, which, together with English, is the official language of the state.

Agriculture and forestry

The economy. Agriculture and forestry are the main sources of income. Rice is the major crop, and the rich soil also supports corn (maize), sugarcane, mustard, tobacco, fruit orchards, and pulses such as peas and beans. Terracing is common in the hills, where the tribes plow the ground with hand hoes. Among some of the hill tribes, domestic animals are kept only for meat and are not milked or used for hauling. The Nāgas are known to use intoxicants to catch fish. Teak and bamboo are major forestry products.

Manufacturing is limited to several well-established cottage industries. The designed cloth produced on handlooms is in demand throughout India and outside the country. Other industries include sericulture, manufacture of bamboo and cane articles, soapmaking, carpentry, and tanning. An industrial complex has been established at Imphāl.

Manipur remains somewhat isolated from the rest of India, and communications within the state are poor. A national highway passes through the state from Tamu on the Burmese border in the south via Imphāl to Dimāpur (Nāgāland) in the north; this highway also connects Imphāl with the Northeast Frontier Railway near Dimāpur. There is no railway line in the state. There are air links from Imphāl to Gauhāti and Silchar in Assam, and to Calcutta in West Bengal.

Administrative and social conditions. The governor is the constitutional head of the state and is appointed by the president of India. The governor functions on the advice of the elected chief minister and the council of ministers. Manipur has a unicameral legislature, consisting of a legislative assembly of 60 members. The state shares a common high court with Assam. Manipur is divided into six administrative districts: Manipur North, Manipur South, Manipur East, Manipur Central, Manipur West, and Tengnoupal.

About two-fifths of the population is literate; the state has a university at Imphāl and more than 20 colleges. Classes in schools are taught in the Manipurī language, and English and Hindī are also taught.

Major health problems include tuberculosis, leprosy, venereal disease, and filariasis. Health facilities, though still inadequate, are expanding.

Cultural life. Polo and hockey are popular sports. Manipur has given birth to an indigenous form of classical dance known as Manipurī dance. It is dissimilar to other Indian dance forms; hand movements are used decoratively rather than as pantomime, bells are not accentuated, and both men and women perform communally. The dance dramas, interpreted by a narrator, are a part of religious life. Themes are generally taken from the life of Krishna, the pastoral god of Hinduism. Long an isolated art form, Manipurī dance was introduced to the rest of India by the poet Rabindranath Tagore in 1917.

HISTORY

Ancient hoes of ground stone reportedly have been found in the region. The earliest recorded history goes back to AD 900. The beginning of Manipur's more recent history dates from 1762, when the raja Jai Singh concluded a treaty with the British to repel an invasion from Burma. Further communication was minimal until 1824, when the British were again requested to expel the Burmese. Disputed successions were a continual source of political turmoil until Chura Chand, a five-year-old member of the ruling family, was nominated raja in 1891. For the next eight years the administration was conducted under British supervision; slavery and forced labour were abolished and roads were constructed.

British intervention

In 1907 the government was assumed by the raja and the durbar, or council, whose vice president was a member of the Indian civil service. Subsequently, the administration was transferred to the raja, and the vice president of the durbar became its president. After an uprising of the Kuki hill tribes in 1917, a new system of government was adopted; the region was divided into three subdivisions, each headed by an officer from the neighbouring government of Assam.

With the accession of Manipur to India in 1947, the political agency exercised by Assam was abolished. Two years later the administration was taken over by the union (central) government, and the region was ruled as a union territory by a chief commissioner and an elected territorial council. In 1969 the office of chief commissioner was replaced by that of lieutenant governor, whose status was changed to governor when Manipur became a full-fledged state of the Indian Union on July 21, 1972.

(B.A.St.)

Meghālaya

Meghālaya—literally Abode (*ālaya*) of the Clouds (*megha*)—situated in the northeast corner of India, was

inaugurated as a full-fledged state of the Indian Union in 1972. Occupying a mountainous plateau of great scenic beauty, it is bounded on the south and southwest by Bangladesh and on the northwest, north, and east by the Indian state of Assam.

The area of Meghālaya is 8,683 square miles (22,489 square kilometres). Its capital is the beautiful hill town of Shillong.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Geologically, Meghālaya is a detached extension of the Indian plateau. From west to east, the Gāro Hills rise abruptly from the Brahmaputra Valley to about 1,000 feet and merge with the Khāsi Hills and Jaintia Hills, two adjacent highland systems that form a single massif consisting of a series of eastward-trending ridges with tablelands between. The massif rises to 4,000–6,000 feet, declining toward the east. There are steep slopes on the southern faces of the plateau and comparatively gentle slopes toward the Brahmaputra Valley in the north. The central plateau acts as a watershed for the many rivers and streams traversing the state, the most important being the Umiam-Barapani, which is the major source of hydroelectric power for Assam and Meghālaya.

Rainfall

One of the world's wettest regions is found in the state—Cherrapunji, with an average of 450 inches (11,430 millimetres) of annual rainfall over a 74-year period, the highest ever recorded in Asia and the second highest in the world. (Rainfall at Cherrapunji may be exceeded, however, by that at Mawsynram, where an average of 700 inches [18,000 millimetres] has been claimed, but there is no official observation station at Mawsynram.)

The climate is mild, with an average minimum temperature in the Gāro Hills of 39° F (4° C) and in the Khāsi and Jaintia Hills, 54° F (12° C). The average maximum temperatures are 93° F (34° C) and 74° F (23.3° C), respectively.

The state is rich in forests—pines, sals, and bamboo are plentiful. Among the other trees are oaks, birches, beeches, and magnolias. The state abounds in elephants, tigers, leopards, deer, wild pigs, wild buffalo, mithun (wild oxen), wolves, anteaters, monkeys, apes, squirrels, snakes, hares, and sambar deer. Peacocks, partridges, pigeons, hornbills, jungle fowls, mynas, parrots, and other birds are found in the forests.

The people. Most of the inhabitants of Meghālaya are Tibeto-Burman (Gāros) or Mon-Khmer (Khāsis) in origin, and their languages and dialects belong to these groups. Whether the Mon-Khmers originally belonged to this region and were subsequently driven away to Southeast Asia or migrated from there, they are the only people in India speaking languages of the Mon-Khmer group. The most commonly spoken languages and dialects are Khāsi and Gāro, followed by Pnar-Synteng, Nepālī, Jaintia, and Haijong, in addition to the plains languages of Bengali, Assamese, and Hindī.

Christianity, Hinduism, and animistic forms of Hinduism are the major religions in the area. There are also some Muslims and a few Buddhists.

The population is predominantly rural and the state has few industries. The important urban centres are Shillong, Jowai, Nongthymmai, and Tura.

Mineral resources

The economy. Meghālaya is blessed with abundant natural but untapped resources. It has rich mineral resources—coal, limestone, kaolin, feldspar, quartz, mica, gypsum, bauxite, and other minerals. The deposits of sillimanite (a source of high-grade ceramic clay) are reputedly the best in the world. The state produces almost all of India's sillimanite output. There are no heavy industries in Meghālaya; the small-scale industries include cement, plywood, and beverage factories. Agriculture is the mainstay of the populace. Land is owned in common, but the primitive *jhum* form of shifting cultivation (burning of trees and planting the clearings in a cyclical operation) has left the people very poor and has eroded the soil. The main products are rice, millet, corn (maize), potatoes, pepper, chilies, cotton, ginger, jute, betel nuts, oranges, mangoes, bananas, pineapples, and other vegetables.

There is little communication among the five districts, and within the districts many of the valleys are isolated. There are no railways in the state. A national highway runs through the state from Gauhāti (Assam) in the north to Karīmganj (Assam) in the south.

Administrative and social conditions. *Government.* Like other states of the Indian Union, Meghālaya has a governor, appointed by the president of India, and a council of ministers, headed by a chief minister. The ministers are collectively responsible to the legislative assembly, which is elected on the basis of adult franchise. Meghālaya shares with Assam a common governor, high court, and some other administrative organs.

Administrative associations with other Indian states

Education and welfare. The state is one of the most underdeveloped areas of the country. About a third of the people are literate. The North-Eastern Hill University is at Shillong; the state has about 15 colleges that offer degrees. The 1947 partition of the subcontinent into India and Pakistan disrupted the tribal populations, many of whom suddenly found themselves on different sides of the international frontier between India and what was then East Pakistan (now Bangladesh). There were also tribal migrations from East Pakistan into India.

Cultural life. The area is rich in tribal culture and folklore. The advent of Christianity and its strict rules of Judeo-Christian ethics and morality have disrupted many of the tribal and communal institutions, however. The folk dances and music of the area are rich and colourful. Drinking and dancing to the accompaniment of music from buffalo horn *singas,* bamboo flutes, and drums are integral parts of religious ceremonies and social functions. Marriages are exogamous. One of the most curious and extraordinary customs among the Gāros is that after marriage the youngest son-in-law comes to live in his wife's parents' house and becomes the father-in-law's *nokrom,* or a representative of the father-in-law's clan in the mother-in-law's family. After the death of the father-in-law, the *nokrom* marries (and the marriage has to be consummated) the widowed mother-in-law, thus becoming the husband of both mother and daughter. The custom is now falling into disuse. The Khāsis once practiced human sacrifice, but this is no longer so.

There is one daily newspaper and several weeklies. The Shillong station of All-Indian Radio serves the region.

HISTORY

Meghālaya consists of five districts, the East and West Gāro Hills, the East and West Khāsi Hills, and the Jaintia Hills. It is inhabited mainly by tribal people, many of whom trace their origins to pre-Aryan times in India. Though the Aryan invaders drove the inhabitants from the plains to the surrounding hills, these people never really came under the rule of any alien group and largely managed to preserve their way of life in seclusion. Even the British, who established their sway over all of India, did not really conquer them and, in return for acceptance of overall British control, left the tribes alone. After Indian independence, the control and rule of the central government was slowly extended and expanded. But India's first prime minister, Jawaharlal Nehru, evolved a policy to preserve and protect the way of life of the tribal peoples. With other tribal areas, the region was given special protection in the Indian constitution, and, while included within the state of Assam, it retained a great deal of autonomy.

Even this did not prove enough, however, and when Assam, in 1960, introduced Assamese as the state's official language, agitation for autonomy and self-rule gathered strength. After several commissions and committees had studied the question, in April 1970 the government of India created Meghālaya as an autonomous state within Assam, and in December 1971 the Indian Parliament enacted a law making the area a separate state and bringing within its ambit the town of Shillong. It officially became a state on January 21, 1972. (C.Ra.)

Mizorām

Mizorām, a union territory in northeastern India after 1972, became a state in 1987. It was formerly the Mizo Hills District of the state of Assam. Up until 1954 it had

still another name, the Lushai Hills District. It has an area of 8,141 square miles (21,087 square kilometres) and is bounded by Burma to the east and south, by Bangladesh and Tripura to the west, and by Assam and Manipur to the north.

PHYSICAL AND HUMAN GEOGRAPHY

Geologically the Mizo Hills form a part of the Arakan arc, which consists of compact parallel ridges with a north–south axis formed of Tertiary sandstone, limestone, and shales. The ridges reach up to 7,000 feet. The hills are covered with thick evergreen forest containing valuable timbers and wild animals including elephant, rhinoceros, tiger, bear, and deer. In the intermontane valleys *jhum* tillage (whereby tracts are cleared by burning and sown with mixed crops) and terrace cultivation are practiced. Rice, corn (maize), cotton, and vegetables are the main crops. There are no major industries. Aizawl and Lungleh are connected by a motor road with Silchār in Assam. With the independence of India in 1947 a road-building program was undertaken with the primary object of completing a north–south route from Silchār through Aizawl to Paletwa in the Arakan state of Burma. There are no railways. Silchār has the nearest airport.

The population is composed of numerous groups, loosely called Mizo, a local term meaning "highlanders." They speak a variety of Tibeto-Burman dialects; some tribes are so isolated that their dialects are unintelligible to groups living in neighbouring valleys. Mizo and English are the principal languages. It is estimated that more than three-fourths of the population is Christian. The literacy rate of 60 percent is among the highest in India.

HISTORY

The area was originally inhabited by a small tribal group of Tibeto-Burman stock. Between 1750 and 1850 the Mizo migrated from the Chin Hills and subjugated the aborigines, assimilating these similar tribes into their own society. The Mizo have a well-developed system of chieftainship, and the state is divided into about 300 hereditary chieftainships. The Mizo are of the same kin as the Kuki, who are their neighbours but are fewer in number. A large percentage of the Mizo are literate, the majority of these being Christians. The most powerful political organization of the Mizo is the Mizorām Union. It stands for the liquidation of the chieftainship and the establishment of a democratic land administration. At first Mizo Hills was divided into two administrative units, the southern unit administered by the government of Bengal and the northern by that of Assam. In 1898 the two parts were amalgamated into one administrative district and put under the government of Assam. Mizorām enjoyed district council administration under the Indian constitution until it became a union territory in 1972. There are three administrative districts; Aizawl, Lungleh, and Chhimtuipui.

(Ed.)

Nāgāland

The state of Nāgāland occupies a region of forested hills and mountains between Burma and the Brahmaputra Valley of upper Assam. It is bordered on the east by Burma, on the south by the Indian state of Manipur, on the west and northwest by the Indian state of Assam, and on the northeast by the Tirāp district of the Indian state of Arunāchal Pradesh. One of the smallest Indian states, its total area is 6,401 square miles (16,579 square kilometres). The capital is Kohīma.

How the word Nāga came into use to designate the many tribes and subtribes that occupy the area is still a matter of conjecture. Some derive the name from Sanskrit *nāga*, "snake"; some from Hindī *naṅgā*, "naked"; and some from *naga*, "belonging to the hills," "highlander." The most widely accepted view is that it is derived from *nok*, meaning "people," or "folk." The people once called themselves by their tribal names, such as Angāmi, Sema, or Āo, but the upsurge of nationalism that led to statehood made Nāga so popular that it became a matter of pride to be so designated.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Except for a few square miles of plain, the entire state is covered with ranges of hills that geologically are a part of the Himalayan system. They consist of sedimentary formations of the Tertiary Period (formed up to 65,000,000 years ago). The Nāga Hills rise abruptly from the Brahmaputra Plain in the north, initially to low hills of about 2,000 feet, which increase in height toward the east to more than 6,000 feet, with isolated peaks reaching 10,000 feet or more. The hills merge with the Pātkai Range, which separates Nāgāland from Burma and which reaches a height of 12,552 feet (3,826 metres) at Mt. Saramati, on the border. The northwest is drained by tributaries of the Brahmaputra. The southeast, comprising most of the Tuensang and Mon districts, and the southeastern part of Kohīma district, has an elevation of more than 4,000 feet and is drained by tributaries of the Chindwin River of Burma. The headstreams of the Barāk flow through the southwest. The hills, clothed for the most part with dense forest, scrub, and grassland, are deeply dissected, though some slopes provide suitable sites for shifting cultivation. The higher slopes are covered with shallow, acidic, brown forest soils whereas there are comparatively deeper soils on the lower slopes. Relief and drainage

Nāgāland has a wet monsoon climate, with the average annual rainfall varying between 70 inches in the southeast and more than 100 inches in the northeast. July and August are the rainiest months. The temperature is relatively low. Varying according to altitude, the summer temperature ranges between 70° F (21° C) and 104° F (40° C), while in winter it rarely goes below 39° F (4° C), although frost is common over higher elevations. Because of the high humidity, the summer is sultry, especially over the lower hills and in the valleys.

The hills are covered with luxuriant forest approximating in type that of the eastern Himalayas, except that there is no true alpine zone. The hill forests may be separated generally into evergreen, broad-leaved, and pine forests. But where the forests have been cleared for *jhum* (shifting cultivation), high grass, reeds, and scrub jungles have sprung up in profusion. There are many species of palm and several of calamus (climbing palms, or rattan). Below 4,000 feet occur wet monsoon and evergreen forests, with mahogany and other valuable timber species and varieties of bamboo. On the higher hills in the east and southeast, oak, pine, and other coniferous vegetation appear. The surveyed forest area of the state is about 712,000 acres.

In the lower hills, especially in the north and northeast, are found rhinoceroses, elephants, tigers, leopards, bears, sambar deer, buffalo, boars, and wild oxen. Also seen are several kinds of monkey, including the long tailed langur, flying lemur, and pink faced monkey. Porcupines, pangolins (scaly anteaters), goats, wild dogs, foxes, civets, and mongooses are also found. Game birds include wild fowl, partridges, woodcocks, and pheasant. The great Indian hornbill is the most treasured bird—for use in decoration and magic. The gayal, a domesticated breed of wild ox, is a prized animal. The colourful smaller animals and the wide variety of birds make the jungle attractive for tourists. Animal life

The people. The Nāgas belong to the Indo-Mongoloid ethnic group. They have long been isolated and they lack cohesion as a single people—there are more than 15 major tribes and several subtribes. Though all have much in common, they differ in physique, dialect, and customs. The tribal groups have formed new affiliations and are using names hitherto unknown. The Chakhesangs, for example, are a combination of Chakru, Khezha, Sangtams, and a few Rengmas. Similarly, the Zeliangs are a mixed group of Zemis, Liangmais, and others. The big Konyāk tribe has two sharply distinguished divisions, one ruled by powerful autocratic chiefs and the other, which now calls itself Shaymnyuyungmang, more democratic.

The Angāmis, Chakhesang, Zeliangs, Rengmas, and Kukis are southern tribes living in the Kohīma and Phek districts. The Tuensang and Mon districts in the east are the home of the Konyāks and the Upper Konyāks, occupying the north and northeast, and the Phoms, Changs, Yimchungers, Khemungans, Sangtams, and a large number of subtribes, occupying the west and south. The major

tribes in the western districts of Mokokchūng, Wokha, and Zonheboto are the Āos, Lhotās, and Semas.

The Nāga tribes do not have a common language, and there are about 60 spoken dialects. In some areas dialects vary even from village to village. The languages all belong to the Sino-Tibetan family. Intertribal conversation generally is carried on through broken Assamese, which is also now commonly used and is called Nagamese. Many Nāgas have also become acquainted with Hindī and English. English is the official language of the state.

In the traditional religions of the Nāga tribes, special status is given to the sun and the moon. There is a conception of a supreme creator, and there are many minor deities. Nature is seen to be alive with invisible forces, with which priests and medicine men mediate. They also lead rites and festivals, bless the marriage bed, and protect the craftsman at his work. Nāga ideas of afterlife vary from tribe to tribe, but there is wide agreement that the soul does not perish at death.

With the advent of British rule, Christianity was introduced. The American Baptist Foreign Mission and various church unions were active, especially in the Kohīma, Phek, Mon, and Mokokchūng districts. Out of the total population, about two-thirds are now Christian, a fair number are Hindu, and less than 1 percent are Muslim, the rest belonging to other faiths.

Nāga villages

Nāgāland is extremely rural. About 80 percent of its population live in isolated villages. The villages are normally large, many having more than 5,000 inhabitants. Their size and singular character is striking even in the wilder Tuensang border ranges. Built on the most prominent points along the ridges of the hills, they were formerly stockaded; some had village gates—massive wooden doors approached by narrow, sunken, winding paths. The Āos and Lhotās set out their houses in streets, frequently along the top of a ridge. The villages are usually divided into *khels,* or quarters, each with its own headmen and administration. Many houses are built on stilts. Different areas are dominated by different tribal groups, which have well-defined portions of lands and sources of water. Tuensang and Mon districts are comparatively densely populated. Kohīma, Mokokchūng, and Dimāpur are the only urban centres.

The economy. Agriculture is the mainstay of the economy. About 90 percent of all workers are cultivators. Autumn and winter rice, small millets, pulses (edible seeds), oil seeds, fibres, sugarcane, potato, and tobacco are the principal crops. Production, however, is inadequate, and Nāgāland has to depend largely on imports from neighbouring states. This deficiency may be attributed to inadequacy of fertile, cultivable land and the common practice of shifting, or *jhum,* cultivation. Under shifting cultivation the jungle on the hill slopes is cleared and the undergrowth is burned down. The ground is then lightly hoed and sown for two successive years. After two harvests, the land is allowed to remain fallow for eight to ten years. This practice has led to soil erosion and permanent loss of fertility of the soil. In higher or otherwise unsuitable regions, *jhum* is replaced by terrace cultivation. The Angāmis of Kohīma district excel in the art of terracing and irrigation. The implements of the cultivator include the light hoe, *dao* (a multipurpose heavy knife), and sickle. Except in the plains, the plow is not used. A number of seed and sericulture farms and fruit nurseries have been established in the state. Agriculture is undergoing slow but steady transformation, and wasteful practices are being progressively replaced by improved methods.

Cultivation practices

Forestry is a primary source of income and employment. Quantities of fine timber, such as mahogany, are produced, as well as timber for fuel and forest products.

It is believed that there are rich deposits of minerals in Nāgāland, but limestone and low-grade coal are the only two known at present. The limestone, ranging in colour from light gray to a deep blue, makes an excellent building stone. Low-grade coal occurs in the Wokha, Mokokkchūng, and Nāginimāra areas. Nāginimāra is the only producing area at present but proposals have been made to mine coal from other areas. Many of the coal measures contain iron ore in the form of clay, ironstone, and impure limonite. Salt is locally produced from brine wells. Petroleum seepages occur in the Dikhu Valley, and the geological homogeneity of northern Nāgāland with the oil and coal-bearing area of Upper Assam suggests the presence of exploitable oil reserves.

Lack of raw materials, financial resources, and power and, above all, poorly developed transport and communication have hindered the development of industry. There are in the state some small bakeries, motor vehicle service and repair shops, rice mills, and cabinet and furniture factories. The sugar mill at Dimāpur also manufactures alcohol for wine production; it was the first manufacturing industry of any importance in the state. A paper and pulp factory and a plywood factory have been set up at Dimāpur. A small fruit-canning centre is located at Chanki.

Electric power generation started only after World War II. Two small hydroelectric schemes begun by Nāga entrepreneurs were subsequently taken over by the state government; Nāgāland's first hydroelectric plant started operating in 1978. All the district headquarters, subdivisional administrative centres, and almost one-half of the villages have electricity.

Even though the state is devoid of manufacturing industries, cottage industry is common. Nāgas are noted weavers and expert in basketry, woodwork, and pottery. They weave beautiful shawls, handsome ties, and colourful skirts. Cottage industry, however, suffers from lack of raw materials and improved tools, competition from outside, and the absence of a sufficiently wide market.

Nāgāland depends mostly on roads for communication. The Northeast Frontier Railway runs along the south bank of the Brahmaputra River and provides the single rail link to Nāgāland. It extends a little way across the Assam border. The total length of railway line within Nāgāland is only five miles; Dimāpur is the main railway station.

There are only a few all-weather roads. A national highway runs from Dimāpur through the southern part of the state, via Kohīma to Imphāl, the capital of Manipur. The second important road runs from Amguri, in the Sibsāgar district of Assam, to Mokokchūng, via Tuli and Merangkong village. The third important road links Tuensang with Mokokchūng. And Mokokchūng, Zonheboto, Kohīma, and Phek are also connected with short all-weather roads.

Administrative and social conditions. *Government.* In common with the other states of India, the administration of Nāgāland is conducted by a council of ministers, which is headed by a chief minister and is responsible to the Nāga legislative assembly. Appointed by the president of India, the governor is the constitutional head of the state. The state is divided into seven administrative districts—Kohīma, Mokokchūng, Mon, Phek, Wokha, Zonheboto, and Tuensang. The deputy commissioner is responsible for the maintenance of law and order, revenue, and criminal administration. Each district is divided into subdivisions headed by subdivisional officers responsible to the deputy commissioner.

The administration of Nāgāland differs from other states in that wider autonomy has been provided in the constitution to the tribal communities. Each tribe has three local councils—the village council, the range council, and the tribal council. These deal with disputes involving breaches of customary laws and usages. Appeals are made to the Nāga Tribunal in respect of cases decided according to customary law. Special provisions were made for Tuensang, which was not placed under the direct jurisdiction of the assembly but was put under a regional council consisting of elected representatives of all the tribes in the district, to which the governor might also nominate members. The deputy commissioner is its ex officio chairman. It elects representatives from among its members to the legislative assembly.

Provisions for tribal autonomy

Education and welfare. This hill and forest land, hitherto an underdeveloped region, has been witnessing a sociocultural renaissance. From its formation in 1963, the popular democratic government rapidly established a good measure of stability and directed its efforts toward improving social conditions through systematic planning. Under a special plan, inaugurated in 1958, the highest

item of expenditure was on roads, and greater emphasis was given to public health and water supply. Simultaneously, elaborate educational and community development programs were undertaken.

Because of the high incidence of tuberculosis, a hospital was established to treat victims of the disease. Malaria is another major health problem, and an antimalaria unit with trained personnel has been deployed. Under extensive community development programs, drinking water has been supplied to several hundred villages.

More than 40 percent of the population is literate. There has been much educational development, and the literacy percentage has been increasing at a rapid pace. There are more than 10 colleges for general education, all affiliated with the North-Eastern Hill University in Meghālaya.

The social and economic welfare activities of the state include consumer trading societies; farming and dairy, marketing, and other cooperatives; and adult literacy centres, libraries, youth clubs, and women's societies.

Cultural life. Nāga society contains varying patterns of local organization, ranging from near dictatorship to extreme democracy. Among the Semas and Changs there is a system of hereditary chieftainship. The Konyāks are ruled by powerful chiefs, or *āngs,* whose word is law and who are held to be sacred. The Āos have councils of elders representing the main family groups in the village. The Angāmis, Lhotās, Rengmas, and others are democratically run.

Tribal institutions and customs

A prominent village institution is the *morung* (communal house, or dormitory), for young unmarried men. Some tribes have communal houses for unmarried girls also. Formerly, skulls and other trophies of war were hung in the *morungs*. The pillars are still carved with striking representation of tigers, hornbills, and human and other figures.

The basic interest of every Nāga is in his family, the clan, the *khel*, and the village, and the Nāgas are passionately attached to their land and traditional customs. A central feature of their lives is the Feast of Merit, consisting of a series of ceremonies that finally culminate with the sacrifices of oxen (called *mithun* by the Nāgas).

Although there is no matriarchy among the Nāgas, women hold a high and honourable position. They work on equal terms with men in the fields and make their influence felt in the tribal councils.

The people have a rich vein of poetic thought that is little known outside the state. They have a passion for beauty and colour that inspires them to decorate almost everything they use or wear, and their poetry, folk songs, tales, and folklore all express an exuberant concern for life, romance, and a love of chivalry.

HISTORY

Nāgāland has no early written history. The first reference to the Nāgas is found in chronicles of the neighbouring Ahom kingdom of Assam. These chronicles tell of an invasion of the region in 1228 (although scholars consider a later date more probable) and contain information regarding the Nāga tribes, their economy and customs, and indicate that the Ahom did not rule the country but maintained friendly relations with the Nāga chieftains. There were occasional clashes between the Nāgas and the Ahoms, and a Burmese invasion of Assam in 1816 led to an oppressive Burmese rule from 1819 until the establishment of British rule over Assam in 1826, after which the Nāga hill areas were gradually annexed to the British Empire. The advent of British administration, which by 1892 encompassed the whole of Nāga territory (except the wild and rugged Tuensang area in the northeast), brought an end to the practices of headhunting and intervillage raids, and the Nāgas settled down to a more peaceful life of cultivation and trade.

Nāga separatism

When India became independent in 1947, a strong nationalist movement sought a union of the Nāga tribes, claiming that Nāgāland was a sovereign state separate from India, and a small armed force was organized among them. At the same time, however, there were many political differences among the Nāgas. In 1957, after negotiations with the first Nāga People's Convention at Kohima, the Nāga Hills district and Tuensang frontier division, a mutual agreement between Nāga leaders and the Indian government, were placed under central administration.

This status, however, proved unsatisfactory to a large segment of the Nāgas, and unrest continued, taking the form of noncooperation with the Indian government, nonpayment of taxes, and armed conflicts. A rebel force engaged in action against the Indian army and committed acts of sabotage. The Indian government and representatives of the Nāgas, however, continued to negotiate, and an agreement reached at the Nāga People's Convention meeting of July 1960 provided that Nāgāland should become a full state in the Indian Union. Nāga representatives found this solution acceptable because of the wide autonomy that would be granted to the new state. Nāgāland accordingly entered the Indian Union in 1963, and in 1964 a democratically elected government took office and proceeded to administer the state. Dissident factions within Nāgāland refused to accept this solution, however, and rebel activity continued, increasingly assuming the form of banditry and often motivated more by tribal rivalry and personal vendetta than by political aspiration. The state received substantial financial help from the central government, but relations were strained somewhat by the increasing presence of the Indian army in the territory. From June 1966 to March 1969 there was a ceasefire, and talks between the dissidents and the government of India continued throughout the period. In clashes that occurred after the breakdown of talks, evidence was produced suggesting that large numbers of Nāga dissidents had received training in China; it can be said with certainty that the Nāga underground had some contacts with the Chinese.

The dissidents became increasingly fragmented in the late 1960s, and several groups surrendered to the Indian army. In the late 20th century, however, a hard core of extremists were continuing their activities. (M.B.)

Orissa

Orissa is situated in eastern India. It is bounded by the Bay of Bengal in the southeast; by West Bengal in the northeast; by Bihār in the north; by Madhya Pradesh in the west; and by Andhra Pradesh in the south. Its area is 60,119 square miles (155,707 square kilometres). Many people speaking Oṛiyā (the language of Orissa) also live in the adjoining areas of neighbouring states. Before India became independent in 1947 Orissa's capital was at Cuttack. The present capital was subsequently built at Bhubaneswar, in the vicinity of its historic temples.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Physiographically, Orissa is of a heterogeneous character. Its geological formations vary from the oldest rocks of the Earth's crust in the stable landmass of the Indian Peninsula to deltaic alluvium or littoral deposits and ridges of windblown sand on the seaboard.

The four geographic regions

Broadly, the state can be divided into four natural divisions: (1) the Northern Plateau, (2) the Eastern Ghāts, (3) the Central Tract, and (4) the Coastal Plains. The Northern Plateau is an extension of the Chota Nāgpur plateau covering the Mayūrbhanj, Keonjhar, and eastern Sundargarh areas. The upper reaches of the rivers Subarnarekha, Baitarani, and Brāhmani lie in the area. Its central portion contains many small hills and forests. The Eastern Ghāts form an undulating plateau that extends over Koraput, parts of Kalāhāndi, Phulbāni, and Ganjām. The Central Tract covers the districts of Sambalpur, Bolāngir, and Dhenkānāl, and parts of Kalāhāndi, Sundargarh, and Phulbāni. The rivers Mahānadi, Brāhmani, and Baitarani and their tributaries flow through the area. The landscape consists of a succession of plateaus, hills, uplands, and valleys. The coastal plains comprise the Balasore, Cuttack, and Puri districts, and parts of Ganjām. This fertile area runs parallel to the coast and contains the river deltas; it is formed of alluvium and silt.

The main rivers are the Subarnarekhā, Burābalang, Baitarani, Brāhmani, Mahānadi, Rushikulya, and Vamsadhāra. Notable mountain ranges are the Mahendra Giri (rising to 4,924 feet), the Malyagiri (3,896 feet), and the

Meghasani (3,823 feet). Orissa's Chilka Lake is the biggest saltwater lagoon in India.

Orissa is situated in the hot climatic belt where the monsoon winds blow. The main seasons are summer from mid-February to June, the rainy season from July to October, and winter from November to mid-February. Places in Sambalpur and Mayūrbhanj districts have an extreme climate, whereas in places like Puri and Gopālpur the climate is equable. The average rainfall is about 60 inches.

The vegetation is of a tropophilous character—that is to say it is able to adapt itself to a rapidly changing environment. Forests cover about 40 percent of the total land area. In forest wealth, Orissa ranks among the top states in India. The forests are of three broad types: northern tropical semievergreen; northern tropical moist deciduous; and northern tropical dry deciduous. The first group includes the mango; the second group includes, for example, teak; while the third includes bamboo.

Orissa has the same animal life as the rest of peninsular India. Monkeys are common. Carnivores include different types of tigers. The elephant, the wild buffalo, the blackbuck, and the four-horned antelope are found in some districts. The peafowl is one of the features of the Orissa forests. Lake Chilka is noted for its marine fauna.

The people. The population includes the Australoid, the Alpinoid, and the Mediterranean racial groups. The Australoids are represented mainly by the Savara, Santāl, Khond, Goṇḍ, Muṇḍā, Ho, Juāṅg, Gaḍabā, Bhuina, Koya, and Oraon tribes. These tribes are divided into three linguistic groups: the Muṇḍā-speaking (the Santāl, Savara, Juāṅg, and others); the Dravidian-speaking (the Khond, Goṇḍ, Oraon, and others); and the Oṛiyā-speaking (the Bhuina, and others). Most tribes live in the hill areas, but they are also found in the plains. The Alpinoid and Mediterranean groups are represented among the nontribal population, which is mainly Oṛiyā-speaking and Hindu.

The tribes have for a long time been going through the process of Hinduization. Tribal chieftains have claimed Kṣatriya (warrior) status, while many of the Khonds, who constitute the largest tribe, have abandoned their Kui language (Dravidian) and speak Oṛiyā. Many tribes are bilingual. Some have become almost indistinguishable from the Hindus and have lost their original language. The Oṛiyā language does not vary in its written form, but it differs in the spoken form from place to place. The purest Oṛiyā is spoken in the Cuttack and Puri districts. The Balasore, Sambalpur, and Ganjām districts have distinct local accents.

The caste structure

A majority of the people are Hindus. The caste structure is the same as in other states of eastern India, with some regional variations. Next to the Brahmins are the Karaṇas (the writer class), who claim Kṣatriya status (with the pen as their weapon rather than the sword). The Khandayats (literally, "swordsmen") are mostly cultivators but call themselves "Khandayat-Kṣatriyas." All castes look to Jagannātha (one of the forms of the Hindu god Vishnu) as the centre of their religious faith. For centuries Puri, the abode of Jagannātha, has been the only place in India where all castes, including the so-called untouchables, eat together.

Orissa has a predominantly rural population. The building of the new capital at Bhubaneswar, of a steel town at Raurkela, of a port at Parādīp, of a dam at Hīrākud, and of a power-generating station at Tālcher have necessitated new urban settlements, and a number of industrial towns are under construction. The major cities are Cuttack, Raurkela, Bhubaneswar, Chowduār, and Berhampur.

The economy. About 80 percent of the rural population depends on agriculture, even though some areas are unproductive and some others are unsuitable for more than a single annual crop. Orissa produces about one-tenth of the total rice output of India. Other important crops are oilseeds, pulses, jute, sugarcane, and coconut. The nonagricultural classes live by handicrafts, by trade, and by the rendering of services. Some rural families depend on village or cottage industries, including weaving.

A number of agricultural families engage in nonagricultural pursuits by way of subsidiary employment. Whatever their occupation, most rural people do not get continuous employment the year around. The amount of agricultural land available per person has been declining because of the growth of population. The search for new lands for cultivation, the partial abolition of large landholdings, and the distribution of available land among landless families have been some of the acute problems faced by the state government in the late 20th century.

Mineral deposits

The industrial resources of Orissa are considerable. Because most mineral deposits are in the former feudatory states, their exploitation began only after 1947. In the late 20th century Orissa was ahead of all other Indian states in the production and export of iron ore and chromite and took a prominent place in the production of manganese. It also produced coal, limestone, dolomite, graphite, mica, and bauxite.

Large-scale industries, mostly based on minerals, include a steel plant and fertilizer plant at Raurkela, ferromanganese plants at Joda and Rāyagada, refractor-producing factories at Raj Gāngpur and at Belpahār, a refrigerator manufacturing plant at Chowduār, and a cement factory at Raj Gāngpur. There are some large scale sugar and paper mills at Rāyagada and Chowduār; other industries include textiles, glass, aluminum ingots and cables, and heavy machine tools.

The Hīrākud Dam Project on the Mahānadi River is one of the biggest multipurpose river-valley projects of India; the Machkund hydroelectric project and the Tālcher thermal project are other important power projects of Orissa.

Communication facilities were undeveloped before 1947, but the merger of the feudatory states with Orissa and the discovery of mineral resources necessitated the construction of a network of good roads. During the five-year plans, bold construction programs were undertaken and bridges were built over most of the major rivers, but the state still lacks adequate railway communications.

An all-weather, sheltered, deep-draft port has been constructed at Parādīp at the mouth of Mahānadi River. The port has an extensive hinterland with mineral deposits and a developing industrial belt.

Administrative and social conditions. The head of the state is a governor appointed by the president of India. The actual administration is conducted by a council of ministers, headed by a chief minister and responsible to the elected legislature, which consists of only one chamber elected at intervals of not more than five years through universal adult suffrage.

The districts of Orissa

There are 13 districts: Bālāsore, Bolāngir, Cuttack, Dhenkānāl, Ganjām, Kalāhāndi, Keonjhar, Koraput, Mayūrbhanj, Phulbāni, Puri, Sambalpur, and Sundargarh—grouped into three revenue divisions, each under a divisional commissioner. A board of revenue is in charge of revenue administration. The district administration is conducted by a deputy commissioner who is also the district magistrate. Each district has a superintendent of police. The districts are divided into *tahsils,* each having a *tahsildar* as its revenue officer. *Tahsils* comprise groups of villages, administered by *pañcāyats* (village councils), to which villagers elect their representatives. A *sarpañc* (elected president) heads the *pañcāyat.* The system represents a democratic decentralization of power for the benefit of the rural population. The towns are administered by municipalities.

Health. At one time the coastal belt was highly malarious and the whole state was subject to epidemics of cholera and smallpox. The incidence of filariasis (a disease caused by the presence of filarial worms in the blood and glands), leprosy, and tuberculosis was high. The high rate of mortality even led to a decrease of population in certain places. Since independence, however, much attention has been paid to health services, and excellent results have been achieved by various programs. The number of hospitals and dispensaries has increased; and the three medical colleges, at Cuttack, Berhampur, and Burla (Sambalpur), have expanded considerably.

Education and welfare. The number of educational institutions increased considerably after 1947. There are five universities including an agricultural university and many colleges for general education.

Orissa has a Tribal Welfare Department. Schemes have

been devised to promote the educational, cultural, economic, and social advancement of the tribes. A Tribal Research Bureau and a Tribal Research Laboratory have been established at Bhubaneswar to collect data about tribes; this information assists the state government in formulating plans and policies regarding tribal welfare. The state Social Welfare Advisory Board, instituted in 1954, cares for the welfare of women and children through courses of instruction, urban-welfare-extension projects, and holiday camps for children.

Cultural life. Orissa has a rich artistic heritage and has produced some of the best examples of Indian art and architecture. Artistic traditions are maintained through mural paintings, stone carving, wood carving, icon paintings (known as *patta* paintings), and paintings on palm leaves. Handicraft workers are famous for their exquisite silver filigree ornamentation and decorative work.

In tribal areas Orissa has a wide variety of folk dances. The music of the *mādal* and flute is common in the countryside. The classical dance of Orissa, known as the *oṛissī* dance, has survived for more than 700 years. Originally it was a temple dance, performed before gods. The modes, movements, gestures, and poses of the dance are depicted on the walls of the great temples, especially at Konārak, in the form of sculpture and in relief carvings. Modern exponents of the dance have made it popular outside the state. The *chhau* dance (performed by groups of masked dancers) of Mayūrbhanj and Saraikela regions is another feature of Oṛiyā culture. For the promotion of dancing and music, the Kala Vikash Kendra centre was founded at Cuttack in 1952 with a six-year teaching course. The National Music Association serves a similar purpose. Other notable dance and music centres in Cuttack are the Utkal Sangit Samaj, the Utkal Smruti Kala Mandapa, and the Mukti Kala Mandir.

The Boita-Bandana ceremony

There are many traditional festivals. A unique festival is the ceremony of Boita-Bandana (worshipping of ships). In October–November for five consecutive days before the full moon, people gather near riverbanks or the seashore and float miniature boats, as a symbolic gesture that they will leave for the faraway lands (Malaysia and the East Indies) to which their ancestors once sailed. The greatest religious festival of Orissa is the Chariot Festival of Jagannātha at Puri, which attracts hundreds of thousands of people from all parts of India.

HISTORY

The land corresponding roughly with modern Orissa, but at times much larger in area, passed under the names of Utkala, Kaliṅga, and Oḍra Deṣa in ancient and medieval times. In origin, these names were associated with peoples. The Okkalā or Utkala, the Kaliṅgā, and the Oḍra or Oḍḍakā were mentioned in literature as tribes. Ancient Greeks knew the latter two as Kalingai and Oretes. Ultimately the names became identified with territories. For centuries before and after the birth of Christ, Kaliṅga was a formidable political power, extending from the Ganges River to the Godāvari. Approximately between the 11th and 16th centuries the name fell into disuse; instead, the name Oḍra Deṣa was gradually transformed into Uḍḍiṣa, Uḍisā, or Oḍisā, which in English became Orissa. The language of Oḍisā came to be known as Oṛiyā or Oṛiā.

At the dawn of Indian history, therefore, Kaliṅga was already famous. Buddhist sources refer to the rule of King Brahmadatta in Kaliṅga at the time of Buddha's death. In the 4th century BC the first Indian empire builder, Mahāpadma Nanda, conquered Kaliṅga, but the Nanda rule was short-lived. In 261 BC the Indian emperor Aṣoka invaded Kaliṅga and fought one of the greatest wars of ancient history. He then renounced war, became a Buddhist, and preached peace and nonviolence in and outside India. In the 1st century BC the Kaliṅga emperor Khāravela conquered vast territories that collectively came to be called the Kaliṅga Empire. In the early centuries of the Christian Era Kaliṅga was a maritime power. Its overseas activities culminated in the 8th century AD with the establishment of the Ṣailendra Empire in Java. During the 8th, 9th, and 10th centuries Orissa was ruled by the powerful Bhauma-Kara dynasty, and in the 10th and 11th centuries by the Soma dynasty. The Temple of Liṅgarāja at Bhubaneswar, the greatest Ṣaiva monument of India, was begun by the Soma King Yayāti.

The Gaṅga dynasty

Medieval Orissa enjoyed a golden age under the Gaṅga dynasty. Its founder, Anantavarma Cōḍagaṅgadeva (1078–1147), ruled from the Ganges to the Godāvari with Cuttack as his capital. He began the construction of the Temple of Jagannātha (Lord of the Universe) at Puri. Narasiṃhadeva I (1238–64) built the Sun Temple of Konārak, widely acknowledged as the finest specimen of Hindu architecture. In the 13th and 14th centuries, when the whole of India was overrun by the Muslims, independent Orissa remained a citadel of Hindu religion, philosophy, art, and architecture. The Gaṅgas were succeeded by the Sūrya dynasty. Its first king, Kapilendradeva (1435–66), won territories from his Muslim neighbours and greatly expanded the Orissan kingdom. His successor, Puruṣottamdeva, maintained these gains with difficulty. The next and the last Sūrya king, Pratāparudradeva, became a disciple of Caitanya, the great medieval saint, and became a pacifist. After his death (1540) Orissa declined, and 1568, when King Mukundadeva was killed by his own countrymen, Orissa lost its independence to the Afghān rulers of Bengal.

The Mughal emperor Akbar conquered Orissa from the Afghāns. When the Mughal Empire fell, part of Orissa remained under the Bengal Nawabs, but the greater part passed to the Marāṭhās. The Bengal sector came under British rule in 1757 after the Battle of Plassey; The Marāṭhā sector was conquered by the British in 1803.

Linguistic basis of unification

Thus, the British did not conquer the entire Oṛiyā-speaking area at the same time; nor did they administer the whole area, when conquered, as one unit. Consequently, there arose a demand for unification on a linguistic basis, and as a result Orissa was constituted a separate province on April 1, 1936. Even then, 26 Oṛiyā princely states remained outside the provincial administration. After independence in 1947, however, all those states, except Saraikela and Kharsāwān (which merged with Bihār), were included in Orissa. The Constitution of India (1950) recognized Orissa's statehood. (M.N.D.)

Sikkim

Sikkim, the smallest constituent state of the Republic of India, covers an area of 2,783 square miles (7,208 square kilometres) in the eastern Himalayan mountains. It is bordered by the independent Himalayan kingdoms of Nepal, to the west, and Bhutan, to the east; by the Tibetan Autonomous Region of the People's Republic of China, to the north and northeast; and by West Bengal state, India, to the south. Gangtok is the state capital.

Although Sikkim was long a sovereign state, its relations with India were governed from 1950 until 1975 by a treaty, under the terms of which India defended Sikkim's territorial integrity and accepted responsibility for its foreign relations and strategic communications. Formal union with India was established in 1975.

Sikkim has great extremes of altitude and climate within its borders; its lowest point is about 700 feet above sea level, and its highest is the peak of Kānchenjunga (28,169 feet [8,586 metres]), the highest mountain peak in India and the third highest mountain in the world.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Known in popular parlance as "a fabulous kingdom," Sikkim is surrounded by precipitous mountain walls on three sides and appears like "a pearl in a craggy oyster." The Singālila Range separates Sikkim from Nepal in the west, while the Dongkya Range forms the border with China to the north and northeast. Along the northeastern border, Sikkim is separated from Bhutan by the Chumbi Valley of Tibet. This Tibetan enclave has made the location of Sikkim strategically and politically sensitive out of all proportion to its size.

Sikkim is small, stretching about 70 miles from north to south and about 40 miles from east to west. Its horseshoe-shaped mountains offer no sizable flat area, and only in the south is the state not separated from its neighbour,

the Darjeeling district of West Bengal state, by a mountainous wall.

About two-thirds of the total territory of Sikkim consists of perpetually snow-covered mountains that act as snowfields for numerous glaciers. These lofty ranges are without permanent human habitation, although seminomadic herdsmen visit these areas with their yaks and sheep during the summer phase of transhumance. The most significant and noteworthy group of mountains is that dominated by Kānchenjunga. An expedition climbed the peak successfully in 1955 but left the last few feet of snow untrampled in deference to the Sikkimese belief that the mountain is both a god and the abode of gods. Tent (24,165 feet), Kabru (24,075 feet), and Panhunri (23,395 feet) are other important mountain peaks of Sikkim.

Kānchenjunga

Within the mountains on the east and west of Sikkim, there are deep valleys, of which the Tista Valley is the largest. On the whole, the altitude in Sikkim varies from 720 feet in the southern foothills to more than 27,800 feet along the northern and northwestern borders. A number of passes connect Sikkim with neighbouring countries.

Elevation, forest cover, and gradient affect the soils of Sikkim, which are extremely varied. The soils of the tropical forest zone are no doubt leached, but the decomposition of tree leaves, grasses, and ferns has enriched their humus content. In the temperate forest belt, alluvial soils are found along riverbeds, while in areas dominated by conifers and oaks, podzolized acidic soils and brown soils have developed, respectively. Active erosion on the steep slopes has led to the formation of immature soils. In higher reaches, fluvio-glacial soils are found on terraced patches.

The Tista River

Much of Sikkim forms the drainage basin of the Tista River. The river originates in the extreme northeast from the Zemu glacier above Lachen Gompha, near the Chinese frontier. From its source the river descends steeply, dropping about 15,700 feet within its 65-mile course to Rongphu (Rangpo). The Rangit, Rongni Chu, Talung, Ragpo, Lachen, and Lachung, all snow-fed torrents, are the major tributaries of the Tista. All these southward-sloping streams have cut deep, beautiful valleys with almost precipitous walls; several hydroelectric power projects supply electricity to Gangtok, Rongphu, Singtām, and Mangan. There are also a number of lakes and hot springs; of the latter, the Phut-Sachu, Ralong Sachu, Youmthang, and Momay springs are the most important.

The extremely varied climatic conditions of Sikkim result from variations in elevation. The climate ranges from the tropical heat of low valley bottoms to the alpine cold of the higher reaches. Up to an average altitude of 2,000 feet, moist tropical climate is experienced. Beyond that, a cool temperate climate prevails to an altitude of 6,600 feet. Thereafter, cold temperate climatic conditions are found up to 9,800 feet. Higher up, cold increases progressively to Arctic and polar conditions. Depending upon altitude and aspect, annual rainfall varies from 50 to 200 inches (1,250 to 5,000 millimetres). Because of the proximity of the Bay of Bengal and direct exposure to the effects of moist summer monsoons, Sikkim is one of the most humid regions in the Himalayas. Most of the annual rainfall is received during the months of May through October.

The copious rainfall causes numerous landslides that interrupt traffic and sometimes result in the loss of human life and property. In winter the higher reaches are covered with snow that, in many places, attains a thickness of about 100 feet. The heavy snow mass results in avalanches that often cause great human and material destruction.

Plant and animal life. Sikkim contains luxuriant, seemingly endless forests. About 4,000 species of plants are found in the tiny state, making it a botanist's paradise. The monsoonal rainfall is often accompanied by lightning, which adds nitric acid, a source of plant nutrients, to the soil and atmosphere.

The type of vegetation varies with altitude, and three broad vegetation categories are recognized in Sikkim. The subtropical type, which extends to a height of 5,200 feet, includes varieties of bamboos, ferns, tree ferns, canes, palms, sals, pandanus, and orchids. Twenty varieties of bamboos alone are found in this zone. Out of the known 5,000 species of orchids, 601 are found in Sikkim.

Bamboo and orchids

In the temperate forests (5,200 feet to 13,100 feet) oak, laurel, maple, chestnut, magnolia, alder, birch, rhododendron, fir, hemlock, and spruce form the predominant vegetation. The alpine forests (13,100 feet to above 16,400 feet) are characterized by such flowering plants as primulas, gentians, blue poppies, aconites, and Himalayan rhubarb. The primula flower is held in very high esteem in Sikkim and Bhutan.

Sikkim has a rich animal life, including about 81 species of mammals, 600 species of birds, 631 species of butterflies, and 2,000 varieties of other insects and creepers. Notable animals include the black bear, brown bear, and panda; the musk deer, sambar, barking deer, and Tibetan antelope; the tiger, leopard, leopard cat, and marbled cat; the goral, wild sheep, and wild goats; the wild boar; and the wild ass. In the upper reaches are tailless mountain rats, rock lizards, and the snow fox. The legendary Abominable Snowman, known as *Nee-gued* in Sikkim, is believed to roam the glaciers and ice slopes.

Birdlife includes pheasants, partridges, teal and other ducks, quail, eagles, barbets, Himalayan cuckoos, babblers, thrushes, Tibetan black crows, and miniwets. Trout and carp abound in Sikkim's streams and lakes.

The people. Most of Sikkim's inhabitants live in the central, western, and southern parts of the state. In the far north, the land is rugged, snowbound, inaccessible, and uninhabitable. As there is little flat land, villages are few, each consisting of many hamlets of small homesteads scattered over the slopes of the hills.

Only about a third of the total population has been classified as literate. Almost all the people are Buddhists and Hindus, although there are also adherents of Christianity and Islām.

Sikkim had no urban area until 1951, when the capital of Gangtok was accorded that status. Other towns today include Singtām, Rongphu, Mangan, Gyalshing, Jorthang-Nayabazar, and Manchi.

The economy. Considering its size, Sikkim is rich in mineral wealth. Copper, lead, zinc, coal, iron ore, garnet, graphite, pyrites, marble, and steatite have been found in the state. Copper mining was a traditional vocation in the region, abandoned long ago for unknown reasons. The Sikkim Mining Corporation, however, was established in 1960 and has done an appreciable job. Copper deposits are found in numerous locations, and coal deposits have been located in Rangit Valley. Graphite has been found near Chattery, while garnet is available in small quantities at Panchikhani. Iron ore has been found at Bhotang and lead-zinc ores in small quantities at Rongphu, Dichhu, and Jugdum. Limestone occurs at Rinchinpong and Changgu and steatite near Chakung. High grade deposits of gold, silver, and zinc in combined form have been located near Dichhu.

Copper mining

Forests, which cover more than a third of the state's territory, have immense potential economic value. There is a sawmill at Rongphu.

The economy of Sikkim is predominantly agricultural, although agriculture has had a checkered history in the state. The original inhabitants were not agriculturists, but were hunters, fishermen, and gatherers of wild roots and fruits. When the Bhutia migrated to Sikkim, they established both a semi-pastoral economy and slash-and-burn cultivation. It was the Nepalese immigrants who started settled cultivation in Sikkim. These sturdy, hardy, energetic, and innovative people cleared large tracts of woodlands to make the land fit for permanent cultivation and introduced terraced farming along the slopes. Corn (maize), paddy (rice), buckwheat, wheat, and barley are the principal cereals of the region. Sikkim is one of the world's main producers of cardamom. The region has also become an important exporter of mandarin oranges, apples, and potatoes. Beans, cabbage, ginger, cauliflower, tea, peas, and tomatoes are also grown.

Cardamom production

Animal husbandry also plays a major role in the economy. Though yet in a crude form, it is mixed farming that provides sustenance for a large number of people. Cattle, buffalo, yaks, sheep, goats, pigs, mules, ponies, and poultry are the important domesticated animals. Cattle and buffalo are limited mainly to the subtropical humid

belt. The dairy industry is assisted largely by the National Dairy Development Corporation.

Among the handicrafts, cotton and wool weaving; the making of carpets, rugs, and blankets; and bamboo work have a name in Sikkim. The Cottage Industries Institute and the Industrial Training Institute offer instruction in these pursuits, as well as in doll-making, ceramics, dyeing and batik, traditional scroll painting, embroidery, the preparation of handmade paper, canework, woodwork, tailoring, and toy-making. Sikkimese caps and ornaments are attractions for visitors. The Sikkim distillery, set up in 1956, has earned a reputation for its high-quality products. Other industries include a roller flour mill, a tannery, a watch assembly unit, a fruit preservation plant at Singtām, and a tea processing factory at Temi.

Roads, though not extensive, are the primary mode of travel. Ropeways have also been provided at many points. Gangtok is nearly 75 miles from the nearest airport at Bāghdogra, in West Bengal state, while Silīguri, 70 miles away and also in West Bengal state, is the nearest railway station to the Sikkim state capital.

Administrative and social conditions. Under the formal feudal system, the *chogyal* was the head of state of Sikkim. Since 1975, the Sikkim government, like that of any other state of India, has been headed by the governor, who is appointed by the president of India and who holds office during his pleasure. The governor appoints as his chief minister the elected leader of the majority party represented in the state's Legislative Assembly. The chief minister is assisted by a council of ministers.

The state Legislative Assembly consists more than 30 seats, which are equally apportioned between the Lepchā-Bhutia population and the Nepalese. One of the seats awarded to the Lepchā-Bhutia group is reserved for the nominee of the lāmās (Buddhist religious leaders), and one seat of the Nepalese group is reserved for the representative of the lower castes.

Local government

The state is divided into four districts with headquarters at Gangtok, Namchi, Mangan, and Gezing. Each district is headed by a district officer, who is also the district collector. A headman elected from each area serves as a liaison between the people and the district administration. Committees (*pañcāyats*) of elected representatives administer the villages and implement welfare schemes. The village system is particularly strong in the northern areas. Each district has a court with a judicial magistrate.

There are many primary and secondary schools, and no student lives more than two miles from a school. Both primary and secondary education is free. There is a teacher-training institute; a government-subsidized degree college located in Gangtok is affiliated with the University of North Bengal at Darjeeling.

Cultural life. Sikkim's cultural life is related to Tibetan religious and aesthetic traditions, but it retains a unique character from the various tribes of Sikkim and their pre-Buddhist customs. The cultural climax of the year is represented by the two-day Phanglhapsol festival, when masked dances are performed in honour of Kānchenjunga, the presiding deity and the mountain. There are also various folk dances. Music is an important part of the school curriculum. The Namgyal Institute of Tibetology has one of the largest collection of Tibetan books in the world. Many monasteries are repositories of artistic treasures, including wall paintings, tankas (religious paintings mounted on brocade), and bronze images.

HISTORY

Three tribes—the Naong, Chang, and Mon—are believed to have inhabited Sikkim in prehistoric times. The Lepchā entered Sikkim sometime later and completely assimilated with the earlier inhabitants, but they are generally thought of as indigenous to the land. The Lepchā were and continue to be peace-loving, deeply religious, and shy. They speak a Tibeto-Burman language, using a script that was devised by King Chagdor Namgyal at the beginning of the 18th century. Most Lepchā have accepted Mahāyāna Buddhism, although the pre-Buddhist Lepchā religion, Bon, has influenced Sikkimese Buddhism.

Other tribes that are either indigenous to Sikkim or that may have entered the country from the trans-Himalayas in the prehistoric period are the Magar and the Tsong. The Magar are renowned warriors who formerly were an important part of the Sikkim Guards. They are mentioned as one of the groups to have celebrated the coronation in 1642 of Phuntsog (Penchoo) Namgyal, the first *chogyal* (temporal and spiritual king) of Sikkim and founder of the Namgyal dynasty, which ruled Sikkim under a feudal system until 1975. The name Sikkim may have been derived from the Tsong term *sukkim,* meaning "new house" or "happy house."

The Bhutia migration

The Bhutia, a Tibetan-speaking group that entered Sikkim in the 13th century, previously lived a pastoral life on the plateaus of Tibet; they belonged to the Rnying-ma-pa (The Old Order), a liberal sect of Mahāyāna Buddhism. Their migration, which was from Kham in Tibet, was led by a prince of the Namgyal dynasty. In adapting to the lusher geography of Sikkim, most of them settled as farmers, although some still prefer to live in the high valleys of northern Sikkim, herding yaks and sheep. An aggressive people, they occupied lands with little resistance from the docile Lepchā. Over the years, the Bhutia came into increased contact with the Lepchā and intermarried with them.

The largest migration to Sikkim took place between the mid-18th and 19th centuries, when various tribes entered from northern and eastern Nepal. The first large Nepalese group was the Newār, a Tibeto-Burman people skilled in metalwork. The ruling *chogyal* granted them the right to mine copper and to mint Sikkimese coinage. The Newār were followed by the Gurung, Tamang, Rai, and Sherpa; these were peoples of Tibetan origin who were adept at farming at high altitudes. To accommodate this influx of settlers, new lands were opened to cultivation. As pressure on the land grew, new, intensive methods of farming were introduced, including the terracing of steep slopes. Other groups that migrated from Nepal included clans of the Chettri, the Bahun, and the Bishu Karma, all of whom were probably of Aryan origin.

The various Nepalese tribes spoke their own languages; at present most of them speak Nepālī (also called Gorkhālī). The newcomers were mainly Hindus of a peculiarly Himalayan type who brought with them a pantheon of both Hindu and Buddhist deities. Many clans worship deities of both religions; in Sikkim, as in Nepal, Buddhist and Hindu beliefs have traditionally interacted or amalgamated over the centuries.

Sikkim, being a small country, was obliged to struggle to maintain its territorial integrity. In the mid-18th century Sikkim fought a series of territorial wars with Bhutan. The expansionism of the Nepalese ruler Prithvī Nārāyan Shāh (reigned 1742–75) and his successors obliged Sikkim to defend its western borders, which had once extended as far as Ilām, Nepal. In a series of wars that were to last for nearly a century, Nepal briefly occupied part of western Sikkim, but was obliged to relinquish part of it in 1793 and another part, the submontane Tarai, in 1816.

Establishment of British influence

In 1839 the British East India Company obtained Darjeeling from Sikkim as a health resort. British advocates of expansionism policy then attempted to bring independent Sikkim within their sphere of influence and succeeded in annexing the submontane regions by 1849. The annexation was resisted but, after the defeat of Sikkimese troops by the British, an Anglo–Sikkimese treaty was signed in 1861. By the terms of that treaty, the British obtained the right to build a road through Sikkim to Tibet, while Sikkim's sovereignty was recognized, thus establishing its role as a buffer state between British India and Tibet. In 1890, after a series of Tibetan incursions across the frontier of Sikkim, a convention was concluded between the British and the Chinese, who then controlled Tibet. Under this agreement, the border between Sikkim and China was defined and China acknowledged the special relationship of British India with the kingdom of Sikkim. A British political officer was subsequently appointed to assist the *chogyal* in the administration of Sikkim's domestic and foreign affairs.

India attained independence in 1947, and a treaty was signed between Sikkim and India in 1950. By the terms

of the treaty, Sikkim was deemed to be an Indian protectorate, and an Indian representative, who was simultaneously accredited to Tibet and Bhutan, was sent to Sikkim. Thus India became responsible for the external relations, defense, and communications of Sikkim.

This period was a difficult one in Sikkim's history. Political parties were formed for the first time, social distinctions were sharply felt, and some factions exerted considerable pressure to persuade Sikkim to accede to India. In the political awakening that then swept Sikkim, political parties began to demand a more popular form of government, demands that the *chogyal* resisted. Prior to integration with India, general elections based on adult franchise were held in 1952, 1958, 1967, 1970, and 1973. In the last of these elections, two rival parties merged to form the Sikkim National Congress, which swept the polls. The winning party launched *satyāgraha,* a movement to obtain greater political liberties and rights that the *chogyal* attempted to suppress. Finding that a solution was out of his control, the *chogyal* asked the government of India to take over the administration. India prepared a constitution for Sikkim that was approved by the Sikkimese national assembly with eight amendments; this became the Government of Sikkim Act of 1974, under which fresh elections to the national assembly were held in 1974. Following passage of a resolution by the national assembly on April 10, 1975, and a special referendum on April 14, 1975, in which more than 97 percent of the electorate voted for the merger of Sikkim with India, the government of India approved the 36th amendment of its constitution on April 26, 1975. The President of India accorded his consent to the Sikkim bill on May 15, 1975, and Sikkim became the 22nd state of the Indian union. Elections to the legislative assembly were held in 1979. (S.S.Ch.)

Accession to India

Tripura

Tripura, a small state in the northeastern arm of India, is bordered on the north, west, and south by Bangladesh, on the east by the Indian state of Mizorām, and on the northeast by the state of Assam. Its area of 4,051 square miles (10,491 square kilometres) is about equal to that of Lebanon but is half the size of nearby Manipur state, India. The capital city of Agartala is the only urban area in the state. Tripura is basically an agricultural area in which rice is the major crop. Since the 1960s the Indian government has fostered the growth of small-scale industries in the state.

PHYSICAL AND HUMAN GEOGRAPHY

The state can be divided physiographically into the four northern valleys—Dharmanagar, Kailāshahar, Kamalpur, and Khowai—and the southern plain of Agartala. The easternmost valley of Dharmanagar is bordered on the east by the long ranges of the Jāmrai Tlāng, which rise to between 2,000 and 3,000 feet (600 and 900 metres) above sea level, and on the west by the Sākhān Tlāng, which rise to between 1,000 and 2,500 feet. It is drained by the northward-flowing Deo River in the south and the Juri River in the north.

The northern valleys

West of the Sākhān Tlāng is the Kailāshahar Valley, which, like the Dharmanagar Valley, runs the full length of the state. Its western limit is defined by the Langtarai Range, which rises to more than 1,500 feet above sea level. The northward-flowing Manu River and its tributaries drain the valley, which is much dissected and forested in its southern half. The northern half of the valley is more open and even marshy.

Between the Langtarai Range on the east and the Atharamura Range, which rises to more than 1,400 feet on the west, lies the Kamalpur Valley. The valley is drained by the northward-flowing Dhalai River; the south is covered with forests, and the north is largely marshy rice fields. West of the Atharamura Range lies the Khowai Valley, which is bordered on the west by the low Deotamura Range. The valley is drained by the northward-flowing Khowai River; its western hills do not rise much above 800 feet in elevation.

The southern half of Tripura is more open and contains an extensive plain to the west of the Deotamura Range. The region is drained by a number of rivers of which the Gumti is the largest. The Gumti receives several southward-flowing streams and cuts across the ranges in a steep-sided valley from east to west before emerging from the hills near Rādhākishorepur.

Climate

More than 80 inches (2,000 millimetres) of rainfall occur during the monsoon season from June to September. Temperatures are hot in the valleys and cooler in the mountains.

Natural vegetation consists largely of evergreen forests, much of which have been cleared for cultivation. The remaining forests contain valuable trees including the sal, an East Indian timber tree that yields hardwood, second in value only to teak. Animal life includes tigers, leopards, jackals, and wild dogs. The Asiatic elephant is fairly common, and there are wild boar, the gaur—the largest of the world's wild oxen—the gayal ox, and wild buffalo.

The people. Tripura is predominantly a rural state, with the most densely populated areas being the Dharmanagar and Khowai valleys.

Of the total population, more than 40 percent belong to scheduled castes and scheduled tribes, groups outside the traditional Indian caste system. More than half the population speaks Bengali; other important languages are Tripurī and Manipurī. Most of the population is Hindu. There are also many Muslims and Buddhists.

Towns are concentrated on the southern plain, where the state capital of Agartala, Sonāmura, Udaipur, Amarpur, Belonia, and Sabrūm are located. The four towns for which the northern valleys are named all serve as local marketing centres.

The economy. Tripura is primarily an agricultural area. The major crop is rice, which is chiefly grown in the northern marshy section of the Kailāshahar Valley. Other cash crops include jute (an East Indian plant that produces fibres used in the manufacture of sacking, burlap, and twine), cotton, tea, and fruit. Sugarcane, rape (a European herb of the mustard family that produces oilseeds), mustard, and potatoes are among the minor crops. Livestock includes cattle, buffalo, sheep and goats, horses, pigs, and poultry. Forestry activities produce timber, firewood, and charcoal. Manufacturing is largely on a small scale and includes many cottage industries, such as weaving, carpentry, basketry, and pottery manufacture. There are small units for the production of tea, sugar, sawn timber, bricks, tin articles, canned fruit, agricultural implements, and footwear; larger establishments include a spinning mill, a jute mill, a steel mill, and a plywood factory.

Financial services include one state cooperative bank and one land mortgage bank. There are also numerous service, agricultural, and marketing societies.

Power plants

Energy is provided by diesel-powered thermal power stations at Agartala, Ambāsā, Khowai, Dharmanagar, Kailāshahar, Udaipur, and Bogafa. The Gumti Hydroelectric Project, completed in 1976, is the primary source of hydroelectricity.

Tripura's topography has rendered communications difficult, and transport outside of the state is made difficult by the surrounding foreign territory of Bangladesh. Two main roads link Agartala with Karīmganj (Assam) and with Sabrūm. There is also a railway link from Dharmanagar to Kalkalighāt, Assam. Most of the rivers carry boat traffic, especially during the rainy season. Long-distance travel is most convenient by air, and a small airport provides regional air service from Agartala.

Administrative and social conditions. Tripura is divided into three administrative districts: North Tripura, South Tripura, and West Tripura. The governor, appointed by the president of India, is the constitutional head of the state. The actual administration is conducted by a council of ministers, headed by a chief minister responsible to the elected unicameral legislative assembly. The jurisdiction of the Gauhāti high court extends over Tripura.

Education and health services

Higher educational facilities include an agricultural school, some industrial training centres, one training centre for nursing, and several professional and technical schools. The state's colleges offer general studies and curricula in music and engineering.

Health services include hospitals, public health centres, and dispensaries. There are specialized clinics for the treatment of leprosy, venereal diseases, and diseases of the eye, chest, and teeth. There are also a large number of family-planning centres.

Cultural life. The major factor in Tripura's cultural life is its isolation from the nations around it. The members of the secluded tribes live in the hills and are divided into numerous groups speaking a variety of languages and dialects. Most of the population, adhering to Hinduism and speaking Bengali, seem to have cultural links to the rest of India, while the smaller Muslim population seems closer in culture to Bangladesh. (S.P.C.)

HISTORY

The history of Tripura includes two distinct periods—the traditional, largely legendary period described in the *Rājamālā,* a chronicle of the supposed early kings of Tripura, and the period since *c.* 1431–62, the reigning years of the great king Dharma Māṇikya. The *Rājamālā* is a history in Bengali verse, compiled by the Brāhmaṇas in this king's court. In his reign and that of his successor, Dhanya Māṇikya. (reigned *c.* 1463–1515), Tripura suzerainty was extended over much of Bengal, Assam, and Burma in a series of remarkable military conquests. It was not until the beginning of the 17th century that the Mughal Empire extended its sovereignty over much of Tripura, with the local kings, or maharajas, still in power.

When the British East India Company obtained the *dīwānī,* or financial administration, of Bengal in 1765, that part of Tripura that had been under Mughal rule came under British control. From 1808, each successive ruler had to receive investiture from the British government. In 1905 Tripura was attached to the new province of Eastern Bengal and Assam and was known as Hill Tippera.

The last ruling maharaja of Tripura, Bīr Bikram Kishore Māṇikya, ascended the throne in 1923 and, before his death in 1947, settled that his state should accede to the Indian Union. It became a part of the Union on October 15, 1949, was made a union territory on September 1, 1956, and became a state on January 21, 1972. (Ed.)

West Bengal

West Bengal, one of the constituent states of the Republic of India, is situated in the northeast of the country. It is bounded in the north by Bhutan and the state of Sikkim, on the east by Bangladesh and the state of Assam, on the south by the Bay of Bengal, on the southwest by the state of Orissa, and on the west by Nepal and the state of Bihār. Its area is 33,920 square miles (87,853 square kilometres). Although in area West Bengal ranks as one of the smaller states of India, it is one of the largest in population. The state capital is Calcutta, India's most populous city.

West Bengal has a peculiar configuration; its breadth varies from 200 miles at one point to hardly 10 miles at another. It is of particular strategic importance to India's defense. The 1,350-mile-long frontier with Bangladesh, neither natural nor well defined, is a perennial source of tension. In addition, West Bengal is separated from the People's Republic of China only by the buffer states of Bhutan and Nepal.

PHYSICAL AND HUMAN GEOGRAPHY

The land. West Bengal may be broadly divided into two natural geographical divisions—the Gangetic Plain and the Himalayan and sub-Himalayan area. The upper limit of the Gangetic Plain is the northern border of the district of West Dinājpur, which is a watershed. The Ganges River, much of which lies now in Bangladesh, has for centuries been moving steadily eastward; very little of its water now goes to the sea via the western distributaries, of which the principal one is the Hooghly. This part of the Gangetic Plain has numerous marshes and shallow lakes formed out of dead river courses. The entire area has deep alluvial soil suitable for agriculture. The elevation of the land increases slowly toward the west; the rise is most marked near the Chota Nāgpur Plateau of Bihār, where extensive coal deposits are found.

The sub-Himalayan tract, known as Duārs or Tarai and comprising the districts of Jalpaiguri and Cooch Behār, consists of low land. Once unhealthy and infested with malaria and kala-azar, or black disease (leishmaniasis), the area is now well-drained, healthy, and cultivated. Some of the finest tea plantations of India are situated there.

The Darjeeling district

From the Duārs, the Himalayan mountain ranges rise abruptly along the northern boundary of the state. Here the Darjeeling district is located near Sikkim. Mount Kānchenjunga, with adjacent high peaks in Sikkim, dominates the Darjeeling landscape, and Mt. Everest can be seen in the distance on a clear day.

The climate in the plains is tropical, with high humidity and a moderately high temperature. The annual rainfall in the sub-Himalayan region varies between 150 and 210 inches (3,800 and 5,300 millimetres), and in the plains between 45 and 75 inches.

The year may broadly be divided into three marked seasons—the hot and dry season (March to early June), with dry sultry days and frequent thunderstorms; the hot and wet season (mid-June to September), when rain-bearing monsoon winds blow from the southwest; and the cold (cool) season (October to February), when days are dry and clear and stable atmospheric conditions prevail.

Forests occupy about one-eighth of the total land, and the region as a whole has a rich and varied plant life. In the sub-Himalayan plains the principal forest trees include *sal* (a timber tree), *sisu* (which yields a valuable wood), and *sisum* (Indian rosewood); the forest are interspersed with reeds and tall grasses. The landscape is mostly rice plain, broken by patches of jungle and shrubs. *Sal* is found on sandy soil while other trees grow in the marshes, where cane jungle is also common. Mango, jackfruit (the large pulpy fruit of the jackfruit tree), and bananas are the main fruit trees. On the Himalayan heights vegetation varies according to the altitude, with coniferous belts occurring at higher levels. The delta of the Hooghly is thick with malarial mangrove forests called Sundarbans.

The forest of the northern districts are inhabited by tigers, panthers, elephants, bison, and rhinoceroses, as well as by other animals of the Indian plain, large and small. Reptiles and birds are of the same species as are common throughout the Indian subcontinent.

The people. Three quarters of the population live in villages; and, of the remaining quarter living in urban areas, about two-thirds live in greater Calcutta.

Of the different religions, Hinduism, with its substrata of castes and aboriginal tribes, claims the adherence of the great majority of the population, while about a fifth of the population is Muslim. The tiny remainder consists of Christians, Buddhists, Jains, and Sikhs.

There are many recognized communities of tribes in West Bengal; of these the better known are the Santāls, the Oraons, and the Muṇḍās in the plains and the Lepchās and Bhutias in the Himalayan region.

Linguistic groups

Bengali is the main language of the state and is spoken by most of the population. Other languages spoken are Hindī, Santālī (a tribal dialect), Urdū, and Nepālī. Small minorities speak Oraon (a tribal dialect) and English. English, together with Bengali, is the language of administration, and English serves as a lingua franca for business purposes.

There are about 39,000 villages in the state. Most of the rural people live in small villages. Four zones of conurbation have developed in association with industrial complexes. The conurbation of Calcutta, including numerous towns on either side of the Hooghly, is the largest of these. The second cluster comprises the coal-mining and industrial zone of Asansol-Durgāpur. The third is in the Murshidābād District on either side of the Bhāgīrathi River. The fourth is in the northern extreme of the Midnapore District, which includes Haldia.

The economy. Small in size, West Bengal nevertheless accounts for about one-sixth of India's net domestic product. The partition of Bengal had left West Bengal in a state of economic chaos. Of the two main industries, the jute mills in West Bengal depended on raw jute grown in East Pakistan (now Bangladesh), and the transport of tea from the plantations in the north was affected by the

The jute and tea industries

closure of the rail and river routes through East Pakistan. To overcome these setbacks and to develop new industries became the principal economic objectives of the state.

About a quarter of the crop-producing area is irrigated. Rice covers four-fifths of the area under cultivation on the plains. Jute, pulses (leguminous plants, such as peas, beans, and lentils), oilseeds, corn (maize), wheat, barley, sugarcane, and vegetables are planted on the remainder of the land. A number of tropical fruits are grown, notably mango in the plains, and temperate fruits in the Himalayan range, but orchards are still not economically exploited. The tea plantation industry accounts for about one-quarter of India's total tea production.

The state is among India's top producers of minerals. There are extensive coal deposits near the Bihār border; coal amounts for more than three-fourths of the state's total mineral production, and about a quarter of the country's coal production. Other minerals include china clay, fire clay, dolomite, limestone, manganese, tungsten (wolfram), and quartz.

The importance of the port of Calcutta and the proximity of coal and steel originally gave Calcutta a lead in establishing a heavy engineering industry; West Bengal has concentrated on such heavy-engineering industries as the structural steel industry and railway wagon building. Two major steel plants of India—a privately owned plant at Burnpur and a publicly owned plant at Durgāpur—are located in the state. Other important industries are oil refining, automobile manufacturing, shipbuilding, chemical and fertilizer production, and machine building.

In 1947 the state was faced with the necessity of restoring rail and river-transport services to North Bengal and Assam after these had been disrupted by the closure of the old routes through East Pakistan. By the late 20th century the challenge had been successfully met.

Of necessity, frequent air services connect West Bengal with North Bengal and Assam. Calcutta possesses an international airport equipped to accommodate jet aircraft of the major international airlines. Calcutta Harbour handles much of India's commerce. After the loss of the East Pakistan rivers, the scope for river transportation of goods has become limited.

Administrative and social conditions. *Government.* Executive power in the state is vested in the governor and is exercised by him either directly or through subordinate officers. The governor is appointed by the president of India. The constitution provides for an elected council of ministers, with a chief minister at its head, to aid and advise the governor. The chief minister is appointed by the governor, and the other ministers are appointed by the governor on the advice of the chief minister. The council of ministers is collectively responsible to the legislature. The state legislature consists of a single house, the legislative assembly; the legislative council was abolished in 1969.

The constitution provides for a High Court, to which the chief justice and other judges are appointed by the president of India. District and sessions judges are appointed by the governor.

The state is divided administratively into 16 districts: Bānkura, Bīrbhūm, Burdwān, Calcutta, Cooch Behār, Darjeeling, Hooghly, Howrah, Jalpaiguri, Mālda, Midnapore, Murshidābād, Nadia, Purūlia, Twenty-four Parganas, and West Dinājpur. There is a collector, who is also the district magistrate, to administer each district except Calcutta. Each district is further divided into subdivisions, within which a subdivisional officer is the principal official. In each subdivision there are a number of police stations; each of the police stations has several *mawza's* (villages) under its administrative control.

With the object of developing rural self-government, *mawza's* have been grouped together and a *pañcātya*—an elected local authority established under the West Bengal Panchayat Act of 1956—entrusted with sanitary and conservation services for the group, and with the supervision of the village police and the development of cottage industries. A three-tiered *pañcātya* system covers the rural area with some 3,300 *grām* (village) *pañcāyats,* about 300 intermediate level *pañcāyats,* and 15 district level *pañcātyas.*

Education. There are primary schools, secondary schools, engineering and medical colleges, and eight universities that offer courses in the arts, sciences, engineering and technology, and agriculture. Several hundred colleges offering baccalaureate degrees are affiliated with the universities. There are also many technical institutes. More than 5,000 adult education centres exist for training in literacy. There is also a central library, together with a number of district libraries, area libraries, and rural libraries.

Health. Medical facilities include hospitals, clinics, health centres, and dispensaries. Family-planning services are available in district bureaus, as well as in urban and rural centres. An employees' state insurance scheme provides factory workers with health, employment, safety, and maternity insurance and also provides a free medical service.

Welfare. A social welfare directorate coordinates various welfare services dealing with orphans, destitute persons, vagrants, the mentally and physically handicapped, and child offenders. The government's social-welfare enterprises are supplemented by private agencies, of which the most prominent is the Ramakrishna Mission, founded by the Hindu reformer and teacher Swami Vivekananda in 1897, and the Order of the Missionaries of Charity (1948), founded by Mother Teresa, recipient of the 1979 Nobel Peace Prize. There are also a large number of hospitals and schools operated by various charity organizations.

Cultural life. Bengalis have always fostered literature, art, music, and drama. Bengali literature, which underwent a vigorous creative synthesis after its first contact with the West in the early 19th century, ranks foremost amongst the literatures of India and has produced, among others, the Nobel prize-winning poet Rabindranath Tagore (1861–1941), whose contribution still dominates the Indian literary scene.

The theatre is popular, and performances—amateur, as well as professional—are sophisticated. *Yātrās,* traditional open-air performances which are now changing from predominantly mythological and historical topics to modern themes, are popular in the countryside, as also is *kavi,* an impromptu duel in musical verse between village poets. The *kathakata,* a religious recital, is another traditional form of rural entertainment, based on folklore.

The film industry

The film industry is another established form of popular entertainment. Bengali films have earned national and international awards for their delicate handling of Indian themes; the works of the directors Satyajit Ray, Tapan Sinha, and Mrinal Sen are particularly notable.

Traditional music takes the form of devotional and cultural songs. *Rabindra Sangeet,* songs written and composed by Tagore, which draw on the pure Indian classical as well as traditional folk-music sources, exert a powerful influence in Bengali cultural life.

The visual arts have, by tradition, been based mainly on clay modelling, terra-cotta work, and decorative painting. In Calcutta, the Indian Association for the Cultivation of Science, the Bose Research Institute, and the science laboratories of the University of Calcutta have made notable contributions to science. The Asiatic Society of Bengal, the best known Indian historical-research body in the 19th century, is located in West Bengal. Viśva-Bhārati University, founded by Rabindranath Tagore in Śantiniketan, is a world-famous centre for the study of Indology and international cultural relations. (D.M.S./Ed.)

HISTORY

The name of Bengal, or Bānglā, is derived from the ancient kingdom of Vanga, or Banga. References to it occur in early Sanskrit literature, but its early history is obscure until the 3rd century BC, when it formed part of the extensive Maurya Empire inherited by Aśoka. With the decline of Maurya power, anarchy once more supervened. In the 4th century AD it was absorbed into the Gupta Empire of Samudragupta. Later it came under Pāla rule. From the beginning of the 13th century to Robert Clive's conquest of the province in 1757, Bengal was under Muslim rule, at times under governors acknowledging the suzerainty of the Delhi sultans but chiefly under independent rulers.

In 1756 Shāh 'Alam, the Mughal emperor defeated by the British, granted to the East Indian Company the *dīwānī* of

Bengal, Bihar, and Orissa—that is, the right to collect and administer the revenues of the area. By the Regulating Act of 1773, Warren Hastings, the governor of Bengal, became the first governor general of Bengal, which was declared to be the supreme government with powers of superintendence over the other two presidencies of Madras and Bombay as well.

In 1854 the governor general was relieved of the direct administration of Bengal, which was placed under a lieutenant governor. Thenceforward, the government of India became entirely distinct from that of Bengal. In 1874 Assam was transferred from the charge of the lieutenant governor and placed under a separate chief commissioner. It was felt in 1905 that Bengal had become too unwieldy a charge for a single administration, and in spite of violent Hindu protests, it was partitioned into two provinces, (1) Western Bengal, Bihār, and Orissa and (2) Eastern Bengal and Assam, each under a lieutenant governor. In 1912, because of continued opposition to partition, Bengal was placed under a governor, Bihār and Orissa under a lieutenant governor, and Assam once more under a chief commissioner. At the same time, Delhi became the capital of India in place of Calcutta. Under the Government of India Act (1935), Bengal was constituted an autonomous province in 1937. This was the situation until the Indian subcontinent was partitioned into the two dominions of Pakistan and India after the British withdrawal in 1947. The eastern sector of Bengal, largely Muslim, became East Pakistan (later Bangladesh); the western sector became India's West Bengal.

The partition of Bengal left West Bengal with ill-defined boundaries and a constant inflow of non-Muslim, mostly Hindu, refugees from East Pakistan (now Bangladesh). More than 7,000,000 refugees entered the already densely populated state after 1947, and their rehabilitation placed an immense burden on the administration. The refugees themselves fell easy prey to exploitation by political parties in opposition.

In 1950 the princely state of Cooch Behār was integrated with West Bengal, and in 1954 the former French enclave of Chandernagore was transferred to the state. After the reorganization of Indian states linguistically in 1956, West Bengal gained 3,157 square miles from Bihār; the new land provided a link between the previously separated northern and southern parts of the state.

(Ed.)

WEST CENTRAL INDIA

Gujarāt

The state of Gujarāt lies on the west coast of India. It is bounded by the Arabian Sea to the west, by Pakistan to the northwest, the Indian states of Rājasthān to the north, Madhya Pradesh to the east, and Mahārāshtra to the south. Its coastline is 992 miles long, and no part of the state is more than 100 miles from the sea. Its area is 75,685 square miles (196,024 square kilometres).

The capital of the state is Gandhinagar on the northern outskirts of Ahmadābād (Ahmedabad), the former capital, largest city in the state, and one of the greatest cotton-textile centres in India. It was here that Mahatma Gandhi built his Sābarmati ashram (*āśrama*; "retreat," or "hermitage") as a headquarters for his campaigns.

Gujarāt derived its name from the Gurjara (supposedly a subtribe of the Huns), who ruled the area during the 8th and 9th centuries AD. The state in its present form came into existence in 1960 when the former Bombay state was divided between Mahārāshtra and Gujarāt on a language basis.

Though most of the people are engaged in agriculture, there is a cohesive and comparatively prosperous merchant community that thrives on the trade and commerce generated in Gujarāt because of its long coastline.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The present-day state of Gujarāt is one of great contrasts; it stretches from the wet, fertile, rice-growing plains of the west coast, north of Bombay city, to the almost rainless salt deserts of Kutch. It is easiest to treat the state on the basis of its natural divisions. In the northwest, Kutch comprises a single district so arid as to be almost desert; it is bounded on the south by the Gulf of Kutch and on the north and east is separated from Pakistan and the mainland of India by the Rann of Kutch, best described as a vast salt marsh covering about 8,000 square miles. The Rann floods during the rainy season, slight though the rains may be, and Kutch is converted into an island; in the dry season it is a sandy, salty plain, plagued by duststorms. To the south of Kutch is the large peninsula of Saurāshtra (Kāthiāwār), lying between the Gulf of Kutch and the Gulf of Cambay. It also is arid and rises from the coasts to a low, rolling area of hill land in the centre, covered with scrub or sparse woodland. The chief towns are found in the more fertile spots and were formerly the capitals of small states. Soils are mostly poor, having been derived from a variety of old crystalline rocks, but among the state's valuable products are the fine building stones of Porbandar. Rivers, except for seasonal streams, are absent. On the southern shores of the peninsula is the former Portuguese territory of Diu. Northeastern Gujarāt is mainly a country of small plains and low hills, through which runs the main line of a railway from Bombay to Delhi. Rainfall is low; January temperatures may drop almost to freezing point, while a temperature of 118° F (48° C) has been recorded in the hot season. Crops grown are millet and some cotton.

The Rann of Kutch

Southward in central Gujarāt the rainfall increases; temperature ranges are less; and soils are more fertile, being derived partly from the basalts of the Deccan region. The focus of this area is the city of Baroda (officially Vadodara) formerly the capital of a rich and powerful state. South of what is now the Vadodara district, the important river, the Narmada, empties into the Gulf of Cambay, and it is the silt deposited by this river and the Tāpti that is responsible for the shallowness of the Gulf of Cambay and the decline of its former ports.

Southern Gujarāt, the districts of Bharūch and Surat, are famed for their rich soils and fine crops of cotton. The great Tāpti River, flowing in a deep trench from the east, cuts through Surat; and in the eastern parts of south Gujarāt the country is mountainous. This is, indeed, the northern extension of the Western Ghāts, which attract a heavy rainfall from the rain-bearing summer monsoon winds. Farther south, the mountains are forested. The small district of the Dangs is in this area. Along the coastal plains conditions begin to approach an equable climate, with rainfall nearing 80 inches.

Rainfall and altitude variations are such that the vegetation changes from north to south and from the east highlands to the west coast. With less than 25 inches of rainfall, the species in the scrub forests are the babul acacia, the caper, the Indian jujube, and the toothbrush tree (*Salvadora persica*). Where rainfall is up to 40 inches the following deciduous species are found: teak, catechu or cutch, bakligum, axlewood, and Bengal kino or butea gum. These are found in the tablelands of Saurāshtra and the east mainland area. Most of the deciduous forests are concentrated in areas with more than 40 inches of rain in the south and east. They produce valuable timbers: wooly tomentosa, Vengai padauk (resembling mahogany), Malabār simal, and the heartleaf adina. The west coast of Saurāshtra is known for its algae, and the east coast produces the paper-making plant *Cyperus papyrus* (a sedge known as the paper reed or paper rush).

The Gir Forest National Park in Saurāshtra contains the last Indian lions, the only remaining numbers of the Asiatic species; in the Little Rann of Kutch the only surviving Indian wild asses are found.

The peacock is common throughout the state, and there is a bird sanctuary at Lake Nal Sarovar, near Ahmadābād,

which attracts about 140 species of birds migrating from the Siberian plains and elsewhere in winter. These include the sarus crane, the Brahmini duck, the bustard (a game bird), the pelican, the cormorant, the ibis (a wading bird related to the heron), the stork, the heron, and the egret. The Rann of Kutch is the only nesting ground of the large flamingo (*Phoenicopterus ruber*) in India.

There is excellent offshore and inland fishing in Gujarāt. Catches include pomfret (a food fish), salmon, hilsa (a type of shad), jewfish (scianid fish), prawn, Bombay duck, and tuna.

The people. Ethnically, there are two dominant racial strains in Gujarāt: a brachycephalic (short-headed) leptorhinian (straight-nosed) element, represented by the Pārsī and the Nāgar Brahmin, Bhāṭiā, Bhādela, Rabāri, and Mīnā castes; and a dolichocephalic (long-headed) mesorrhinian (broad-nosed) element, represented by the aboriginal Bhīl and the Bhaṅgī, Koḷī, Dubla, Naikda, and Mācchi-Khārwā tribes. The rest of the population are in intermediate groups; Gujarāt as a whole is a mesocephalous region (*i.e.*, one in which heads are approximately midway between brachycephalic and dolichocephalic). Members of the former untouchable castes and of the aboriginal tribes form nearly one-fifth of the state's population. There is one entirely tribal district of Dangs. Ahmadābād district has the highest proportion of scheduled (untouchable) castes.

The Gujarātī language

The Gujarātī language belongs to the Indo-Aryan family of languages and is derived from Sanskrit through Prākrit, ancient Indic languages other than Sanskrit, and Apabhraṃśa, a language spoken in northern and western India from the 10th to the 14th century. Gujarāt's contact by sea with foreign countries has also led to the introduction of Persian, Arabic, Turkish, Portuguese, and English words. The prodigious writings in Gujarātī of Mahatma Gandhi, the architect of India's freedom, are noted for their vigour and simplicity and have had a strong influence on modern Gujarātī prose.

Hinduism is the religion of most of the population, with a minority of adherents to Islām and Jainism. The policy of the state has always been marked by the religious tolerance of its people.

About one-third of the population is urban. The most highly urbanized part of the state is the Ahmadābād–Vadodara (Baroda) industrial belt. The major industrial towns are Ahmadābād, Vadodara, and Surat. Major towns that were once capitals of princely states are Rājkot, Junāgadh, Bhavnagar (Bhaunagar), and Jāmnagar.

The role of agriculture

The economy. About two-thirds of the population is engaged in agriculture, the gross area cropped amounting to about half of the total land area. Wheat and millets are the staple food crops, with rice production being concentrated in the wetter areas. Sugarcane production is increasing, while cotton, tobacco, and oil seeds (especially peanuts [groundnuts]) are profitable cash crops. Gujarāt produces about one-third of India's cotton and peanut crop and about one-sixth of the country's tobacco. Cash crops are, indeed, a characteristic feature of the state's agricultural economy, slightly exceeding food crops in value.

Gujarāt occupies a high place in the industrial economy of India. The state is rich in minerals such as limestone, manganese, gypsum, calcite, and bauxite; there are also deposits of lignite, quartz sand, agate, and feldspar. Gujarāt is India's major petroleum producing state along with Assam. Output of soda ash and salt amounts to most of the national production; the cement, vegetable oil, chemical, and cotton textile industries are highly important. With its many textile mills, Ahmadābād, has been called "the Manchester of India." The pharmaceutical industry, concentrated at Vadodara, Ahmadābād, and Atul (Valsād), manufactures a good percentage of the national product. The oil refinery at Koyali has created a nearby fast-growing petrochemical industry. Cooperative commercial dairying also is important in Gujarāt. The steady growth of small industries has been significant. The Gandhian approach to labour problems—strict reliance on the truth, nonviolence, settlement by arbitration, minimal demands, and the use of the strike only as a last resort—has had a great impact in the field of industrial relations in Gujarāt, which has remained free from labour unrest.

A thermal-power station using coal, natural gas, and oil is located at Dhuvaran. The state also receives power from the Tārāpur nuclear station in Mahārāshtra state.

There is a good system of railroads and public roads. Transport by land is supplemented by water transport, particularly by coastal shipping. There are many ports besides the major international port of Kandla. Several airports provide air transportation within the state and with the major cities of India outside Gujarāt.

Administrative and social conditions. *Government.* A governor, appointed by the president of India, is the head of the government of Gujarāt. A council of ministers, led by the chief minister, aids and advises the governor in the exercise of his functions. The unicameral legislative assembly is an elected body.

The High Court is the highest judicial authority in the state. There are also city courts, the courts of district and sessions judges, and courts of civil judges in each district.

The state is divided into 19 administrative districts; the revenue and general administration of each district is looked after by the district collector, who also functions as the district magistrate for the maintenance of law and order. The 19 districts are: Ahmadābād; Vadodara; Banās Kāntha; Bharūch; Mehsāna; Kheda; Pānch Mahāls; the Dangs; Surat; Sābar Kāntha; Jāmnagar; Bhavnagar; Amreli; Junāgadh; Kutch; Rājkot; Surendranagar; Valsād; and Gandhinagar.

The *pañcāyats*

With a view to associating the people living in the remotest villages with the administration, rule by *pañcāyat* (local council) was introduced in 1963. The scheme comprised (1) a *grām* (village) *pañcāyat* for each village or group of villages or a *nāgar* (town) *pañcāyat* for towns with a population of between 10,000 and 20,000; (2) a *taluka* (an administrative unit similar to a county) *pañcāyat* for each *taluka* (called a *tahsil* in some states); and (3) a district *pañcāyat* for each district.

Health. Health and medical services are still being extended to remote and backward areas. There are programs for the eradication of malaria, tuberculosis, leprosy, and other communicable diseases as well as for improving supplies of drinking water and preventing food adulteration. Steps have also been taken to expand primary health centres, hospitals, and medical colleges.

Education. Primary schools for all children between seven and 11 have been opened in nearly all villages with 500 or more inhabitants. Besides these, special schools are teaching arts and crafts. Special schools are run for tribal children. There are many secondary schools, as well as eight universities and a large number of institutions for higher education. Technical education is provided by engineering colleges, junior technical schools, and technical schools. Research institutions include the Physical Research Laboratory, the Ahmadābād Textile Industry's Research Association, the Bholabhai Jesinghbhai Institute of Learning and Research, the Indian Institute of Management, the National Institute of Design, and the Sardar Patel Institute of Economic and Social Research at Ahmadābād, the Oriental Institute at Vadodara, and the Central Salt and Marine Chemicals Research Institute at Bhavraga. There are also institutes doing research in agriculture and fisheries.

Welfare. At the same time, there are state institutions under the director of social welfare for the welfare of children, women, the physically handicapped, the aged, infirm, and destitute, for delinquents, beggars, orphans, and released prisoners. There is also a Backward Class Welfare Department for the education, economic uplift, health, and housing of the backward class.

Cultural life. The folklore and folk culture of Gujarāt can be traced to the mythology of Lord Krishna (Kṛṣṇa; an incarnation of the god Vishnu, or Viṣṇu), as told in the *Purāṇas*, a class of Hindu sacred literature. The *rāsnṛtya* and *rāslīlā* dances in honour of Lord Krishna have survived in the form of the popular folk dance, the *garabā.* This dance is performed at the *Navarātrī* festival, which honours the goddess Durgā, the dancers moving in a circle and singing and keeping time by clapping their hands or clashing together sticks called *daṇḍa.* A folk drama, the *bhavai,* has also survived.

Śaivism (Shivaism), the cult of the god Śiva (Shiva), characterized by ritual devotion, has long flourished in Gujarāt; so, too, has Vaiṣṇavism (Vishnuism, the worship of the god Vishnu), the devotional practices of which, with the emergence of the cult of *bhakti* ("devotion"), inspired its devotees to compose songs to be sung in temples or other religious gatherings. Vaiṣṇavism has encouraged the emergence of a number of saints, poets, and musicians, such as Narsiṃha, who composed *padas* (verses) in the 15th century; Mīrā Bāī, a 16th-century Rājput princess who renounced her royal home and composed *bhajanas* ("devotional songs"); Premanand, an 18th-century poet and writer; and Dayārām, an 18th-century composer of songs who popularized the *bhakti* cult. Jainism, with its nonviolence and vegetarianism, gained a stronger hold in Gujarāt than in any other part of India.

The architectural style of Gujarāt is well known for its luxuriousness and perfection. The style is preserved in such monuments and temples as those at Somnāth, Dwārkā, Modhera, Thān, Ghumli, Girnār, and Pālitāna. A happy synthesis of Hindu-Muslim culture has also been preserved in many ancient monuments. Gujarāt is famous, too, for some of its art and craft products. These include the *jari* (gold and silver embroidery) of Surat; the *bāndhaṇī* work (a tie-and-dye technique) of Jāmnagar; the *paṭolā,* a fine silk fabric of Pātan; the toys of Idar; the perfumes of Pālanpur; the hand-loomed products of Kondodar; and the decorative woodwork of miniature temples and mythological figures at Ahmadābād and Surat.

The *mahājans*

Among the most durable and effective of the state's cultural institutions are the trade and craft guilds known as the *mahājans*. Often coterminous with castes and largely autonomous, the guilds have in the past solved disputes, acted as channels of philanthropy, and encouraged arts and culture.

HISTORY

Human settlements have been traced back to the Stone Age period in the valleys of the Sābarmati and Mahi rivers in the eastern part of the state. The historic period is linked with the spread of the Harappan (Indus Valley) civilization, which flourished in the 3rd and 2nd millennia BC. Centres of this civilization have been found at Lothal, Rangpur, Amrī, Lakhabaval, and Rozdi (mostly in the Saurāshtra [Kāthiāwār] Peninsula).

The Mauryan period

The known history of Gujarāt began with the extension over the area of the rule of the Mauryan dynasty, as is evidenced by the edicts of the emperor Aśoka (*c.* 250 BC), carved on a rock in the Girnār Hills of the Saurāshtra Peninsula. After the fall of the Mauryan Empire, Gujarāt came under the rule of the Śakas (Scythians), or western Kṣatrapas (AD 130–390). The greatest of these, Mahākṣatrapa Rudradāman, established his sway over Mālwa, Saurāṣṭra (Saurāshtra), Kutch, and Rājasthān.

During the 4th and 5th centuries, Gujarāt formed a part of the Gupta Empire, until the Guptas were succeeded by the Maitraka dynasty of the kingdom of Valabhī, which ruled over Gujarāt and Mālwa for three centuries. The capital Valabhīpura (near the eastern coast of the Saurāshtra Peninsula) was a great centre of Buddhist, Vedic, and Jaina learning. The Maitraka dynasty was succeeded by the Gurjara-Pratihāras (the imperial Gurjaras of Kannauj), who ruled during the 8th and 9th centuries; they, in turn, were followed shortly afterward by the Solaṅki dynasty. The boundaries of Gujarāt reached their furthest limits in the time of the dynasty, when remarkable progress was made in the economic and cultural fields. Siddharāja Jayasiṃha and Kumārapāla were the best known Solaṅki kings; the famous writer Hemacandra flourished during this period (12th century). Karṇadeva Vāghelā, of the Vāghelā dynasty that followed, was defeated in about 1298 by ʿAlāʾ-ud-Dīn Khaljī, sultan of Delhi; Gujarāt then came under Muslim rule. It was Aḥmad I, the first independent sultan of Gujarāt, who founded Ahmadābād (1411). The end of the 16th century saw Gujarāt under Mughal rule; this lasted until the mid-18th century, when the Marāṭhās overran the state.

The founding of Ahmadābād

Gujarāt came under the administration of the British East India Company in 1818. After the Indian rebellion of 1857, the area became a province of the British crown and was divided into Gujarāt province, with an area of 10,000 square miles, and numerous native states. With Indian independence in 1947, all of Gujarāt except the states of Kutch and Saurāshtra was included in Bombay state; the province was enlarged in 1956 to include the two states. On May 1, 1960, Bombay state was split into present-day Gujarāt and Mahārāshtra states.

In April 1965 fighting broke out between India and Pakistan in the Rann of Kutch, an area that had long been in dispute between them. A ceasefire came into force on July 1, and the dispute was submitted to arbitration by an international tribunal. The tribunal's award published in 1968 gave nine-tenths of the territory to India and one-tenth to Pakistan.

(D.N.P./Ed.)

Madhya Pradesh

Madhya Pradesh, with an area of 171,220 square miles (443,459 square kilometres), is the largest of the states of the Republic of India. As its name implies—*madhya* means central and *pradesh* means region—it is situated in the heart of India. It has no coastline and no international frontier. The state capital is Bhopāl.

Madhya Pradesh is bordered by the states of Rājasthān to the northwest, Uttar Pradesh to the north, Bihār to the northeast, Orissa to the east, Andhra Pradesh to the southeast and south, Mahārāshtra to the south, and Gujarāt to the west.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Forested hills with steep slopes, extensive plateaus, and river valleys characterize the physiography of Madhya Pradesh. The state occupies the Deccan Plateau and in the west and north includes the Vindhya and Kaimur ranges, which in places rise to 1,500 feet (460 metres) above sea level; the Sātpura, Mahādeo, and Maikala ranges in the south have elevations of more than 3,000 feet.

Northwest of the Vindhya Range is the Mālwa Plateau (1,500 to 2,000 feet). Other features are the Bundelkhand Pathār, north of the Vindhyas; the Madhya Bhārat Pathār, in the extreme northwest; the Baghelkhand Plateau, in the northeast; the Chhattīsgaṛh Plain, in the east; and the Danda Kāranya region, in the extreme southeast.

The rivers

Madhya Pradesh is the source region of some of the most important peninsular rivers: the Narmada, the Tāpti, the Mahānadi, and the Wainganga (a tributary of the Godāvari). The Chambal forms the state's northern border with Rājasthān and Uttar Pradesh. Other rivers include tributaries of the Yamuna and the Son (itself a tributary of the Ganges).

Fertile black soils are found in the Mālwa Plateau, in the Narmada Valley, and in parts of the Sātpuras. Eastern Madhya Pradesh, except for the Chhattīsgarh Plain, has mainly less fertile and sandy red and yellow soils.

The climate is monsoonal, with much of the rain falling from June to October. The season preceding the rains (March to May) is hot and dry, with temperatures everywhere higher than 85° F (29° C). The temperature drops during the rainy season. Winters are usually pleasant and dry. Generally speaking, rainfall decreases westward and northward, being 60 inches (1,524 millimetres) and above in the east and southeast, but less than 40 inches in the extreme west and less than 30 inches in the Chambal Valley in the north.

About two-fifths of the state is under forest; another, much smaller percentage consists of permanent pasture or other grazing land. The main forested areas include the Vindhya-Kaimur ranges, the Sātpura and Maikala ranges, the Baghelkhand Plateau, and the Dandakāranya. Economically, the most important trees are teak, sal, bamboo, and salai (which yields a resin used for incense and as a medicine).

The forests abound in wild animals such as tigers, panthers, bison, chital (spotted deer), bears, wild buffalo, sambar (deer), and black bucks. Birds of many species are found. There are national parks at Kanha (famous for

its swamp deer) and at Shivpuri; another park has been established at Bandhogarh (Rewa) and is known for its white tigers.

The tribal groups

The people. A variety of tribes lives in Madhya Pradesh, including Bhīl, Baigā, Goṇḍ, Korku, Kol, Kamar, and Maṛiā. About 20 percent of the total population are classified as members of scheduled tribes, the highest percentages of tribal population among the states of India. The distribution of the rural population is very uneven, high densities being found in the Mahānadi Valley, the Upper Wainganga Valley, the Lower Chambal Valley, and the Narmada Valley, as well as scattered patches on the Mālwa Plateau in western Madhya Pradesh.

Languages

Hindī, the state language, is also the language most widely spoken. Eastern Hindī, represented by the Awadhī, Baghelī, and Chhattīsgaṛhī dialects, is spoken in Baghelkhand, Sātpura, and Chhattīsgaṛh and also in the upper Narmada Valley. Bundelī, a Western Hindī dialect, is spoken in the central and northwestern districts of Madhya Pradesh; Malvi, recognized by some as a Western Hindī dialect, is the speech of western Madhya Pradesh. The Bhīl speak Bhīlī, and the Goṇḍ speak Goṇḍī.

The second most important language is Marāṭhī. Urdū, Oṛiyā, Gujarātī, and Punjābī are each spoken by sizable numbers. Also spoken are Telugu, Bengali, Tamil, and Malayālam.

Most of the people are Hindus. There are, however, sizable minorities of Muslims, Jains, Christians, and Buddhists. There is also a small Sikh population.

The major urban agglomerations are Indore, Jabalpur, Gwalior, Bhopāl, Ujjain, Raipur, Durg-Bhilai, Sāgar, Bilāspur, Ratlām, and Burhānpur.

The economy. Agriculture is the basis of Madhya Pradesh's economy. The distribution of the sown area is very uneven, being governed by variations in topography, rainfall, and soils. The chief cultivated areas are found in the Chambal Valley, the Mālwa Plateau, the Rewa Plateau, and the Chhattīsgaṛh Plain. The Narmada Valley is another fertile area. Agriculture in Madhya Pradesh is characterized by low productivity and the use of traditional methods of cultivation. Only about one-tenth of the sown area is irrigated. Irrigation is carried out chiefly by means of canals, tanks (village lakes or ponds), and wells; the main development in this respect has been through medium-sized or small projects executed during the successive Indian five-year plans.

The most important crops are rice, wheat, *jowār* (sorghum), pulses (leguminous plants, such as peas, beans, or lentils), and peanuts (groundnuts). Rice is grown principally in the east, where the average annual rainfall is more than 50 inches; in western Madhya Pradesh wheat and *jowār* are more important. The state is the largest soybean producer in India. Other crops include linseed, sesame, sugarcane, cotton, corn (maize), and inferior millets which are grown in hilly areas.

Madhya Pradesh is rich in minerals, though these resources have not yet been fully exploited. There are large reserves of coal and important deposits of iron ore, manganese ore, bauxite, limestone, dolomite, copper, fireclay, and china clay. Diamond reserves at Panna are of particular interest.

The state is well endowed with hydroelectric-power potentials. A pioneer hydroelectric scheme is the Chambal Valley project, which has been jointly developed with Rājasthān. Other schemes include the hydroelectric projects of Narmada Sagar, Bargi Sardar, Bansagar, and Bango, as well as the Birsinghapur Thermal Power project.

Before planned development took place, western Madhya Pradesh was the main industrial area—primarily producing consumer goods. There are now many large- and medium-

Industrial centres

scale industries, important centres being Indore, Gwalior, Bhopāl, Ujjain, and Jabalpur. The principal heavy industries are government-sponsored and include the iron and steel plant at Bhilai Nagar, the newsprint mill at Nepa Nagar, the cement factory at Kaymore (Katui), and the heavy electrical factory at Bhopāl. As part of the planned development, industrial estates have been established at Indore, Gwalior, Bhopāl, Raipur, Bhilai, and Jabalpur. In the private sector, there are cement works, paper mills, sugar mills, and textile mills (cotton, wool, silk, and jute), besides flour, oil, and saw mills. There are also plants that manufacture fertilizer, synthetic fibres, and chemicals, in addition to some general engineering industries.

There are many registered small-scale industrial units in the state. The handloom industry is flourishing, and various traditional crafts are practiced such as the handweaving of saris in Chanderi, carpet weaving and pottery making in Gwalior, and gold and silver thread embroidering in Rhopāl.

In comparison with other Indian states, Madhya Pradesh is rather poorly served with transport and communications facilities. The main railroads that pass through the state were originally laid down to connect the ports of Madras, Bombay, and Calcutta with their hinterlands. Inadequate transport facilities have hindered the development of the state's rich resources.

There are both broad-gauge and narrow-gauge rail lines. The five main broad-gauge lines are Bombay–Delhi via Bhopāl, Bombay–Howrah (Calcutta) via Jabalpur, Bombay–Howrah (Calcutta) via Raipur, Madras–Delhi, and Bombay–Delhi via Ratlām.

The state is connected with Delhi and Bombay by air via Bhopāl Gwalior, and Indore.

The roads

The chief national highways connect Āgra with Bombay, Allāhābād with Nāgpur, Jhānsi with Lakhnādon, and Raipur with Vishākhapatnam. Many districts have fewer than ten miles of roads for every 100 square miles of territory. In the absence of a dense rail network, roads play an important part in the developing economy. The construction of bridges across the Narmada and other rivers has greatly helped the development of all-weather traffic routes.

Administrative and social conditions. Executive power in the state is vested in the governor, the constitutional head of state. The elected chief minister and the council of ministers aid and advise the governor in the exercise of his functions. The legislative assembly fluctuates in size from election to election.

Divisions and districts

The state has been divided administratively into 11 divisions and 45 districts, each headed by a collector. The collector exercises executive and magisterial power. There is a high court presided over by a chief justice.

Since the State Panchayat (village council) Act of 1962, local administration has been entrusted to village *pañcāyats*.

Health. Malaria, which was formerly endemic throughout the state, has been largely eradicated. Diseases common in Madhya Pradesh are filariasis (a parasitic disease of the blood), tuberculosis, venereal diseases, leprosy, and dysentery. Besides general hospitals and dispensaries there are several eye hospitals, mental hospitals, and sanatoriums and clinics for the treatment of tuberculosis. There are antirabies inoculation centres in every district, and several venereal-disease clinics.

The universities

Education. More than one-quarter of the state population is literate. There are schools for primary, middle, and high school education, and there are universities at Sāgar, Jabalpur, Raipur, Ujjain, Gwalior, Indore, Rewa, Bhopāl, and Khairāgarh (music); there are more than 100 colleges in the state. Jabalpur also has an agricultural university.

Welfare. The government has undertaken several social-welfare schemes, including adult literacy classes, *yuvak* ("youth") and *mahilā mandals* ("women's clubs"), and youth competitions. Grants-in-aid are given to social welfare and physical welfare institutions, while the Directorate of Panchayats and Social Welfare also runs rescue homes, beggars' homes, and leprosy clinics. Scholarships are granted to crippled children, and there is a scheme to give aid to old people needing care. The government also runs schools for the deaf and dumb.

Cultural life. There are a number of temples, fortresses, and cave works in Madhya Pradesh that have left fascinating evidence for historical studies, both on the early story of mankind in general and on local dynasties and kingdoms. One of the earliest monuments was the *stūpa* (Buddhist mound forming a memorial shrine) at Bhārhut (*c.* 175 BC), near Satna, the remains of which are now in the Indian Museum at Calcutta. Another such monument is the *stūpa* at Sānchi (six miles south of Vidisha),

originally constructed by Aśoka, emperor of India from about 265 to 238 BC; additions to the *stūpa* were made by the Śuṅga kings. The Bāgh caves near Mhow, with their wonderful paintings on Buddhist topics, merit a special mention; the Udayagiri caves (Brahmanical and Jaina monasteries) near Vidisha exhibit art and architecture in the rock-cut tradition. Among the temples known the world over for their erotic art are those at Khajurāho in the Chhatarpur district in the north of the state; dating from AD 1000, they were built by the Chandelā (Candellā) kings. The temples at Gwalior and in its vicinity should also be mentioned. The palaces at Māndu (near Dhār) and the Gwalior Fort, perhaps the most impressive of the residences of the former princes of Madhya Pradesh, represent other notable architectural achievements.

Many of the tribal traditions in Madhya Pradesh are still vital and strong, though they have been exposed in varying degrees to outside cultural influences. A great deal of tribal mythology and folklore is also preserved. The *pardhān* (bards of the Goṇḍ) still sing of the legendary deeds of Lingo-pen, the mythical originator of the Goṇḍ tribe. The *Pandwāni* is the Goṇḍ equivalent of the *Mahābhārata* (one of the two great Hindu epics), while the *Lachmanjati* legend is the Goṇḍ equivalent of the *Rāmāyaṇa*, the other great epic. Other literary works are also popular. All tribes have myths and legends regarding their origin. There are songs for the ceremonies of birth and marriage, while various dance styles have songs to accompany them. Folktales, riddles, and proverbs are other features of the cultural heritage.

Tribal mythology and folklore

HISTORY

The history of Madhya Pradesh extends to the remote past. Several remains of prehistoric cultures, including rock paintings and stone and metal implements, have been found in various parts of the state.

One of the earliest states that existed in the territory of present Madhya Pradesh was Avanti, of which Ujjain was the capital. This was part of the Mauryan Empire (4th–3rd century BC); it was later known as Mālwa and was situated in the western part of the state.

Among the various dynasties that ruled part or the whole of Madhya Pradesh between the 2nd century BC and the 16th century AD were the Śuṅgas (185 to 73 BC), who ruled in eastern Mālwa; the Andhras (Sātavāhanas; 1st century BC–3rd century AD); the Kṣatrapas (2nd–4th century AD); and the Nāgas (2nd–4th century AD). The whole of Madhya Pradesh lying north of the Narmada River formed part of the Gupta Empire (4th–5th century AD) and was the scene of a power struggle against the nomadic Hūṇas (Hephthalites, or White Huns). Yaśodharman was an important Mālwan king who wrested power from the Hūṇas in the 6th century. Mālwa was annexed by the emperor of northern India, Harṣa, during the first part of the 7th century.

The 10th century was a period of confusion: of the various dynasties, the Kalacuris ruled over the Narmada Valley and some other regions; their contemporaries were the Paramāras at Dhār in what is now the western region, the Kachwāhās at Gwalior in the north, and the Chandelās (Candellās) at Khajurāho, about 100 miles southeast of Jhānsi. Later the Tomaras ruled at Gwalior and the tribal Goṇḍs ruled over several districts.

The Muslim invasion of the area took place in the 11th century, and Gwalior was conquered. The Hindu domains there were incorporated into the Delhi sultanate in 1231 by the sultan Iltutmish. Later the Khaljī sultans of Delhi overran Mālwa, which was subsequently annexed into the Mughal Empire by Akbar (1556 to 1605), the greatest of the Mughal emperors. Marāṭhā power extended into Mālwa at the beginning of the 18th century and a large part of what is now Madhya Pradesh had come under Marāṭhā rule by 1760. With the defeat of the peshwas (hereditary Marāṭhā prime ministers who centralized Marāṭhā rule) in 1761, the Sindhia dynasty of the Marāṭhās was established at Gwalior in the north and the Holkar dynasty (also Marāṭhā) at Indore in the southwest.

The period of sultanate rule

In 1817–18 the "Saugor-Nerbudda territories," as they were known, were ceded to the British; 43 years later this area and certain other districts were grouped into the Central Provinces, Berār being added in 1903. On the north and west, the Central India Agency was formed in 1854, with Mālwa, Bundelkhand, and Baghelkhand as the main units.

When India became independent in 1947, the British Indian province of Central Provinces and Berār formed Madhya Pradesh. Boundary changes followed: the old Central Indian Agency (divided after independence into Madhya Bhārat and Vindhya Pradesh) was added to Madhya Pradesh, and eight Marāṭhī-speaking districts were detached from it and joined to Bombay (later Mahārāshtra) state. Bhopāl state was added to Madhya Pradesh in 1956, when its present boundaries were defined. (N.P.A./Ed.)

Boundary changes in 1947

Mahārāshtra

The state of Mahārāshtra occupies a substantial portion of the Deccan Peninsula. It is triangular in shape, with a base running north–south for nearly 450 miles along the western shores of India and jutting out in a blunted apex toward the east about 500 miles away. It has an area of 118,828 square miles (307,762 square kilometres). It is bounded by the states of Gujarāt, Madhya Pradesh, Andhra Pradesh, and Karnātaka (formerly Mysore) and on the west by the Arabian Sea.

Mahārāshtra's capital, Bombay, is an island city, connected to the mainland by roads and railways. Aptly called the gateway of India, it is one of India's biggest commercial and industrial centres and has played a significant role in the country's social and political life.

Mahārāshtra is one of the foremost states in agricultural and industrial production, in trade and transport, and also in the growth of education and culture. Its ancient culture, at one stage considerably obscured by the dominance of the British, survives largely through the medium of a strong literary heritage. A common Marāṭhī literature (Marāṭhī is the predominant language) has in fact played an important role in nurturing a sense of nationality and unity among the Maharashtrians.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Mahārāshtra presents an interesting range of physical diversity.

To the west is the narrow Konkan coastal lowland. This is widest near Bombay. Numerous minor hills dominate the relief. There are numerous small, swift, west-flowing streams, most of them less than 50 miles long. The biggest, Ulhās, rising in the Bhor Ghāt, joins the sea after an 80 mile course.

The Western Ghāts (which are mountains at the western edge of the Deccan Plateau; *ghāt* means "step" in Hindī) run almost continuously 398 miles north–south, the foothills reaching to within three–four miles of the Arabian Sea. Elevations increase northward to peaks of some 4,720 feet (1,440 metres). There are a few gaps through which roads and railroads link the coast with the interior. The eastern slopes of the Ghāts descend gently to the Deccan Plateau and are sculptured by the wide, mature valleys of the Krishna, Bhīma, and Godāvari rivers.

The Western Ghāts

Between Narmada Valley in the north, Krishna Basin in the south, and from the western coast as far east as Nāgpur, the Ghāts and the triangular plateau inland are covered with extensive lava outpourings called "traps." These reach a maximum thickness of 10,000 feet near Bombay. The differential erosion of lava has resulted in characteristic steplike slopes, uniform crest lines, and a table-top appearance of many hills.

East of Nāgpur, the Deccan Trap gives way to undulating uplands (890–1,080 feet high) underlain by ancient crystalline rocks. The Wardha–Wainganga Valley, part of the larger Godāvari Basin, trends southward and abounds in lakes.

The lava rock breaks down into a heavy, black, fertile soil. Very old crystalline rocks in the Wardha-Wainganga Valley have been eroded into light-coloured, sandy soils. Saline soils in the river valleys are a result of heavy evaporation.

The climate is characteristically monsoonal, with local

variations. India's southwest monsoon rains break on the Bombay coast usually in the first week of June and last until September, during which period they account for 80 percent of the annual rainfall. Four seasons are normal: March–May (hot and dry); June–September (hot and wet); October–November (warm and moist); and December–February (cool and dry).

The Western Ghāts and the ranges on the northern borders greatly influence the climate and separate the wet Konkan Coast from the dry interior upland, called the Desh. Rainfall is very heavy in Konkan, averaging 100 inches, with some of the wettest spots recording 250 inches rapidly diminishing to one-tenth of this east of the Ghāts. Rainfall increases again eastward, reaching 40–80 inches (1,000–2,000 millimetres) in the extreme east.

The coastal districts enjoy equable temperatures, monthly averages ranging, as at Bombay, only a few degrees above or below 80° F (27° C). A range of more than 13° F (7° C) between day and night temperatures is unusual. Towns such as Pune (Poona), high up on the plateau, benefit from cooler temperatures throughout the year but remain equable.

Forests cover less than one-fifth of Mahārāshtra and are confined to the Sahyādri Range, its transverse ranges, the Sātpura Range in the north, and the Chandrapur region in the east. On the coast and adjoining slopes, plant forms are rich with lofty trees, variegated shrubs, and mango and coconut trees. The forests yield teak, bamboo, myrabolan (for dyeing), and other woods.

Thorny, savanna-like vegetation occurs in areas of less than 30-inch rainfall, in upland Mahārāshtra.

Subtropical vegetation is found on higher plateaus having heavy rain and milder temperatures. Bamboo, chestnut, and magnolia are common. In the semiarid tracts, wild dates are found. Mangrove vegetation occurs in marshes and estuaries along the coast.

Wild animals include tigers, leopards, and even bison. There are several species of antelope. Striped hyena, wild hog, and sloth bear occur frequently. Monkeys and snakes are of many types. Game birds, the peacock, and ducks are numerous. There are national parks at Tādoba and Borivli.

The people. Mahārāshtra contains a mixture of races. Ethnically, it is a conglomeration of people—indigenous and immigrant. On the Sahyādri and Sātpura perches live Australoid aboriginals—Bhīls, Warlis, Goṇḍs, Korkus, and Gowaris. They are dark, short statured, and have long heads and flat noses. Almost ubiquitous are Kuṇbī Marāṭhās—brown, medium-statured, and supposedly descendants of waves of settlers who came from the north around the beginning of the Christian Era. Through an agelong mixture of aboriginals and immigrants, many castes have developed, most of them in the east. The tall and fair Pārsīs form another group, having fled from Iran, in the 11th century, to safeguard their religion.

Marāṭhī, the state language, is spoken by over 90 percent of the people. Other important languages are Gujarātī, Hindī, Telugu, Kannaḍa, Sindhī, Urdū, Bengali, Malayālam, and English. Sindhī has spread in urban centres as a result of an influx of refugees after the 1947 partition. English is still used in universities, in administration (along with Marāṭhī), and in metropolitan areas. Urdū is spoken by Muslims though a large number use local dialects. There are also many dialects, such as Koṅkaṇī on the west coast and Goṇḍī and Muṇḍārī in the eastern and northern forests.

Marāṭhī is one of India's more developed languages. Despite various regional spoken forms, its written style is uniform.

A variety of faiths is found within Mahārāshtra. Hindus predominate, followed by Muslims and Buddhists. There are many Christians in the metropolitan cities. Pārsīs (a religious minority adhering to Zoroastrianism) are mainly confined to Bombay and its environs; some are found in other cities. Other religious minorities include Jains and Sikhs, who are found in small numbers everywhere.

Major urban centres

About three-fifths of the population is rural, living in numerous villages. Greater Bombay's immense population reflects its industrial growth and commercial success. Textiles dominate, but engineering, food processing, chemical and pharmaceutical, and transport equipment manufacturing also are major industries. Bombay is India's best equipped port and handles an enormous foreign trade. It is a centre of business, finance, and administration.

Nāgpur, Pune, and Sholāpur are other major cities. Its historical and cultural importance apart, Pune has developed many industries because of its proximity to Bombay. Nāgpur was once the capital of the Bhonsle kingdom and then, until 1956, of the erstwhile Madhya Pradesh. The city still enjoys status as Mahārāshtra's second capital. Nāgpur and Sholāpur have textile and other agriculturally based, market-oriented industries. Pune and Nāgpur are additionally important as educational centres. Of particular historical interest is the Mughal city of Aurangābād, which contains several monuments and other historic buildings and which is in close proximity to the famous caves of Ajanta and Ellora.

The economy. Agriculture is the bulwark of Mahārāshtra's economy, and the peasantry forms some two-thirds of the population.

Major crops

Recent progressive measures to combat food deficits include electrifying irrigation, the use of hybrid seeds, ultra-intensive cultivation, and offering incentives to the farmer. A major continuing problem, however, is water scarcity. Millet and pulses (edible seeds) dominate the cropped area. Rice grows where rainfall exceeds 40 inches, and wheat is a winter crop in fields that retain moisture. Cotton, tobacco, and peanuts (groundnuts) are major crops in areas having 24–39 inches of rainfall. Irrigation dams in rain-shadow areas have resulted in a rich sugarcane crop. Mangoes, cashew nuts, bananas, and oranges are popular crops. Considerable success has been achieved in overcoming many problems relating to soil erosion, storage, transport, and marketing.

Forest products include timber, bamboo, sandalwood, and tendu leaves (for cheap cigarettes).

Meagre mineral resources are found mostly in the eastern and, occasionally, in the western districts. Manganese, coal, iron ore, limestone, copper, bauxite, silica sand, and common salt are extracted. Eastern districts such as Bhandāra, Nagpur, and Chandrapur are rich with reserves of coal. Mostly bituminous, it is used by railways and power stations.

Mahārāshtra produces both hydroelectricity and thermal electricity, the former in western areas, the latter in the east. The atomic power station at Tārāpur, 70 miles north of Bombay, was India's first nuclear power plant.

The Bombay–Pune complex is the state's major industrial area; Nāgpur Aurangābād, Sholāpur, Thāne, and Kolhāpur are also important.

The oldest and largest industry is the manufacture of cotton textiles. Bombay, Nāgpur, and Sholāpur are the principal textile manufacturing centres.

Oil refining and the manufacture of such items as agricultural implements, transport equipment, rubber products, electric and oil pumps, lathes, compressors, sugar-mill machinery, typewriters, refrigerators, electronic equipment, and television and radio sets are assuming increasing importance. The petrochemical industry has developed rapidly in the state since the development of India's first offshore oil fields near Bombay in 1976. The centre of India's motion picture industry is in the city of Bombay. There is also an incipient automobile industry in Bombay.

The foremost means of travel is a rail network focussing on Bombay. The nation's east–west and north–south trunk routes intersect between Nāgpur and Wardha. Passenger amenities are increasing, with sleepers, diesel-electric engines, and air-conditioned coaches.

A 20-year plan inaugurated after independence provided for a major road within five miles of every village. Missing links on the 86 state highways were thus completed, and district roads developed in inaccessible areas. Five national highways connect the state with Delhi, Calcutta, Allāhābād, Hyderābād, and Bangalore.

Water transport plays a limited role in Mahārāshtra, although there are minor ports on the western coast.

Daily air services connect Bombay with Pune, Nāgpur, Aurangābād, and Nāsik. Bombay is located on interna-

tional air routes, and Nāgpur is the centre of India's airmail scheme.

Administrative and social conditions. *Government.* Mahārāshtra has had always three conventional components: western Mahārāshtra, Vidarbha, and Marāthwādā. It is divided administratively into 30 districts.

In common with other states, Mahārāshtra is administered by a governor and a council of ministers elected from members of the legislature and headed by the leader of the majority party.

Executive authority in the state is exercised by the cabinet in the name of the governor. The district collector and chief executive officer—responsible for collection of land revenue, special taxes, and for coordinating the work of other departments, are the key figures within the local administrative areas.

The judiciary, headed by the chief justice and a panel of judges, is based in Bombay. There is a bench at Nāgpur.

The legislature has two houses: Vidhān Parishad (legislative council) and Vidhān Sabhā (legislative assembly). Both meet in Bombay and once annually in Nāgpur. Mahārāshtra is represented in the Lok Sabhā and the Rājya Sabhā (which are, respectively, the lower and upper houses of the Indian Parliament).

The Public Service Commission and a State Selection Board select candidates for appointment to all state services, largely by means of competitive examinations.

Health. There are several medical colleges, the most notable being in Bombay and Pune. There are dental colleges in Bombay and Nāgpur. Indigenous medicine is taught at Āyurvedic colleges in Bombay and Nāgpur. Most district hospitals maintain nursing schools. Regional blood banks exist in Bombay, Pune, Aurangābād, and Nāgpur, and resuscitation centres are found in all districts. The state has repeatedly received a national award for the best family-planning work. In Bombay the Haffkine Institute and the Indian Cancer Research Centre (located in the Tata Memorial Hospital) are well-known research centres.

Welfare. Welfare activities are directed toward children, women, workers, and delinquent citizens but not as yet to the unemployed. A Children's Act takes care of youthful offenders, and remand homes do correctional work. A number of state homes shelter women in distress. The aged have rest homes, and numerous hostels accommodate working women. Training facilities exist for the physically handicapped, and juvenile guidance centres function in slum areas.

Education. About one-half of the population is literate. Free compulsory education is provided for children between the ages of six and 14. Vocational and multipurpose high schools have also grown in importance.

There are two universities (including one university for women) at Bombay and others at Nāgpur, Pune, Aurangābād, Ahmednagar, Akola, and Kolhāpur. There are also three agricultural universities and several engineering and medical colleges. More than 100 colleges affiliated with the universities offer baccalaureate degrees.

Artistic traditions

Cultural life. Mahārāshtra is a distinct cultural region. Its long artistic tradition is manifested in the ancient cave paintings of Ajantā and Ellora, in medieval architectural masterpieces, in its classical and devotional music, and in its theatre. Pune, where numerous organizations sustain these great traditions, is the undisputed cultural capital.

In rural Mahārāshtra, the foremost diversion is *tamāshā*, combining music, drama, and dance. About seven artists form a *tamāshā* troupe. A female dancer is the central figure, a clown providing the humour. The theatre and the cinema are popular in urban areas.

Mahārāshtra has many festivals throughout the year: Holī is the spring festival and Ranga panchamī, which follows, is a riot of colours. The Dassera festival has a special significance, being the day on which Marāṭhā warriors started on their campaigns. Dīwālī, coming next, is a festival of lights and fireworks. Polā is a festival in which bullocks are given a holiday and decorated for a race. The Ganesh festival is by far the most popular in Mahārāshtra. Its public celebration was first sponsored by a nationalist political leader, B.G. Tilak. Muharram commemorates the great martyrs of Islām, although even Hindus participate. Unique to Mahārāshtra is the Hurda party, in which the farmer invites village folk to partake of fresh ears of *jowār* (sorghum). Folk songs and traditional dances accompany all these celebrations.

HISTORY

The name Mahārāshtra first appeared in a 7th-century inscription and in the account of a contemporary Chinese traveller, Hsüan-tsang. It then denoted the western upland, which has since been continuously known as Mahārāshtra. The origin of the name is still controversial. One view, however, is that it is derived from the word *raṭhī* (chariot driver), which refers to the skillful builders and drivers of chariots who formed a fighting force (*mahārāṭhis*) and migrated southward, settling in the area occupied by the modern state. Their language, intermingled with the speech of the earlier Nāga settlers, became Mahārāṣṭrī, and this by the 8th century developed into Marāṭhī. There was also a continuous influx of people from remote Greece and Central Asia.

Emergence of Marāṭhā culture

The people of this early period were separated among several Hindu kingdoms: Sātavāhana, Vākāṭaka, Kalacuri, Rāṣṭrakūṭa, Cālukyā (Chalukya), and Yādavas. After 1307 there was a succession of Muslim dynasties. Persian, the court language of the Muslims, had a far-reaching effect on Marāṭhī. By the mid-16th century Mahārāshtra was again fragmented among several independent Muslim rulers, who waged endless wars among themselves. It was in the midst of this chaos that a great leader, Chhatrapati Śivājī, was born. Śivājī (Shivaji) showed astonishing prowess by founding a large empire that shook Mughal rule to its foundations.

During the 18th century almost all western and central India, as well as large segments of the north and east, were brought under Marāṭhā suzerainty. It was this empire that succumbed to the British from the early 19th century on. When India became independent in 1947, the province, long known as Bombay Presidency, became Bombay state. The following year a number of former princely states (notably Baroda [Vadodara]) were merged into the new state, and, on November 1, 1956, a major reorganization of the states of peninsular India resulted in the addition of large parts of the former Hyderābād and Madhya Pradesh. In the north, it now comprised most of the Gujarātī-speaking peoples and, in the south, most of the Marāṭhī-speaking. As a result of the demands of the two language groups, the state was split in two on May 1, 1960, with Gujarāt, in the north, and Mahārāshtra, in the south. (S.Mo.)

SOUTH AND INSULAR INDIA

Andaman and Nicobar Islands

The Andaman and Nicobar Islands constitute a union territory of the Republic of India, centrally administered from Delhi. Formed in 1872 by uniting the Andaman islands and the Nicobar Islands under one British administration, the territory was included under Part D of the first schedule of the Indian Constitution of 1950. It is administered by the president of the republic through a lieutenant governor, who is advised by a local council of five members. The union territory is divided into two administrative districts, Andamans in the north, and Nicobars in the south; the administrative headquarters are at Port Blair on South Andaman, and an assistant commissioner is stationed on Car Nicobar in the Nicobar Islands. The total area of the territory is 3,202 square miles (8,293 square kilometres).

ANDAMAN ISLANDS

The Andaman Islands, a group of 204 islands and islets in the Bay of Bengal, have a total area of 2,448 square

miles (6,340 square kilometres). They lie 650 miles (1,050 kilometres) from the mouth of the Hooghly River and 120 miles from Cape Negrais in Burma, the nearest point on the mainland. The extreme length of the Andamans is about 310 miles (500 kilometres), the greatest width is less than 30 miles (48 kilometres). There are three chief islands in the main group—North, Middle, and South Andaman—so closely adjoining each other that they have long been known collectively as Great Andaman. Grouped with the chief islands are, on the extreme north, Landfall Island; Interview Island, off the northwestern coast of Middle Andaman; the Labyrinth Islands, off the southwest coast of South Andaman; Ritchie's Archipelago (or the Archipelago Islands), off the eastern coast of South Andaman; Baratang, wedged in between and on the east of Middle and South Andaman; and Rutland Island, at the southern end of South Andaman. Little Andaman forms the southern extremity of the whole Andaman group and lies 31 miles south of Rutland Island across Duncan Passage, in which lie the Cinque and other islands. There are a great number of islets lying off the shores of the main islands. The Andaman Islands are separated from the Nicobar Islands to the south by the 90-mile-wide Ten Degree Channel.

Physical and human geography. *The land.* The islands forming Great Andaman consist of a mass of hills enclosing very narrow valleys, the whole covered by dense tropical forest. The hills rise, especially on the east coast, to a considerable height; the highest is Saddle Peak (2,461 feet [750 metres]) on North Andaman. Little Andaman, with the exception of the extreme north, is practically flat. There are no rivers and few perennial streams in the islands. The scenery is everywhere strikingly beautiful and varied, and the coral beds of the more secluded bays in its harbours are conspicuous for their exquisite colouring.

The Andaman coasts

The coasts of the Andamans are deeply indented, forming a number of safe harbours and tidal creeks, often surrounded by mangrove swamps. The principal harbours are Port Blair in South Andaman, Diglipur in North Andaman, and Rangat and Māyābandar in Middle Andaman.

Rarely affected by a cyclone, though within the influence of practically every one that blows in the Bay of Bengal, the Andamans are of the greatest importance because of the accurate information relating to the direction and intensity of storms that can be communicated from them, better than from any other point in the bay, to the vast amount of shipping in this part of the Indian ocean. A well-appointed meteorological station was established at Port Blair as early as 1868. Generally, the climate of the Andamans may be described as normal for tropical islands of similar latitude. It is always warm but tempered by pleasant sea breezes, and it is very hot when the sun is northing. The rainfall is irregular, but usually there is very little during the northeast and a great deal during the southwest monsoon.

The chief timber of indigenous growth is Andaman redwood, or padauk (*Pterocarpus dalbergioides*), used for buildings, boats, furniture, fine joinery, and all purposes to which teak, mahogany, hickory, oak, and ash are applied. This tree is widely spread and forms a valuable export to European markets. Animal life is generally deficient throughout the Andamans, especially as regards mammalia, of which there are only 20 separate species in all, including the dugong of the surrounding areas. Fish are very numerous and many species are peculiar to the Andaman seas. Turtles are abundant and supply the Calcutta market. Principal crops are rice, coconuts, and area nuts. Small industries produce plywood and matchwood.

The people. The Andamanese belong to the Negrito stock, which is represented also by the Semang of Malaya and the pygmies of the Philippines. Their remoteness and their practice of killing all foreigners preserved them from any considerable physical or cultural modification. They lived by hunting and gathering. The dog, introduced in the 19th century, was the only domestic animal. The only indigenous weapon was the bow. Turtle, dugong, and fish were caught with nets and harpoons; the latter were used from single-outrigger canoes. The Andamanese were unique in knowing no method of making fire.

There were several tribes, each consisting of a number of small independent groups with a common language. Most Andamanese are now detribalized, but the old life survives among the Jarawa and Onge of the interior of South Andaman, Rutland Island, and Little Andaman.

History. Because of their position athwart the trade routes from India to Burma and the Far East, the existence of the islands has been known from earliest times. They are mentioned by the 7th-century Chinese Buddhist monk I'Tsing, the Arab travellers of the 9th century, Marco Polo (1286), and others. The derivation of the name Andaman is uncertain. It seems possible that the Agathou Daimonos Nesos (Good Spirit Island) of Ptolemy was a misunderstanding of some form like Agdaman. It has been suggested that Angamanain (the name used by Marco Polo) is an Arabic dual indicating "the two Andamans." The name is probably derived from the Malay Handuman, coming from the Hindu mythological figure Hanumān (meaning "monkey god"). Later travellers repeat the stories, too well founded, of the ferocious hostility of the people.

In September 1789 Capt. Archibald Blair, acting under instructions from the government of Bengal, established a penal colony on Chatham Island, in the southeastern bay of the Great Andaman, now called Port Blair but then Port Cornwallis. Two years later, urged by Adm. Sir William Cornwallis, the government transferred the colony—together with its name of Port Cornwallis—to the northeastern part of Great Andaman, where a naval arsenal was to be established. But the scheme did poorly, and in 1796 the government put an end to it. In 1844 the troopships "Briton" and "Runnymede" were driven ashore there. The natives showed their usual hostility, killing all stragglers. Outrages on shipwrecked crews continued to be so frequent that the question of occupation was revived, and in 1855 a project was formed for a settlement and a convict establishment. This scheme was interrupted by the Indian Mutiny of 1857, but as soon as the revolt was broken it was deemed more than ever urgent to provide such a resource because of the great number of prisoners falling into British hands. A new settlement was established at the beginning of 1858 near the site of Blair's original Port Cornwallis and, to avoid confusion, was named Port Blair. For some time, sickness and mortality were excessively high, but the reclamation of swamp and clearance of jungle on an extensive scale by Col. Henry Man was ultimately successful. The Andaman colony obtained a tragic notoriety from the murder, by a Muslim convict, of the viceroy, the Earl of Mayo, when on a visit to the settlement on February 8, 1872.

Penal colony established

The islands were occupied by the Japanese on March 23, 1942, and reoccupied by the government of India on October 8, 1945, when the abolition of the penal settlement, which had been going on gradually since 1926, was finally effected. As a result of the foundation of the penal settlement, there is a considerable Indian population. Vinayak Damodar Savarkar, the celebrated Indian revolutionary, was imprisoned for life (1911–37) in the Cellular Jail (declared a national monument in 1979) at Port Blair.

NICOBAR ISLANDS

The Nicobar Islands, a group of 12 inhabited and seven uninhabited islands in the Bay of Bengal, lie between Sumatra and the Andaman Islands. Their aggregate area is 754 square miles (1,953 square kilometres), and Great Nicobar (Loöng), the largest and southernmost of any size, has an area of 333 square miles. Principal of the central group of islands are Camorta and Nancowry (Nankauri), between them being a landlocked shelter called Nancowry Harbour. Chief island of the northern group is Car Nicobar (Pu), with an area of 49 square miles.

Physical and human geography. *The land.* The Nicobars form part of a great submarine mountain chain linking the Arakan Yoma range of Burma, through the Andamans and Nicobars, with the main mountain chain of Sumatra. Some of the islands have mere flat coral-covered surfaces; others, again, are hilly, Great Nicobar rising to 2,105 feet (642 metres). On that island there are numerous fast flowing streams, but the other islands generally lack sufficient fresh surface water.

Being situated between latitude 6° and 10° N of the Equator, the islands have a nearly equatorial climate. They are exposed to both monsoons, and smooth weather is experienced only from February to April and in October. Rain falls throughout the year, generally in sharp, heavy showers. The rainfall varies from 90 to 135 inches (2,300 to 3,400 millimetres), and the shade temperature from 64° to 92° F (18° to 33° C).

Snakes, lizards and chameleons, crocodiles, turtles and an enormous variant of the edible Indian crab are numerous; butterflies and other insects have not yet been systematically collected. The freshwater fish are reported to be of the types found in Sumatra.

The people. The Nicobarese are probably a mixture of Malay and Burmese (Talaing) strains. They are of relatively short stature and sturdy build; hair may be slightly wavy or curly, scanty on face and body; brachycephaly is usual. The language, which has several dialects, belongs to the Mon-Khmer group.

Articles of European clothing are now often worn. Houses vary in shape; they are built on piles, with thatched roofs, and are reached by a notched pole. Much cooking is done in them on clay hearths, but there are also separate cookhouses. Special houses are built by communal labour for feasts and meetings. The coconut is the main crop, and yams and other vegetables are grown; fish, turtles, pigs, and fowl are also important in the diet.

Ancient trade route

History. The situation of the Nicobars along a line of an ancient trade route caused them to be reported by traders and seafarers through historical times. In the 17th century the islands began to attract the attention of missionaries. At various periods, France, Denmark, Austria, and Great Britain all had more or less obscure rights to the islands, the Danes being the most persistent in their efforts to occupy the group, until in 1869 they relinquished their claims in favour of the British, who at once began to put down the piracies of the islanders and established a penal settlement, which was withdrawn in 1888. Car Nicobar in the north and Camorta in the centre became the principal ports of the group.

In 1942, during World War II, the Nicobar Islands were occupied by Japanese forces, who developed Car Nicobar as a big supply base. In 1945 the islands were reoccupied by the Allies. With the Andamans they passed to India on the establishment of Indian independence. (Ed.)

Andhra Pradesh

Andhra Pradesh is located in the eastern part of the Indian subcontinent, and it has an area of 106,204 square miles (275,068 square kilometres). It is bounded by the states of Tamil Nādu (formerly Madras) to the south, Karnātaka (formerly Mysore) to the west, and Mahārāshtra, Madhya Pradesh, and Orissa to the north; to the east it is bounded by the Bay of Bengal. The state capital is Hyderābād.

The state derives its name from the Andhra people, who have inhabited the area since antiquity, and who have developed their own language, Telugu. The state in its present form came into existence in 1956 as a result of the demand of the Telugu-speaking Andhras for a separate state. Although the state is primarily agricultural, there is some mining and a significant amount of industry.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The state has three main physiographic regions: the coastal plain to the east, extending from the Bay of Bengal to the mountain ranges; the mountain ranges themselves, the Eastern Ghāts, which form the western flank of the coastal plain; and the plateau to the west of the Ghāts.

The coastal plain runs almost the entire length of the state and is watered by several rivers, flowing from west to east through the hills into the Bay. The deltas formed by the most important of these streams—the Godāvari and the Krishna—comprise the central part of the plains, an area of fertile alluvial soil.

The Eastern Ghāts, which do not form a continuous range, being broken by the great river valleys, are part of a larger mountain system extending from central India to the far south and running parallel to the east coast. They have highly porous soils on their flanks.

The plateau to the west of the ranges is composed of gneissic rock (a foliated rock formed deep in the earth under conditions of heat and pressure); it has an average elevation of 1,600 feet above sea level. As the result of erosion, this is a region of graded valleys, with red, sandy soil and isolated hills, which support a thorny vegetation. Black soil is found in certain parts of the area.

Coastal vegetation

Near the coastal plains mangrove swamps and palm trees are found, while the deltas constitute a rich agricultural belt where food crops, fruit, and tobacco are grown. The soil on the plateau and the uncertain rainfall allow only dry crops to be raised, although with the extension of irrigation, rice and sugarcane are also being grown. Of the state's total area, about one-quarter is forested. The forests consist of both moist deciduous and dry forest vegetation; teak, rosewood, bamboo, and wild fruit trees are found. Cashews are grown in the coastal districts. Common trees such as the neem (a tree which produces an aromatic oil), banyan, mango, and pipal (*Ficus religiosa*) also occur, as well as flowering plants—jasmine, rose, and many local varieties. Animal life, apart from common domestic types (dogs, cats, and cattle), includes tigers, leopards, bears, and deer, which abound in the hills and forest areas.

A summer that lasts from March to June, a July-to-September season of tropical rains, and a winter from October to February constitute the three seasons. The range of maximum and minimum temperatures is from 74° F to 82° F (23° C to 28° C) and from 50° F to 53° F (10° C to 12° C), respectively. Summers are very warm in the coastal plains, with temperatures in places reaching 108° F (42° C). On the plateau, summers are cooler and winters colder. Rainfall is to a large extent caused by the southwest monsoon (rain-bearing) winds; some places receive a maximum of 55 inches of rain and others only 20 inches. In coastal areas, rainfall is heavy; in some areas on the plateau, especially in the north and west, it is sparse.

Languages and religions

The people. The state's cultural pattern is revealed by the distribution of its languages and religions. Telugu is spoken by most of the population, and a small minority speak Urdū, the language of the Muslim population. The remaining groups consist of people speaking border area languages, including Tamil, Kannaḍa, and tribal languages such as Lambadī. The great majority of the population are Hindus, and there are both Muslim and Christian minorities. The Christians live in urban centres and coastal areas, and the Muslims are concentrated in the Telangāna area.

Somewhat less than one-quarter of the population are urban dwellers, and a third of these are concentrated in the industrial and manufacturing areas around the three main cities—Hyderābād, Vishākhapatnam, and Vijayawāda. With increasing industrial development, these cities are becoming linked with neighbouring urban areas, forming agglomerations. Some important industries and several smaller enterprises, such as sugar factories and service industries, are scattered in medium and smaller urban areas.

The literacy rate is about 40 percent among men but only about 20 percent among women. In the Hyderābād, Krishna, and Godāvari districts literacy is higher than elsewhere.

The economy. Agriculture is the primary sector of the state's economy especially in the production of food grains. The state produces about four-fifths of India's Virginia tobacco and is one of the largest rice-producing states in the country. Andhra Pradesh is also known for such cash crops as tobacco and sugarcane, in addition to chilies, oil seeds, and pulses (leguminous crops, such as peas, beans, and lentils).

Principal crops

The state's rivers, particularly the Godāvari and the Krishna, account for its agricultural importance; for a long time their benefits were restricted to the coastal districts of Andhra region, which had better irrigation facilities. Since independence, great efforts have been made to tap the waters of these and other rivers for the benefit of dry regions, such as Rayalaseema and Telangāna. The Nāgarjuna Sagar multipurpose project, diverting the waters of the Krishna for irrigation, has increased substantially the

production of rice, food grain, and sugarcane. Half of the total investment for development is allotted to agricultural irrigation. The state's forest areas annually yield high quality timber and forest produce.

Among the state's mineral resources are asbestos, mica, manganese, barite, and high-grade coal. Low-grade iron ore is found in the southern parts of the state. Andhra Pradesh produces almost all of the country's chrysolite asbestos and about three-quarters of India's barite. It is the only state in southern India that possesses coal reserves. Production at the once world-famous diamond mines of Golconda, where the Koh-i-noor diamond and other famous stones were found, is being renewed. Quartz, limestone, and graphite are also found. For the exploitation of its mineral resources the state has established a mining corporation.

Andhra Pradesh, once industrially underdeveloped, has in the latter half of the 20th century become one of the highly industrialized states in India. Since independence, industries such as shipbuilding, aeronautics, and plants producing electrical equipment, machine tools, and drugs have been established by the central (union) government in the Vishākhapatnam and Hyderābād areas. Private enterprises include chemical, textile, cement, fertilizers, food products, petroleum products, and cigarette manufacturing, and light industry; many of these enterprises are located in the Vijayawāda-Guntūr region. An oil refinery is located in Vishākhapatnam as is the largest shipbuilding yard in the country. Mining and manufacturing industries, however, account for a small percentage of the state's income. The central government has sanctioned the construction of a mammoth steel plant at Vishākhapatnam, where there is easy access to raw materials and port facilities. The phenomenal increase in recent years in power generated by hydroelectric and thermoelectric projects augurs well for industrialization and irrigation. With its long coastline and many rivers, the state also has a significant and expanding fishing industry.

Federal investment in industrialization

There are four civilian airports—at Hyderābād, Vijayawāda, Tirupati, and Vishākhapatnam. The rail system serves a major part of the state and connects it with other parts of India. There is an extensive road system, which includes three national highways connecting all parts of the country. Other roads are maintained by state and local governments. Bus transportation is mostly nationalized and offers facilities for express travel between various cities.

The river canals in coastal areas, especially the saltwater Kommamur (Buckingham) Canal running parallel to the coast from the Krishna River to Madras, are used for cargo transportation. Vishākhapatnam is a major seaport used by international shipping.

Administrative and social conditions. *Government.* Andhra Pradesh is a constituent unit in the Indian Union. A governor, appointed by the president of India, is the executive head of the state administration, but the real power is in the hands of a chief administrator and a cabinet of ministers responsible to the state legislature. The state has a unicameral legislature consisting of a legislative assembly elected by adult suffrage from territorial constituencies.

The administration is conducted by various departments, each under the direction of a minister, assisted by a staff of permanent civil servants. The State Secretariat at Hyderābād supervises the administration of the state's 23 districts. Local administration in each district is the responsibility of a district collector. Rural local government has been democratically decentralized by the introduction of a system in which local authorities operate at the village, block (a unit consisting of a group of villages), and district levels. Urban areas are governed by municipal bodies.

Features of local government

A special feature of state government is the regional committee for Telangāna (which consists of the nine Telugu-speaking districts that were formerly a part of Hyderābād state); the duty of the committee is to ensure that the views of the people of Telingāna are given adequate consideration. The committee was established when the region was joined to Andhra Pradesh in 1956, in order to protect Telangāna interests, since the area was economically and educationally less advanced than the Andhra areas.

The state judiciary is headed by a High Court, located in Hyderābād; the High Court has original jurisdiction in some cases and exercises appellate and administrative control over the district and lower level courts. The High Court is itself subject to the appellate authority of the Supreme Court of India in certain matters. Within the state are located many federal instrumentalities, including the railways administration, the Income Tax Department, and other establishments. A number of defense establishments are located at Hyderābād (Secunderābād), and Vishākhapatnam is the headquarters of the Eastern Naval Command.

Health. Medical and health facilities, under the supervision of the Director of Medical and Health Services, have expanded rapidly in recent years. Under the Primary Health Centres scheme, medical help, both curative and preventive, is being brought to rural areas. An increase in the number of in-patient dispensaries and in the number of beds available in hospitals has augmented urban medical facilities. The state's principal hospitals, such as the Osmania Hospital at Hyderābād and the King George Hospital at Vishākhapatnam, for example, have been improved; and new institutes, including those for treating particular diseases, have been opened. There is also a family-planning program. Medical aid is free to low-income groups, and several medical-insurance schemes cover various categories of employees.

Education. The literacy rate stands at about 30 percent. The state's educational system provides for ten years of schooling followed by a two-year junior college course leading to undergraduate and postgraduate education. Primary education has been free and compulsory since 1961. High school education, primarily the concern of local bodies, is also free. Privately run high schools also exist, mostly in urban areas. The state has nine universities—at Hyderābād (four), Waltair, Warangal, Guntur, Anantpur, and Tirupati—including the agricultural university at Hyderābād. Affiliated to each university are a number of colleges. Postgraduate instruction and research facilities are available at the three university centres as well as at other postgraduate centres. The Central Institute of English and Foreign Languages is at Hyderābād. For several years technical education has been receiving special attention in order to meet the demands of industrialization. Various industrial-training institutes provide for vocational training, while the engineering colleges of the three universities train advanced technical personnel.

Higher education

Welfare. Before independence, social welfare work was mainly undertaken by private agencies. After independence, however, the magnitude of the problem and scarcity of resources, both organizational and financial, were acknowledged; and this led the government to accept primary responsibility in this field. Public investment in social welfare accounts for a very large proportion of all that is spent on planning. There are social-welfare programs for disabled and handicapped persons, scheduled castes (formerly called "untouchables"), and tribal or other groups that are not fully integrated into the social structure. Social programs include those for education, economic advancement, health, and housing. Scholarship programs are available for all scheduled caste, tribal, and other disadvantaged groups in all educational institutions that receive substantial financial assistance from state and federal agencies. In addition to the policy of reserving places for these groups in educational institutions and in employment, other ameliorative measures include housing schemes and land-distribution schemes. Progress toward social advancement has, nevertheless, been rather slow. A separate government department for women's welfare exists. There are also many private social organizations, including the Andhra Mahilā Sabhā, an organization that works for the cause of women's welfare.

Cultural life. The Andhras' contribution to India's cultural heritage is substantial. Their legacy in art and architecture has already been noted. The *kuchipudy* style of dance is unique in the Indian tradition, while Carnatic (South Indian) music has derived much from Andhra roots. Many of southern India's major classical composers have been Andhras, and Telugu has been the language of

most of the compositions. Telugu, one of the four literary languages of the Dravidian family, occupies a prestigious place among Indian languages, being renowned for its antiquity as well as for its mellifluous quality. Andhras also are prominent in the modern Indian literary renaissance, and their writing reflects the impact of the contemporary revolution in literary forms and expression. There are many periodicals in English, Telugu, and Urdū. Muslim culture in the Telangāna region further enriches the state's cultural diversity.

Before independence, arts and literature thrived mostly under the sponsorship of private organizations, many of which still function. The state has also created autonomous academies to revive, popularize, and promote fine arts, dance, drama, music, and literature.

Folk culture

The conscious pursuit of culture is more an urban than a rural phenomenon, for cultural performances, literary meetings, and religious discussions occur mostly in towns or cities. In rural areas, folk culture predominates. Cultural evolution in different parts of the state under different historical circumstances resulted in the occurrence of recognizable variations in dialect, in the caste structure, and in habits, thus diversifying the folk arts. Folk media such as minstrel-ballad performances, puppet shows, and mythological storytelling, are indigenous to the area; use of these media in social and political communication is also common.

The penetration of the mass media, especially of radio, to rural parts has helped to bring an awareness of classical culture to the rural folk, and of folk arts to the urban population. Andhra Pradesh is among the few major movie-making states of India.

HISTORY

Though Sanskrit writings dating to around 1000 BC contain references to a people called "Andhras" living south of the central Indian mountain ranges, historical evidence of the Andhras dates from the times of the Mauryan dynasty, which ruled in the north in the 3rd century BC.

The mission from Aśoka

The great Mauryan emperor Aśoka (*c.* 265–238 BC) sent Buddhist missions to the Andhras in the south. Around the 1st century AD, the Sātakarṇi, or Sātavāhanas, one of the most renowned of the Andhra dynasties, came to power. Its members ruled over almost the whole of the Deccan Peninsula and even established trade relations with Rome. They were patrons of diverse religions and were also great builders: their principal city, Amarāvatī, contained Buddhist monuments inaugurating a new style of architecture. Experts ascribe parts of the famous paintings in the Ajantā caves of the Deccan to the Andhra painters of that period. Buddhism prospered under the Andhras and in their capital flourished the great Buddhist university of antiquity, where Nāgārjuna (*c.* AD 150–250), the founder of the Mahāyāna school of Buddhism, taught. The ruins of the university at Nāgārjunakoṇḍa still attest to its past glory. In the 11th century, during the reign of the eastern Cālukyas, a dynasty that unified most of the Andhra area, the first of the Telugu poets, Nannaya, began translating the Sanskrit epic, the *Mahābhārata,* into Telugu, thus inaugurating the birth of Telugu as a literary medium. The dynasty of the Kākatīyas of Warangal in the 12th and 13th centuries extended Andhra power militarily and culturally; it was during their regime that the commercial expansion of the Andhras toward Southeast Asia reached its peak. By this time, however, the Muslims had established themselves in the north, and their invasion of the south led to the fall of Warangal in 1323. But the rise of the kingdom of Vijayanagar, to the southwest of Warangal, arrested the expansion of the Muslim power for some time. Acclaimed as the greatest kingdom in Andhra history and as one of the greatest in Indian history, Vijayanagar, under the rule of its great king Krishna Deva Rāya, who reigned from 1509 to 1529, became synonymous with military glory, economic prosperity, good administration, and artistic splendour. During this period, Telugu literature flourished. The formation of an alliance between the various neighbouring Muslim principalities, however, led to the fall of Vijayanagar in 1565, leaving the Muslims in control of the Andhra areas.

Arrival of the British and French

In the 17th century British and French traders began to involve themselves in Indian politics. As a result, successive Niẓāms (rulers) of Hyderābād, seeking to consolidate their kingdom against rivals, obtained first French and later British support. In exchange for their help, the British acquired from the ruling Niẓām the coastal Andhra districts lying to the north of Madras and later the hinterland districts. Thus, the major part of the Andhra country came under British rule. Part of the Telugu-speaking areas, known as the Telangāna region, remained under the Niẓām's dominion. The French, for their part, acquired a few towns in the Andhra country.

During the 19th century, which witnessed the rise of Indian nationalism, the Andhras came to the forefront of the movement. Leaders such as Kandukuri Veerasalingam pioneered in social reform. In the struggle against British rule, Andhra leaders played decisive roles. Pride in their historical and linguistic achievements led them to demand a separate province. Simultaneously, a movement was also organized to unite the Telugu-speaking peoples living under British rule and those under the Niẓām's administration. The Andhras demand for separate statehood became so insistent that when the central government refused to accede to the demand, a local leader, Potti Sreeramulu, fasted to death in 1953 to dramatize the issue. The government finally acceded to the people's request by creating on October 1, 1953, the Andhra state, which included the Telugu-speaking districts of the former Madras state, thus paving the way for the formation of linguistic states throughout India in 1957. The erstwhile state of Hyderābād, which had joined independent India in 1949, was split up and its nine Telugu-speaking districts (Telingāna) were joined to the Andhra state on November 1, 1956, to form the new state of Andhra Pradesh. In 1960, 221.4 square miles (573 square kilometres) in the Chingleput and Salem districts of Madras (now Tamil Nādu) were transferred to Andhra Pradesh in exchange for 410 square miles (1,062 square kilometres) from Chittoor district.

Demand for Andhran statehood

(R.V.R.C.R.)

Dādra and Nagar Haveli

The two former Portuguese enclaves of Dādra and Nagar Haveli, seized by Indian nationalists in 1954, were constituted as a union territory of India effective in 1961. The territory, with an area of 190 square miles (491 square kilometres), lies at the southernmost point of the coastal border between the states of Gujarāt and Mahārāshtra. The capital is Silvassa.

PHYSICAL AND HUMAN GEOGRAPHY

The central area of the territory consists of plains cut by a number of rivers, only one of which, the Damānganga, is navigable. (Damānganga, at its mouth, was once famous for its docks.) The northeastern and southeastern sectors of the area are hilly. The climate is typical of the region—extremely hot in summer, with an average annual rainfall of about 120 inches (3,050 millimetres). Forests cover nearly four-fifths of the territory.

About 90 percent of the population consists of Ādivāsi tribal peoples, the main language being Bhīlī (Bhīlodī). Farming is the chief occupation, much of it done on terraced land, with rice and ragi as the major crops. Some wheat and sugarcane also are grown. Timber production is mainly centred on the valuable native teak. There is very little large-scale industry; industrial estates have been set up in Piparia, Massate, and Khadoli for producing such items as chemicals and fertilizers.

HISTORY

Dādra and Nagar Haveli came under Portuguese control in the late 18th century. Nagar Haveli was obtained in 1783 as compensation for a vessel destroyed by the Marāṭhā navy. Dādra was received by treaty in 1785 as a kind of fief of the local ruler. After India achieved independence in 1947, Goan nationalists sought to break away from Portugal; and their first successes were the seizure of Dādra on the night of July 21, 1954, and their capture of Nagar Haveli two weeks later. A pro-Indian administra-

tion was formed in these enclaves, and on June 1, 1961, it requested accession to the Indian Union. The Indian government had already acknowledged their incorporation into the Union from the day of liberation but made it official on August 11, 1961. On September 24, 1962, an informal consultative council of 29 members was formed.

Daman and Diu

Daman and Diu, formerly part of the Indian union territory of Goa, Daman, and Diu, became a separate union territory when Goa was made a state in 1987. Daman is on the Gujarāt coast and has an area of 28 square miles (72 square kilometres). The island of Diu lies off the southeastern coast of Gujarāt in the Gulf of Cambay and has an area of 15 square miles.

PHYSICAL AND HUMAN GEOGRAPHY

Both Daman and Diu have small harbours, and the economy is based primarily on fishing and agriculture. The majority of the population is Hindu, with small Muslim and Christian minorities. Most of the people speak Gujarātī.

HISTORY

The Portuguese, under treaty with Bahādur Shāh of Gujarāt, built a fort at Diu in 1535. Daman (known in Portuguese as Damão) was conquered by the Portuguese in 1559. Both areas were subject to the governor general of Goa as part of the Portuguese overseas province Estado da India (State of India). They remained under Portuguese rule until 1961, when they were occupied by Indian forces.

Goa

Before it became a state in 1987, Goa, together with Daman and Diu, constituted a union territory of India. Goa is bounded by the Terekhol, or Araundem, River (north), the Western Ghāts (east), the North Kanara district of Karnātaka state (south), and the Arabian Sea (west). Formerly a Portuguese possession, it was annexed by India in 1962. The total area of Goa (district and island) is 1,429 square miles. The capital is Panaji (Panjim).

PHYSICAL AND HUMAN GEOGRAPHY

Goa, with a coastline of 62 miles, is hilly, especially in the Novas Conquistas, including a portion of the Western Ghāts rising to nearly 4,000 feet. The two largest rivers are the Mandāvi and Juari, which together encircle the island of Goa (Ilhas), being connected on the landward side by a creek. The island is triangular, the apex (called the cape) being a rocky headland separating the harbour of Goa into two anchorages. There are three principal cities in Goa: Old Goa, Panaji (New Goa), and Margao. Old Goa is, for the most part, a city of ruins. Panaji, originally a suburb of Old Goa, is built (like the parent city) on the left bank of the Mandāvi estuary. It is a modern port and contains the archbishop's palace and the government house; it has a medical college, a postgraduate centre, a teachers' training college, and several secondary and primary schools. Margao, with its modern breakwater and quay and sheltered by a promontory, is the best port between Bombay and Kozhikode (Calicut). A railway connects it with Castle Rock on the Western Ghāts and so with the southern railway.

Goa exports coconuts, fruit, spices, manganese and iron ores, bauxite, fish, and salt; its trade is small but its manufacturers produce fertilizers, sugar, textiles, chemicals, iron pellets, and pharmaceuticals. Rice is the staple product, with fruit, salt, coconuts, pulses, and betel (areca) nut.

The Goan population is a mixture of Christian and Hindu. The Christians generally spoke Portuguese but now tend to speak English and Koṅkaṇī. The Hindus speak Koṅkaṇī and Marāṭhī. Economic conditions in Goa caused emigration on a large scale, mainly to the eastern coast of Africa and to other parts of India. Large Goanese colonies have consequently been formed in Bombay, Mozambique, Natal, and elsewhere. Many Goanese bear Portuguese names as a result of intermarriage between early Portuguese settlers and the local inhabitants.

HISTORY

The ancient Hindu city of Goa, of which hardly a fragment survives, was built at the southernmost point of the island, and it was famous in early Hindu legend and history. In the *Purāṇas* and certain inscriptions, its name appears as Gove, Govapuri, and Gomant. The medieval Arabian geographers knew it as Sindabur, or Sandabur, and the Portuguese as Goa Velha. It was ruled by the Kadamba dynasty from the 2nd century AD to 1312 and by Muslim invaders of the Deccan from 1312 to 1367. It was then annexed by the Hindu kingdom of Vijayanagar and later conquered by the Bahmani dynasty, which founded Old Goa in 1440.

With the subdivision of the Bahmani kindgom after 1482, Goa passed into the power of Yusuf Adil Shah, the Muslim king of Bijapur, who was its ruler when the Portuguese first reached India. At this time Goa was important as the starting point of pilgrims from India to Mecca, as a market with no rival except Kozhikode on the western coast, and especially as the centre of the import trade in horses (Gulf Arabs) from Hormuz. It was easily defensible by any power with command of the sea and was attacked in March 1510 by the Portuguese under Afonso de Albuquerque. The city surrendered without a struggle, and Albuquerque entered it in triumph.

Three months later Yusuf Adil Shah returned with 60,000 troops, forced the passage of the ford, and blockaded the Portuguese in their ships from May to August, when the cessation of the monsoon enabled them to put to sea. In November, Albuquerque returned with a larger force and, after overcoming a desperate resistance, recaptured the city, massacred all the Muslims, and appointed a Hindu, Timoja, governor of Goa.

Goa was the first territorial possession of the Portuguese in Asia. Albuquerque and his successors left almost untouched the customs and constitutions of the 30 village communities on the island, only abolishing the rite of suttee (*satī*). A register of these customs (*Foral de usos e costumes*) was published in 1526.

Goa became the capital of the whole Portuguese empire in the east. It was granted the same civic privileges as Lisbon. In 1542 St. Francis Xavier mentions the architectural splendour of the city; but it reached the climax of its prosperity between 1575 and 1600. The appearance of the Dutch in Indian waters was followed by the decline of Goa. In 1603 and 1639 the city was blockaded by Dutch fleets, though never captured, and in 1635 it was ravaged by an epidemic. In 1683 a Mughal army saved it from capture by Marāṭhā raiders, and in 1739 the whole territory was attacked by the same enemies and saved only by the unexpected arrival of a new viceroy with a fleet.

The seat of the government was moved to Mormugão (now Margao) and in 1759 to Pangim (now Panaji). Cholera epidemics were one of the chief reasons for the migration of the inhabitants from Old Goa to New Goa. Between 1695 and 1775 the population of Old Goa dwindled from 20,000 to 1,600; and in 1835 it was inhabited by only a few priests, monks, and nuns.

During the 19th century, events of importance affecting the settlement were its temporary occupation by the British in 1809 as a result of Napoleon's invasion of Portugal; the governorship (1855–64) of Conde de Torres Novas, who inaugurated a great number of improvements; and the military revolts of the second half of the century. The most notable of these was the revolt of September 3, 1895, which necessitated the dispatch of an expeditionary force from Portugal. The infante Affonso Henriques, duque de Oporto, accompanied this expedition and exercised governor's powers from March to May 1896.

After Indian claims on Goa in 1948 and 1949, Portugal came under increasing pressure to cede Goa, with its other possessions in the subcontinent, to India. A crisis was reached in 1955 when *satyāgrahis* (nonviolent resistors) from India attempted to penetrate the territory of Goa. At first the *satyāgrahis* were deported; but later, when large numbers attempted to cross the borders, the Portuguese authorities resorted to force, and casualties were inflicted. This led to the severance of diplomatic relations between Portugal and India on August 18, 1955. Tension between

India and Portugal came to a head when on December 18, 1961, Indian troops supported by naval and air forces invaded and occupied Goa, Daman, and Diu. Portuguese India was, by constitutional amendment, incorporated into the Indian Union in 1962. (Ed.)

Karnātaka

The state of Karnātaka (formerly Mysore) is situated on the western coast of India. With an area of 74,051 square miles (191,791 square kilometres), it coincides approximately with the area in which the Kannaḍa, or Kanarese, language is spoken. Karnātaka is a Kannaḍa name meaning "lofty land." The state is bounded by the states of Goa and Mahārāshtra to the north; by the states of Andhra Pradesh to the east, Tamil Nādu to the southeast, and Kerala to the south; and by the Arabian Sea to the west. The state extends for about 420 miles from north to south and for about 300 miles from east to west. It has a coastline 200 miles long. The state capital is the city of Bangalore, near the southeastern border.

Before the independence of India in 1947, Mysore was a prosperous and progressive but landlocked princely state, with an area of less than 30,000 square miles and a population of about 7,000,000 situated on the Karnātaka Plateau. The transfer of additional territories to the state in 1953 and 1956 united the Kannaḍa-speaking peoples, gave the state an outlet to the sea, and greatly extended its boundaries. The state took its present name in 1973. Karnātaka has abundant hydroelectric power and extensive forests. It is also India's chief source of gold and silver and coffee, and provides the bulk of the world's supply of sandalwood.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Physiographically, Karnātaka is divided into four distinct regions—the coastal plain, the hill ranges (the Western Ghāts), the Karnātaka Plateau to the east, and the black cotton-growing soil tract to the northwest. The coastal plain represents a continuation of the Malabār Coast and experiences heavy rainfall from the southwest monsoon during June through September. Coastal sand dunes give place inland to small alluvial plains with coconut-fringed lagoons. The coast is difficult of access, except by sea. To the east the land rises sharply to the slopes of the Western Ghāts, which have an average height of 2,500 to 3,000 feet above sea level. Covered with forest, the Ghāts form a region known as the *malnād;* it is also a watershed, and from its crest numerous swift streams flow to the plains, including the Sharāvati River, which is associated with the Jog (Gersoppa) Falls (830 feet [253 metres] in height). Other streams flow over the undulating Karnātaka Plateau, which slopes gently eastward. The plateau region has an average height of about 1,500 feet above sea level; its soils are generally porous and infertile. The basins of its rivers, however, which include the Cauvery to the south and the Tungabhadra, which is a tributary of the Krishna River, to the north, are loamy and of some fertility. From early times the smaller streams have been dammed to form irrigation reservoirs; in recent times much hydroelectric power development has taken place. In the northwestern part of the state, underlying volcanic rock produces a soil known as regur, the humus-rich, black cotton-growing soil of India.

Passing from west to east, the *malnād* gives way on the plateau to more open plain country, known as *maidān* where rainfall is less and where monsoon forests are replaced by scrub forests and scrubland. The monsoon forests are rich in wildlife, which includes tigers, elephants, bison, and deer. In the *maidān*, wild boar, bears, panthers, and leopards may be encountered. Among the common birds is the peacock. The state has a wildlife sanctuary located at Daudeli and national parks at Baudihur and Nagarhole.

The Dravidian population

The people. Ethnologically, the population of Karnātaka is predominantly Dravidian. The Dravidians were the original inhabitants of India and are related to the peoples of the Malay Peninsula, Sumatra, and Madagascar. Over the centuries, however, much intermixture with other races has occurred. In the north of the state, Aryan traits, mixed with Dravidian, occur. Some Anglo-Indians, representing a mixture of European and Dravidian, are occasionally encountered, particularly in Bangalore.

Kannaḍa is the language spoken by more than 70 percent of the population, although Hindustanī (a mixture of Hindī and Urdū) is often used in trade and business. Toward the borders of the state other languages, such as Tamil, Telugu, Marāṭhī, and Koṅkaṇī, are also spoken. Koṅkaṇī is associated particularly with the city of Mangalore. The major religion is Hinduism, but Jainism and Buddhism—once widespread—still survive. Small percentages of the population follow Islām and Christianity.

Although about three-quarters of the population is rural, the trend toward urbanization is becoming more marked as industrialization increases. The major cities are Bangalore, the capital, and Mysore. Other important towns are Dhārwār-Hubli, Mangalore, Belgaum, Gulbarga, Bellary, Dāvangere, Bijāpur, Shimoga, and Bhadrāvāti.

The economy. The economy of Karnātaka is based on its forests, its rivers, its coasts, and its mineral wealth.

The forests of the *malnād* region in the west produce teak, sandalwood, bamboo, and such other products as tanning dyes, gums, and lac (a resinous substance used in the manufacture of varnishes). Other trees include eucalyptus and rosewood. Oil processed from sandalwood at Mysore city is a leading state export.

Agriculture is the occupation of 80 percent of the population. The coastal plain is intensively cultivated, with rice grown as a food crop and sugarcane as a cash crop. The second and third major food crops after rice are *jowar* (sorghum) and *ragi* (millet). Other cash crops are cashews, cardamom, areca nut, and grapes. In the Western Ghāts coffee and tea plantations are established. Much of India's coffee is produced there, especially in the southern part; tea is grown on the higher hills. In the eastern region, irrigation makes possible the cultivation of sugarcane, a little rubber, and such fruits as bananas and oranges. The black soil of the northwest is devoted to the cultivation of cotton, oilseeds, and peanuts (groundnuts).

Tea and coffee plantations

Karnātaka has sufficient hydroelectric power to supply surplus energy to neighbouring states. The Sharāvati project near Jog Falls is the largest of several hydroelectric plants that provide power to Karnātaka's industries.

In places the Precambrian rocks of Karnātaka, which are from 4,600,000,000 to 570,000,000 years old, are highly mineralized. Karnātaka is the largest producer of chromite in India; it is also one of only two states in India (the other being Tamil Nādu) producing magnesite. High-grade iron ore reserves are largely located in Chikmangalūr and Chitradurga districts; mica, copper ore, bauxite, garnet, and emery are obtained in small quantities. From the region of the Kolār Gold Fields, all of India's gold is obtained. The principal gold mines, now nationalized, have workings up to 9,000 feet deep. The mineral resources of the state provide the basis for the iron and steel industry at Bhadrāvati and heavy engineering works at Bangalore. Other industries in the state include cotton milling, sugar processing, and the manufacture of textiles, food products, electrical machinery, fertilizers, cement, and paper. Both Mysore city and Bangalore have old established silk industries producing most of India's mulberry silk.

The obstacle formed by the Western Ghāts has prevented the building of railroads linking the ports to the plateau in the interior. Bangalore, in the southeast, is the main focus of rail transportation. The port of Mangalore, in the southwest, is connected with Bombay through the state of Kerala by tracks running parallel to the coast.

The import and export trade relies primarily on road transport, but many roads in the western part of the state become impassable during the rainy season. National highways run from Bangalore east to Madras, north to Hyderābād, northwest to Bombay, and west through Hassan to the coast of Mangalore. Airports are located at Bangalore, Belgaum, and Mangalore.

Administrative and social conditions. *Government.* The state is divided into 19 districts. The head of state is the governor, who is appointed by the president of India. The bicameral legislature consists of a legislative assembly of directly elected members. The chief minister is assisted by

a council composed of Cabinet ministers. The state high court is subordinate to the Supreme Court in New Delhi; it consists of a chief justice and several additional judges, who are appointed by the president of India in consultation with the chief justice of India and the governor of the state. There are also district and subordinate courts. A public service commission, whose members are appointed by the governor, functions in an advisory capacity.

Education. The state, which has a population that is about two-fifths literate, is one of the educationally advanced states of India. There are a large number of schools and educational institutions, including the University of Mysore at Mysore, the Karnātaka University at Dhārwār, and the Gulbarga University, the Mangalore University, Bangalore University, the University of Agricultural Sciences, and the Indian Institute of Science, all at Bangalore. Nearly half of the schools and institutions are managed by the government, and the remainder by local boards and private bodies. Compulsory free primary education is provided in most towns and villages in the state.

Welfare. The Employees State Insurance Scheme covers sickness, maternity, and employment injury risks and provides free medical treatment to factory workers and their families in Bangalore. Welfare schemes are run by the government for tribal and other scheduled groups, such as the Harijans. The State Social Welfare Board also sponsors family welfare centres.

Cultural life. Karnātaka possesses a rich cultural heritage, compounded by the contributions of successive dynasties, which have fostered various religions and philosophies, which, in turn, have influenced literature, architecture, folklore, music, painting, and minor arts.

Mauryan architecture is to be seen at Sravana Belgola, 56 miles from Mysore, where there is a giant stone figure, believed to be 1,000 years old, of Bāhubali (Gommateśvara), the Jaina saint. Huge monolithic Jaina statues are peculiar to the Kannaḍa culture. The influence of the Cālukya (Chalukya) and the Pallava dynasties is still apparent in temple architecture dating from the 7th century AD.

HISTORY

The name Mysore (from the Hindu word for "buffalo town") derives from the destruction of the buffalo-demon Mahiṣāsura by the goddess Cāmuṇḍā. The prehistory of Mysore is lost in legends that concern the struggle that took place in southern India between invading Aryan peoples and the original inhabitants; in legendary form this struggle is represented as a conflict between devils and demons on the one hand and gods and goddesses on the other. The subsequent history of the region deals mainly with the princely state of Mysore as it was before 1953, for no dynasty succeeded in ruling the whole region occupied by the Kannaḍa-speaking peoples.

After the reign of Aśoka, the Mauryan emperor of India, in the 3rd century BC, the principal dynasties in the area of Mysore were the Kadambas of Banavasi, the Western Gaṅgas (who were in power from the 3rd to the 11th century AD), the Bānas, and other feudatories of the Pallava dynasty, which ruled from about the 4th to the 9th century. The rich lands of the upper Tungabhadra region and the land between that river and the Krishna River were taken from the Kadambas in the 6th century by a central Karnātaka family, the Cālukyas. Their efforts and those of their competitors of the Rāṣṭrakūṭa dynasty, to unite the plateau and exploit the softer lands of the coastal plains enriched Mysore but led to reprisals from the Tamils to the east and south. By the 12th century the Hoysaḷa dynasty had absorbed Gaṅgavāḍi (as the state of Mysore was then called); but after the Hoysaḷas had been obliged to submit to the sultan of Delhi, Mysore gradually came under the sway of the state of Vijayanagar, whose capital of the same name stood on the site now partly occupied by the village of Hampi on the Tungabhadra River in contemporary Karnātaka. In the latter part of the 16th century the Vijayanagar Empire faded, giving place to Mughal power north of the Tungabhadra River and to the rajas of Mysore in the south. In the 17th century the *wadiyars* ("rulers") of Mysore profited from the conflict between the Mughal Empire and the Marāṭhās in western India, as well as from the internal power struggles that occurred in the Mughal Empire after the death (1707) of Aurangzeb, to expand their rule. In 1610 the *wadiyar* of Mysore seized Srirangapatnam (Seringapatam); later, Bangalore was also acquired and *wadiyar* power consolidated. Later, maladministration at home and interference in wars of succession in the plains led to the usurpation of power in 1761 by the military adventurer Hyder Ali. His invasions of Malabār and the Karnātaka plains extended Mysore's dominion but eventually led, after the resulting Mysore Wars, to the death of his colourful and energetic son Tippu Sultan in 1799 and to the conquest of Mysore by the British, who sponsored the restoration of *wadiyar* rule. Mysore was governed by a British commissioner from 1831 to 1881, when administration was once again restored to the *wadiyars*. The last of the *wadiyars* became governor of the state after the territorial reorganizations of 1953 and 1956. The state was renamed Karnātaka in 1973. (G.K.G./Ed.)

Kerala

Kerala, one of the smallest states in the Republic of India, was formed in 1956. It has an area of 15,005 square miles (38,863 square kilometres), about one percent of the area of India.

The state stretches for about 360 miles along the Malabār Coast on the western side of the Indian peninsula; its width varies from 20 to 75 miles. It is bordered by Karnātaka state to the north, Tamil Nādu (formerly Madras) state to the east, and the Arabian Sea to the west. The capital is Trivandrum. Isolated from the Deccan Plateau by the mountainous belt of the Western Ghāts, but with a long coastline that opens it to foreign influences, Kerala has evolved a unique culture. It is a highly politicized region, but it has a long tradition of religious amity. It is an educationally advanced state with its own language, Malayālam, and has the highest rate of literacy among India's states. Women in Kerala enjoy a high social status, thanks perhaps to its historic matrilineal system. But the hills of Kerala are inhabited by some of India's least advanced tribes.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Kerala is a region of great natural beauty. From the majestic heights of the Western Ghāts the country undulates westward, presenting a vista of silent valleys clothed in the richest green. Among the many rivers that glide across the plains to merge their waters with the Arabian Sea, the more important are the Periyār, the Pamba, and the Bharatapuzha. Along the coast, sand dunes shelter a linked chain of lagoons and backwaters the still waters of which are studded with sea gulls and country canoes plying at a snail's pace. The silence of their clear skies is broken only by the coos of koels, a type of cuckoo, and the frequent flutter of cranes perched on the embankments. The highest peak of peninsular India, Anai Mudi (8,842 feet [2,695 metres], is located in the state.

The climate is equable and varies little from season to season. The temperature normally ranges from 80° to 90° F (27° to 32° C) in the plains but drops to about 70° F (21° C) in the highlands. The state gets its due share of both the southwest and northeast monsoons, and the rainfall is heavy, averaging about 118 inches annually.

Land reclamation

The alluvial lowlands are covered with coconut groves and rice fields. A unique feature of agriculture in the Kutanad area of Travancore is the annual reclamation of land from deep lakes and backwaters for rice cultivation. The midland region between the coast and the plateau is rich in a variety of crops such as rice, tapioca, and millet, and in palm, mango, jackfruit, and cashew trees. The highlands are rocky in the higher reaches, but at lower levels there are extensive plantations of pepper, rubber, coffee, tea, and cardamom. Primeval forests cover the mountainous belt and are inhabited by a variety of wildlife including elephants, tigers, panthers, bison, and cobras. The Periyār Wildlife Sanctuary is a tourist attraction.

The people. The racial distribution of the state's population is representative of the subcontinent's history. Several

of the hill tribes exhibit Negroid features, while others appear to be related to the Veddas, the forest nomads of Sri Lanka (Ceylon). These racial types occur only in southern India and are probably remnants of the region's earliest peoples. The bulk of Kerala's inhabitants, however, are of the Dravidian stock. The Nairs and Ezhavas, who form the leading Hindu communities, belong to the Dravidian group of the early Cēra kingdom, while the Nambudiri Brahmins are descendants of the later Aryan migrants from the north. The Malayālis—Malayalam-speaking people—are a product of race admixture between the predominant Dravidian and the smaller Aryan groups.

Kerala is among the most densely populated states in India. The proportion of urban population is about 16 percent, but this low figure is deceptive because the villages in Kerala are semiurban and form a continuous chain. The major industrial complexes are at Alwaye, Quilon, Alleppey, Calicut, and Cannanore. Other demographic features reflect the status of a burgeoning population.

Religious harmony

Kerala has a unique record in India of harmonious coexistence of diverse religions. The majority of the Malayālis belong to the Hindu community, which has absorbed Buddhists and Jains into its fold. There is no conflict between the Dravidian *nāga* (serpent-god) worship and that of Kālī (the mother goddess) on the one hand and the Hindu pantheon on the other, nor between the rival sects of Śaivism (worship of the god Śiva) and Vaiṣṇavism (worship of the god Vishnu). The concept of unity in diversity is exemplified in the various Kerala temples by giving all these deities their allotted places. The Jewish community remains a small, exclusive sect. The Christians, forming about a fourth of the population, belong broadly to the Orthodox Syrian, Roman Catholic, and Protestant churches, but the denominations among them are no less numerous than the subcastes among Hindus. The Muslim minority have lived in habitual amity with the Hindus.

The economy. The economy is largely circumscribed by geographic and geologic constraints. The amount of arable land is far below the needs of the crowded population. The state is devoid of major deposits of minerals and fossil fuels, except for such minerals as ilmenite (the principal ore of titanium), rutile (titanium dioxide, used for coating welding rods), and monazite (a mineral composed of cerium and thorium phosphates), which are found in the beach sands. It has, however, enormous hydroelectric potential in its rapid rivers. The educational system, a developed banking system, and excellent transport facilities provide conditions favouring further economic development.

Agriculture is the state's main economic activity. Cash crops, which account for less than half of the total land under cultivation, earn a sizable amount of foreign exchange but have also resulted in a shortage of food for local consumption. The principal cash crops are rubber, cardamom, pepper, coconut, tea, and coffee, and the major food crops are rice, pulses, and jowar (sorghum). Though livestock is of poor quality, commercial poultry farming is the best developed in India, and eggs are exported. The forests yield industrial raw materials such as bamboo (used in the rayon industry), wood pulp, charcoal, gums, and resins, as well as ivory and rare timbers such as teak and rosewood. Kerala ranks first in fish production in India; fishing trawlers have been bought, and the industry has been mechanized with Norwegian assistance.

Most of the population is unaffected by industrialization. The traditional low-wage cottage industries, such as the processing of coconut fibre and cashews or weaving, employ most workers. Many of those employed by larger industrial enterprises are engaged in food processing. The state participates in most major industrial concerns; products include fertilizers, chemicals, electrical equipment, titanium, aluminum cables, plywood, ceramics, and synthetic fabrics. More than one-fourth of the workers are employed in providing services. Unemployment is acute, and the problem is accentuated by a high level of education among the jobless.

In proportion to its area the state has the most developed road system and traffic flow in India; it is connected with the states of Tamil Nādu, Karnātaka, and Mahārāshtra by the national highway and by a west coast road. The railway stretches from the northern border near Mangalore down to Trivandrum, connecting the principal towns. There are three major ports—Cochin, Alleppey, and Calicut—which handle both coastal and foreign traffic, and several minor ports. More than 1,100 miles of inland waterways constitute the main arteries of commerce and carry the bulk of freight from and to the ports. Trivandrum has an international airport and Cochin has one for domestic flights.

Administrative and social conditions. *Government.* The administration conforms to the uniform pattern adopted by states in the Indian union, with a governor at the head, a singlechamber legislature, a cabinet executive presided over by a chief minister, and an independent judiciary. The legislative assembly has a membership of 134. The structure of the judiciary is hierarchical, with a high court at the apex from which appeals to the Supreme Court of India may be made. For provincial administration the state is divided into districts that are subdivided for revenue purposes into *talukas* and villages.

Kerala's modern political experience has been one of unstable government, with popular ministries rising and falling in quick succession. The proliferation of political parties has made coalition governments inevitable.

Social issues. The Aryan-Dravidian ethnic synthesis created a caste-ridden society characterized by peculiar social customs, manners, and institutions. The matrilineal joint family has given way to nuclear families with paternal inheritance. Women in Kerala enjoy an honoured place in society. Both polygamy and polyandry were widely prevalent but today are illegal except for Muslim polygamy. The simple marriage ceremony of the Nairs consists merely of the presentation of cloth.

Education and welfare. In education, Kerala is one of the most advanced states in India. The level of literacy is substantially higher than the national average. Elementary education is free and compulsory between the ages of 6 and 11. There are primary, basic, and secondary schools; polytechnical and industrial training institutes; arts and science colleges; professional colleges; and four universities. The state maintains a relatively high standard of health service. There is a comprehensive health insurance plan for factory workers and free medical treatment in many hospitals, health centres, and dispensaries.

A high level of education

Cultural life. Kerala's cultural heritage contains elements of ancient Hindu culture that have been enriched by centuries of contact with both East and West. Literature and learning, in both Tamil and Sanskrit, flourished from the 2nd century AD onward. The Malayālam language, though an offshoot of Tamil, has absorbed the best in Sanskrit and has a prolific literature. Notable names in Malayālam poetry are Tuñchattu Eḻuttaccan and Kuñcan Nampiyār among classical poets, and Vallathol and Kumaran Asan in the present century.

Dance and drama

Dramatic productions in the Kerala style, which combines poetry, music, and dance, have entered folk tradition. The themes are usually stories from the great Indian epics. The Kalamandalam, a recently established cultural institution, seeks to revive and promote traditional arts. A display of various visual arts takes place during Ōṇam, the major festival of Kerala. Ōṇam is also the occasion for a snake-boat race at Aranmulai, in which hundreds of participants in each boat row rhythmically to the accompaniment of music and dancing.

In the fine arts, Swati Thirunal, the 19th-century royal composer, made signal contributions to classical Carnatic music of South India, while the artist Ravi Varma was renowned for his oil paintings. The splendid murals that adorn the walls of palaces, temples, and churches reveal a tradition and style that is distinct from classical Indian painting. Wood carving has attained a high level of excellence. The "Malabar gable"—a triangular projection at the top extremities of tiled or thatched roofs—is the characteristic feature of the Kerala style of architecture, which is noted for its beauty and utilitarian simplicity.

HISTORY

Kerala is first mentioned on a 3rd-century BC rock inscription left by the emperor Aśoka. It was then the independent kingdom of Kerlaputra and was ruled by the Cēra dynasty until the 5th century AD. It traded with

Egypt, Babylonia, Greece, Rome, and China. In this period Aryans migrated into Kerala and introduced Hinduism, Buddhism, and Jainism into this Dravidian kingdom. It is believed that St. Thomas the Apostle founded the Syrian Church—and Jewish immigrants established a colony—at Muziris (Cranganore) in the 1st century AD.

The history of Kerala from the sixth to eighth centuries AD is obscure. It was probably during this period that Arab traders introduced Islām. From AD 825—the year in which the Kollam era, regarded as still in progress, began—Kerala rose to prominence under the Kulasekhara dynasty. The ninth and tenth centuries formed a golden age marked by a remarkable development in philosophy, literature, education, and the fine arts. This period marks the origin of Malayālam as a separate language.

The Malabār era

During the 11th century a hundred years' war with the Cōḷas led to the disintegration of the Cēra kingdom into innumerable petty principalities. By the beginning of the 14th century, however, the ruler Ravi Varma Kulasekhara established a short-lived supremacy over the whole of southern India. His empire fell to pieces, however, soon after his untimely death in 1314, after which Kerala became a conglomeration of warring chieftaincies, among which the most important were Calicut in the north and Travancore in the south.

In 1498—when Vasco da Gama landed at Calicut—the era of foreign intervention began. In the 16th century the Portuguese superseded the Arab traders and dominated the commerce of Malabār. Their attempt to establish sovereignty was thwarted by the *zamorin* of Calicut. The Dutch ousted the Portuguese in the 17th century but were defeated by King Martanda Varma of Travancore in 1741. The invasions of Kerala between 1766 and 1790 by Hyder Ali and Tippu Sultan, both of Mysore, paved the way for the establishment of British power. Cochin and Travancore were obliged to accept British suzerainty (1791 and 1795, respectively); the *zamorin*'s territories were annexed by the British and became the Malabār district of the Madras presidency under their direct rule. Unsuccessful revolts against British rule were led by Pazhassi Raja of Malabār from 1800 to 1805 and by Velu Thampi of Travancore in 1809. Two years after India's independence was achieved in 1947, the two states were united as the Travancore-Cochin state. In 1956 the present state of Kerala was constituted on a linguistic basis, thus also including other Malayālam-speaking areas in the Malabār region. (V.R.P.)

Lakshadweep

Lakshadweep, a union territory of India since 1956 and formerly called, until 1973, the Laccadive, Minicoy, and Amindīvī Islands, is a group of islands lying scattered over 30,000 square miles of sea, about 185 miles (300 kilometres) off the western coast of the Indian state of Kerala. The name Lakshadweep, or Laksha Dweep, means "hundred thousand islands." There are 20 separate coral reefs and 14 islands, with a total area of 12 square miles (32 square kilometres). Ten of the islands are inhabited. The group once was attached to Madras state of India, the northern portion, known as the Amindīvī Islands, being within the South Kanara district and the remainder within the Malabār district. The administrative centre is Kavaratti Island in the Laccadive group.

PHYSICAL AND HUMAN GEOGRAPHY

The Aminidīvī are the northern islands of the group. Minicoy (Miniḳoi) lies between the Laccadives and the Maldives. Munyal, Coradive, and Sesostris are submerged banks lying north of the islands. Most of them have emerged from the eastern and protected side of the reef, the western being completely exposed to the southwest monsoon. The islands are small, none exceeding a mile in breadth, and much of the total area is occupied by coconut plantations. The soil is light coral sand, beneath which, a few feet down, lies a stratum of coral stretching over the whole of the islands, generally a foot to a foot and a half in thickness. It has been wholly excavated in the principal islands, the underlying damp sand being made available for growing cereals. These excavations, a work of vast labour, were made at a remote period—according to the native tradition, by giants.

Coarse grains, pulses, bananas, and vegetables are cultivated; coconuts grow abundantly and occupy most of the remainder of the land. For rice the inhabitants must depend upon the mainland.

The people are Moplahs—that is, of mixed Hindu and Arab descent—and are Muslims. The language spoken is Malayālam (but written in the Arabic character), except in Minicoy, where Mahl (akin to old Sinhalese) is spoken. The chief industry is fishing; there is a cannery at Minicoy island. The manufacture of coir (coconut husk fibre), largely employing women, is the major cottage industry. The islanders also are occupied with boatbuilding and trade with the mainland. Because the frequent coral reefs in the vicinity of the islands make navigation dangerous, commerce is carried on in native vessels manned by skilled sailors. There is a college affiliated with the Calicut University in Kerala. The jurisdiction of the Kerala High Court extends over Lakshadweep.

HISTORY

The Arab scholar al-Bīrūn (*c.* 1030) divided the whole archipelago (Dibajat) into the Divah Kuzah, or Cowrie Islands (the Maldives), and the Divah Kanbar, or Coir Islands (the Laccadives). The islanders were converted to Islām by an Arab apostle name Mumba Mulyaka, whose grave at Androth still gives a peculiar sanctity to that island. The kazi of Androth was in 1847 still a member of his family and was said to be the 22nd who had held the office in direct line from the saint.

The Portuguese first sighted the Laccadives in May 1498 and built forts upon them, but in about 1545 the local inhabitants rose against them. The islands subsequently became a suzerainty of the raja of Cannanore, and after the peace of Seringapatam in 1792 the southern group was permitted to remain under the management of the Moplah chief at a yearly tribute. Since this payment was often in arrears, the islands were sequestrated by the British in 1877. Sovereignty was transferred to India upon Indian independence in 1947. The islands were constituted a union territory in 1956.

Pondicherry

Pondicherry is a union territory formed in 1962 out of the four former colonies of French India: Pondicherry, Kārikāl, and Yanam, scattered along the Coromandel Coast, or eastern seaboard, of India; and Mahe, lying on the other side of India on the Malabār Coast, or western seaboard. The total area of the combined territory is 190 square miles (492 square kilometres). The capital is the town of Pondicherry.

PHYSICAL AND HUMAN GEOGRAPHY

All four areas of the territory are seaside tourist resorts. The town of Pondicherry is divided into two parts by a canal, and all the main streets, running parallel to one another, lead to the open roadstead offshore. The port of Pondicherry does not have a harbour, and ships are forced to lie a mile or two offshore; but its roadstead was once considered the best on the Coromandel Coast. There is a promenade and landing place for cargo, and in the 1960s a new pier was constructed. In and around the town are artesian wells that supply a large quantity of water for irrigation, the chief local crops being paddy, sugarcane, cotton, and peanuts (groundnuts). The main industries are food processing, electrical appliances, textiles, paper, and lumber. The Pondicherry area has about 300 villages and hamlets.

The commune of Kārikāl is in the fertile Thanjavūr delta, in one of the most important rice-producing areas of India. The exceptional fertility of the region is to some extent reflected in the unusually high density of its rural population. The town is on a branch line, the Mayavaram–Peralam route, of the southern railway.

The Mahe sector consists of two parts: the quaint, picturesque town of Mahe, with all its buildings situated on

the left bank of the Mahe River close to its mouth; and the isolated tract known as Naluthrara, on the right bank, comprising the four villages of Chambara, Chalakara, Palour, and Pandaquel. Rice is the chief crop.

Yanam is a small town on the bank of a branch of the Godāvari River, about 400 miles (650 kilometres) north of the city of Madras, near Kākināda.

The major languages spoken in the areas are Tamil, Malayālam, and Telugu. Tamil is predominant in the southern settlements of Pondicherry and Kārikāl; Malayālam is predominant in Mahe, and Telugu in Yanam. Other significant languages include Urdū, French, Kannaḍa, Hindī, Gujarātī, English, and Marāṭhi. Hindus form the majority in all the four regions; Muslims are an important minority in Kārikāl, Mahe, and Yanam; and Christians are numerous in Pondicherry. There are also a few Sikhs, Buddhists, and Jainas.

There are no heavy industries or mining in the union territory; it purchases its entire power requirement from nearby states. Pondicherry is governed by a lieutenant governor who is advised by a chief minister and a council of ministers. The jurisdiction of the Madras High Court extends over the union territory.

Pondicherry contains the Hindu *āśrama* (religious retreat) of Sri Aurobindo, and Auroville—the international township and study centre. The Romain Rolland Public Library houses some rare French volumes. A medical college, a law college, and several other colleges for general education are affiliated with the University of Madras.

HISTORY

French India was first founded by the French East India Company (formed by Jean-Baptiste Colbert in 1666), which established a settlement at Surat in 1668. This was followed, five years later, by another at Pondicherry, whose director, François Martin, made it the capital of the French posts. Mahe was founded in 1725, Yanam in 1731, and Kārikāl in 1739. French concerns multiplied in Bengal, with Chandernagore as centre, especially after 1730 under the direction of Joseph-François Dupleix, who in 1742 was appointed general director.

From 1763 the French establishments in India, which were under the authority of the king after the abolition of the company in 1769, comprised—apart from a few small posts (*loges*)—no more than five settlements of moderate size: Chandernagore in Bengal; Yanam, Pondicherry, and Kārikāl on the Coromandel Coast; and Mahe on the Malabar Coast. The English conquest of India lessened the commercial activity of the settlements. They were occupied by the English in 1778 and then again in 1793 but were returned to France in 1816. The Second Republic of France gave them local government and representation in the French parliament. Under the Second Empire, commercial liberalism and Anglo-French understanding gave these settlements a moment of passing prosperity.

In 1947 the *loges* were given back to independent India. Chandernagore was finally transferred in 1951. De facto transfer of the four remaining French possessions to the Union of India took place on November 1, 1954, and de jure transfer on May 28, 1956. Instruments of ratification were signed on August 16, 1962, from which date Pondicherry, consisting of the four enclaves, became a union territory. (Ed.)

Tamil Nādu

One of the southern states of India, Tamil Nādu (Land of Tamil) is located in the extreme southwest of the Indian peninsula. Representing the Tamil-speaking area of what was formerly Madras Presidency, it has an area of 50,215 square miles (130,057 square kilometres). It is bounded by the Indian Ocean to the east and south, by the states of Kerala to the west, Karnātaka (formerly Mysore) to the northwest, and Andhra Pradesh to the north. Its capital is Madras.

While it has an industrial core in Madras, the state is essentially agricultural. Proud of their Dravidian language and culture, the Tamils have resisted attempts by the union government to make Hindī the national language.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The state is divided naturally between the flat country along the eastern coast and the hilly regions in the north and west. The broadest part of the eastern plains is the fertile Cauvery Delta; farther south are the arid plains of Rāmanāthapuram and Madurai. The high Western Ghāts range runs all along the state's western border. The lower hills of the Eastern Ghāts and its outliers—locally called the Javādis, Kalrāyans, and Shevaroys—run through the centre of the region.

The rivers

The important rivers include the Cauvery, the Ponnaiyār, the Pālār, the Vaigai, and the Tāmbraparni, all of which flow eastward from the inland hills. The Cauvery and its tributaries are Tamil Nādu's most important sources of water and power.

Apart from the rich alluvial soil of the river deltas, the predominant soils are clays, loams, sands, and red laterites (soils with a high content of iron oxides and aluminum hydroxide). Black cotton soil is found in parts of Salem and Coimbatore in the west, Rāmanāthapuram and Tirunelveli in the south, and Tiruchchirāppalli in the central region of Tamil Nādu.

The climate is essentially tropical. The temperature in summer seldom exceeds 109° F (43° C) and in winter seldom falls below 64° F (18° C). The lowest temperatures are recorded during December and January, and the highest in April to June. The average annual rainfall, falling mainly between October and December, depends on the southwest and northeast monsoons and ranges between 25 and 75 inches a year. The most precipitation falls in the Nīlgiris and other hill areas, the lowest in the Rāmanāthapuram and Tirunelveli districts.

Forests cover about 15 percent of the state. At the highest altitudes of the Western Ghāts—the Nīlgiri, Anamalai, and Palni Hills—the mountains support subalpine vegetation. Along the eastern side of the Western Ghāts and in the hills of the northern and central districts, the plant life is a mixture of evergreen and deciduous plants, some of which are markedly adapted to arid conditions. Timber products extracted from the forests include sandalwood, pulpwood, and bamboo. Rubber is also an important forest product. The state's aquatic birds are represented in the bird sanctuary at Vedantangal; other forms of wildlife may be seen in the game sanctuary at Mudumalai.

The people. The present population has changed little over the centuries. Population characteristics include short to medium stature, dark to brown skin colour, and wavy hair. Dark pigmentation is more general in the southern districts and among the lower castes and tribes. Most of the hill tribes are of Veddoid origin. The caste system is still strong, though untouchability has been banned by law.

Tamil, the official state language, is spoken by most of the people. For a considerable number of the population long resident in the state, Tamil has almost become a mother tongue. Telugu is spoken by almost 10 percent of the population; Kannaḍa, Urdū, and Malayālam are spoken by much smaller percentages. In the Nīlgiris district in the west, Kannaḍa (and its dialect Baḍaga) and Malayālam are stronger. English is spoken as a subsidiary language.

The main religions are Hinduism, Christianity, Islām, and Jainism. Followers of the first three are found in all districts, but Jainas are confined to North and South Arcot and Madras city. Hindus constitute the overwhelming majority of the population. The largest concentration of Christians is in the Tirunelveli and Kanniyākumāri districts. The growth of atheism is a recent development, possibly as a protest against Brahmin ritualism.

Though Tamil Nādu is one of the most urbanized states of India, it is still a rural land. Most of the people live in more than 64,000 nucleated villages. The poorest low-caste villagers live in segregated areas called *cēri*. The Madras metropolitan conurbation, covering the industrial areas, townships, and villages surrounding Madras city, has the largest population, but there are other conurbations, of which those around Madurai, Coimbatore, and Tiruchchirāppalli are the most important.

The economy. Agriculture is the mainstay of life for about three-quarters of the rural population. From very

The irrigation system

early times Tamil farmers have skillfully conserved scarce rainwater in small and large irrigation reservoirs, or "tanks." Government canals, tube wells, and ordinary wells also form part of the irrigation system. Because several of the river valley projects depend for water on the erratic northeast monsoon, the government is trying to tap subsoil water sources.

Agricultural practices have undergone radical improvement since 1950. Multiple cropping, the use of new and better strains of rice, cotton, sugar, and millet, and the use of chemical fertilizers have been widely adopted. By 1967 the state was self-sufficient in the production of food grains. Rice, peanuts (groundnuts), cotton, chillies, bananas, coffee, tea, rubber, and sugarcane are important cash crops.

Improved port facilities and the effective use of electric power resources have helped industrial development. The state is one of the most industrialized of the Indian states. The important minerals are limestone, bauxite, gypsum, lignite, magnesite, and iron ore. Cotton ginning, spinning, and weaving continue to be the major industries, followed by the production of automobiles, motorcycles, transformers, sugar, agricultural implements, fertilizers, cement, paper, chemicals, and electric motors. The railway-coach factory at Perambur is one of the largest in Asia; the Heavy Vehicles Factory, producing tanks, is at Āvadi, near Madras. There is an oil refinery at Madras and a larger thermal power project at Neyveli; both are public sector ventures. The state ranks second only to Kerala in the production of fish.

Tamil Nādu is rich in handicrafts; notable among them are handloomed silk, metal icons, leather work, *Kalamkari* (hand painted fabric, using natural dyes), brass, bronze, and copper wares, and carved wood, palm leaf, and cane articles.

The transport system of the southern Indian states converges on Madras. Many railways run through Tamil Nādu. The artificial harbour of Madras handles seaborne traffic. Apart from the international airport at Meenāmbakkam near Madras, there are three other airports at Tiruchchirāppalli, Madurai, and Coimbatore. There is a network of motorable roads. Passenger bus transport is being nationalized; express buses take passengers to all important towns and places of interest.

Administrative and social conditions. *Government.* The governor, the legislative assembly, the legislative council, and the chief minister and his council of ministers together constitute the legislative and executive wings of the state government. The ministries are housed in Ft. St. George in Madras, though the offices of several heads of departments are in multistoried buildings outside the fort.

The state has 16 administrative districts, each administered by a district collector. Lower administrative units are divisions called *taluka, firka,* and villages. All of these units are responsible to the Revenue Department and the Board of Revenue. After independence, new units—*pañcāyat*s (village councils)—were established for purposes of local self-government and rural development. Above the *pañcāyat* there are *pañcāyat* unions and development councils. Villagers get their first lessons in democracy and administration in the *pañcāyat*s.

The judiciary is headed by the High Court at Madras; there are district judges and magistrates at the lower levels.

Health. Cholera, malaria, and filariasis (disease caused by infestation of the blood and tissues by parasitic worms) are the chief endemic diseases. Urban sanitation and drainage are substandard. There are a large number of public and private hospitals, dispensaries, and primary health centres in the state.

Education. The literacy rate is about 45 percent. There are primary and middle schools, high schools, and arts and science colleges, as well as medical colleges, engineering colleges, polytechnic institutes, and industrial training institutes. There are universities located at Madras, Chidambaram, Coimbatore, and Madurai. The Dakshina Bharat Hindī Prachār Sabhā (1918) at Madras, and the Gāndhīgrām Rural Institute (1956) at Gāndhīgrām, are the two institutes of national importance engaged in popularizing the Hindī language, and Mahatma Gandhi's concept of rural higher education, respectively. A vigorous effort is being made to make Tamil instead of English the medium of instruction at the university level.

Welfare. The lowest castes, called Harijan (Children of God), continue to be the most miserable section of the Hindu population. The Department of Harijan Welfare is in charge of programs to improve their housing, education, and economy. Government positions are reserved for them, and there are also reserved constituencies for them for elections to the state legislature and Parliament. The State Social Welfare Board and the Department of Women's Welfare are concerned with the problems of the handicapped, widows, and children. The state government provides free midday meals for children at the elementary schools.

Cultural life. *The cultural milieu.* Despite attempts to break the religious and political power of the priestly Brahmin caste, Hinduism remains at the core of the culture. The more than 9,300 larger temples come under the administrative control of the Hindu Religious and Charitable Endowments Department. In most towns—and particularly in Chidambaram, Kānchipuram, Thanjāvūr (Tanjore), and Madurai—the *gopuram*s, or gateway towers, of the temples are dominant. The cycle of temple festivals attracts large congregations of devotees. Noteworthy are the car festivals during which large chariots decorated with icons are taken in procession around the temple. Hindu families owe allegiance to a number of sectarian monastic institutions, or *maṭh*s, of which the most important is the Śaṅkara Maṭh at Kumbakonam.

The cultural role of Hinduism

The arts. Classical *bhārata-nāṭya* dancing and Carnatic music are widely practiced. Painting and sculpture, however, are less developed, although there are schools that teach the art of sculpture in stone and bronze. Tamil literature rapidly adapted to the Western literary forms of the novel and the short story. The poet Subrahmanya Bharati (who died in 1922) was one of the first to modify traditional Tamil poetry. Since the 1940s, the motion picture has become the most popular form of mass entertainment. There are both touring and permanent cinema theatres; and sentimental and spectacular films, often featuring light music and dancing, are produced by the film studios situated largely around Madras.

The press. There are hundreds of periodicals published in Tamil, of which most are daily newspapers. The *Dina Thanthi* is the leading paper. Among English newspapers, *The Hindu* of Madras maintains excellent journalistic standards.

HISTORY

The history of Tamil Nādu begins with the establishment of a trinity of Tamil powers in the region—the Cēra, Cōḷa, and Pāṇḍya kingdoms. By about AD 200 the influence of northern Aryan powers had progressed, and the Aryan sage Agastya had established himself as a cultural hero. The use of Roman gold and lamps and the consumption of Italian wine attest to the extensive foreign trade of the period. From the mid-6th century until the 9th century, the Cālukyas of Bādāmi, the Pallavas of Kāñcī, and the Pāṇḍyas of Madurai fought a long series of wars in the region. The period, nonetheless, was marked by a revival of Hinduism and the advance of the fine arts. From about AD 850, Tamil Nādu was dominated by the Cōḷas, of whom Rājendra I (reigned 1014–44) was the most distinguished ruler. In the mid-14th century the Hindu kingdom of Vijayanagar, including Tamil Nādu, came into prominence. During the 300 years of Vijayanagar rule, Telugu-speaking governors and officials were introduced in the administration.

The three Tamil kingdoms

In 1639 the British East Indian Company opened a trading post at the fishing village of Madraspatnam (now Madras) with the permission of the local ruler. The history of Tamil Nādu from the mid-17th century to 1946 is the story of the Madras Presidency in relationship to the rise and fall of British power in India. After 1946 the Madras Presidency was able to make steady progress, as it had a stable government. In 1953 the Telugu-speaking state of Andhra Pradesh was formed, and in 1956 the presidency was further divided into the states of Kerala, Mysore (now Karnātaka), and Tamil Nadu. (A.Ai.)

BIBLIOGRAPHY

Physical and human geography. *Land and people:* The following titles, though produced largely by government sources, contain much useful information and statistical data: S.N. AGARWALA, *Population* (1967); M.R. CHAUDHURI, *Economic Geography* (1969); *The Wealth of India: A Dictionary of Indian Raw Materials and Industrial Products,* 8 vol. to date, issued by the INDIAN COUNCIL OF SCIENTIFIC AND INDUSTRIAL RESEARCH (1948–69); *India: A Physical Geography* (1968); and *The Gazetteer of India,* vol.1 (1965), both issued by the PUBLICATIONS DIVISION OF THE MINISTRY OF INFORMATION AND BROADCASTING of the Government of India; S.L. OGALE, *The Tragedy of Too Many* (1969); *The Census of India* (1951, 1961, 1971, 1981) by the Registrar General's Census Commissioner for India. Two useful books are D.N. WADIA, *Geology of India,* 3rd ed. rev. (1957); and (ed.), *Minerals of India* (1966). See also R.L. SINGH (ed.), *India: A Regional Geography* (1971).

Government: There has been a spate of books on government since the constitution came into force in January 1950. The following are among the more informative and authoritative books: GRANVILLE AUSTIN, *The Indian Constitution: Cornerstone of a Nation* (1966); P.B. GAJENDRAGADKAR, *The Constitution of India* (1969); ALAN GLEDHILL, *Fundamental Rights of India* (1955); SIR MAURICE GWYER and ANGADIPURAM APPADORAI, *Speeches and Documents on the Indian Constitution, 1921–1947,* 2 vol. (1957); ROBERT L. HARDGRAVE, *India—Government and Politics in a Developing Nation* (1975); M. HIDAYATULLAH, *Parliamentary Privileges: The Press and the Judiciary* (1969); SIR BENEGAL NARSINGA RAU, *India's Constitution in the Making,* ed. by B. SHIVA RAO, rev. ed. (1963); B. SHIVA RAO, *Framing India's Constitution,* 5 vol. (Indian Institute of Public Administration, 1966–68); R.J. VENKATESWARAN, *Cabinet Government in India* (1967).

Legislature: Among the better known books on the legislature are the *Reports on the General Elections in India* published since 1955 (on the 1951–52 elections) by the Election Commission in India; RAJNI KOTHARI *et al., Party System and Election Studies* (1967); W.H. MORRIS-JONES, *Parliament in India* (1957); K.C. RAY, *The Law of Elections in India* (1969); and MYRON WEINER and RAJNI KOTHARI (eds.), *Indian Voting Behaviour* (1965).

Executive and judiciary: Among the better known and more useful sources that may be consulted are PAUL H. APPLEBY, *Public Administration in India* (1953); ASOK CHANDA, *Indian Administration* (1958); A.D. GORWALA, *Report on Public Administration* (1953); N.C. ROY, *The Civil Service in India,* 2nd ed. (1960); GHOPAL DAS KHOSLA, *Our Judicial System* (1949); and R.G. CHATURVEDI, *Judiciary Under Constitution* (1967).

Defense: Authoritative works on this subject are few. These include *Defence and Security in the Indian Ocean Area* by the INDIAN COUNCIL OF WORLD AFFAIRS (1958); LORNE J. KAVIC, *India's Quest for Security* (1967); RAVI KAUL, *India's Strategic Spectrum* (1969); S.S. KHERA, *India's Defence Problem* (1968); T.S. MURTY, *Paths of Peace: Studies on the Sino-Indian Border Dispute* (1983); H.M. PATEL, *The Defence of India* (1963); GAUTAM SEN, *India's Defense Policy: An Analysis, 1947–81* (1983); and K.M. PANIKKAR, *Problems of Indian Defence* (1960).

Education: Education is a popular topic that has been subjected to wide scrutiny. The official government publications cover an extensive field. A valuable report is one on *Indian Universities* by C.P. RAMASWAMY AIYAR (1964). The University Grants Commission, the Institute of Constitutional and Parliamentary Studies, and the Inter-University boards of India and Ceylon have produced various reports and pamphlets. The Ministry of Education has also issued several reports and reference books.

Cultural activities: Some of the best known books on India's cultural activities have been written by foreigners, mainly English and American. Those worth consulting include KAY AMBROSE, *Classical Dances and Customs of India* (1950, reprinted 1965); and the two volumes by PERCY BROWN, *Indian Painting,* 7th ed. (1960) and *Indian Architecture,* 5th ed., 2 vol. (1965–68). KAMALADEVI CHATTOPADHYAYA, *Carpets and Floor Coverings of India* (1969), is a standard work. A.K. COOMARASWAMY, *The History of Indian and Indonesian Art* (1927, reprinted 1965), is regarded as a classic. C.B. GUPTA, *Indian Theatre* (1953); STELLA KRAMRISCH, *The Art of India* (1954); M.L. MATHUR, *Indian Miniatures* (1983); R. NATH, *Some Aspects of Mughal Architecture* (1976); and CALAMBUY SIVARAMAMURTI, *The Art of India* (1977), are all standard works.

Scientific research: The most authoritative sources on this subject consist mainly of government reports published by the Council of Scientific and Industrial Research and the Department of Atomic Energy. A good book on agriculture, though somewhat outdated, is M.S. RANDHAWA, *Agricultural Research in India* (1958).

Health and social welfare: There are several informed books on family planning. One of these is SRIPATI CHANDRASEKHAR, *Population and Planned Parenthood in India,* rev. ed. (1961). There are a number of official publications by the Ministry of Health and Family Welfare, including *Family Welfare Programme Yearbook.* A good book on social welfare is A.R. WADIA (ed.), *History and Philosophy of Social Work in India,* 2nd ed. (1968). There are also many works on rehabilitation and some books on the underprivileged classes. The two best known books on this subject are MAHATMA GANDHI, *The Removal of Untouchability,* ed. by BHARATAN KUMARAPPA (1954); and B.R. AMBEDKAR, *The Untouchables,* 2nd ed. (1969), by the well-known Harijan leader. An informative book on tribal life is G.S. GHURYE, *The Scheduled Tribes of India* (1980).

Press, radio, and films: There are several reports by various governmental commissions and committees on the press, broadcasting, and information. See the annual governmental report *Film in India.* Quite a few books are also available, both by Indian and foreign authors, on the freedom of the press in India. Among the best known is that of the Dutch author JOSEPH MINATTUR, *Freedom of the Press in India* (1961).

Economy: R.C. DUTT, *The Economic History of India,* 2 vol., 2nd ed. (1970); ASOKA MEHTA, *Economic Planning in India* (1970); J.N. BHAGWATI and P. DESAI, *India: Planning for Industrialization* (1979), industrialization and trade policies since 1951. Official government publications include *Gazetteer of India,* vol. 3: *Economic Structure and Activities* (1975); *Statistical Abstract of India* (annual); and *India: A Reference Annual. The Times of India Directory and Yearbook* is a useful source for current statistics. See also MOONIS RAZA *et al.* (eds.), *Sources of Economic and Social Statistics of India* (1979). An account of Indian agriculture is found in JASBIR SINGH, *An Agricultural Atlas of India: A Geographical Analysis* (1974).

History. *The emergence of civilization in North India:* The following are general surveys of the prehistory of the subcontinent of India: BRIDGET and RAYMOND ALLCHIN, *The Birth of Indian Civilization* (1968), with extensive bibliography and list of radiocarbon dates; DOUGLAS H. GORDON, *The Prehistoric Background of Indian Culture,* 2nd ed. (1960); D.D. KOSAMBI, *The Culture and Civilisation of Ancient India in Historical Outline* (1965), a somewhat idiosyncratic, synthetic account that sets out to integrate divergent types of evidence; H.D. SANKALIA, *Prehistory and Protohistory in India and Pakistan* (1962); BENDAPUDI SUBBARAO, *The Personality of India,* 2nd ed. (1958); SIR MORTIMER WHEELER, *Early India and Pakistan, to Ashoka,* rev. ed. (1968), a short and lucid survey by a former director general of archaeology in India; SHEREEN RATNAGAR, *Encounters: The Westerly Trade of the Harappa Civilization* (1981), a comprehensive work with an extensive bibliography on Indus Valley-Mesopotamian and Persian Gulf contacts; S.R. RAO, *Lothal and the Indus Civilization* (1973); J.C. SHAFFER, *Prehistoric Baluchistan* (1978); and G.L. POSSEHL (ed.), *Ancient Cities of the Indus* (1979), a collection of papers covering Indus civilization, archaeology, and history. General accounts of the Indus Civilization are SIR MORTIMER WHEELER, *The Indus Civilization,* 3rd ed. (1968), an up-to-date and comprehensive account; STUART PIGGOTT, *Prehistoric India to 1000 B.C.* (1950), still indispensable in spite of much recent work; WALTER A. FAIRSERVIS, "The Origin, Character, and Decline of an Early Civilization" (1967), a short pamphlet that includes much profound and original thought, and *The Roots of Ancient India* (1971); JEAN-MARIE CASAL, *La civilisation de l'Indus et ses énigmes* (1969); and D.P. AGRAWAL, *Copper Bronze Age in India* (1971).

The development of Indian civilization from c. 1500 BC to c. AD 1200–1300: On historiography, a much-neglected area of Indian history, the standard work is CYRIL H. PHILIPS (ed.), *Historians of India, Pakistan and Ceylon* (1961). FREDERICK E. PARGITER, *Ancient Indian Historical Tradition* (1922, reprinted 1962); V.S. PATHAK, *Ancient Historians of India* (1966); and ROMILA THAPAR, "Interpretations of Ancient Indian History," in *History and Theory,* 7:318–335 (1968), all deal with the various aspects of the historiography of ancient India. Among survey histories of the period may be mentioned, ARTHUR L. BASHAM, *The Wonder That Was India,* 3rd ed. (1967); D.D. KOSAMBI (*op. cit.*); and ROMILA THAPAR, *A History of India,* vol. 1 (1966). For political history, H.C. RAYCHAUDHURI, *The Political History of Ancient India,* 6th ed. rev. (1953), remains a standard work. More recent writing is included in the series *The History and Culture of the Indian People,* 11 vol. (1951–77). On economic history, U.N. GHOSHAL, *The Agrarian System in Ancient India* (1930), raises the relevant issues, as does ALINDRANATH BOSE, *Social and Rural Economy of Northern India, c. 600 B.C.–200 A.D.,* 2nd ed. rev., 2 vol. (1961–67). See also G.L. ADHYA, *Early Indian Economics: Studies in the Economic Life of Northern and Western India 200 B.C.* to *A.D. 300* (1966). A useful introduction to political ideas and

institutions is R.S. SHARMA, *Aspects of Political Ideas and Institutions in Ancient India,* 2nd rev. ed. (1968). (*1500–c. 600 BC*): N.R. BANERJEE, *The Iron Age in India* (1965), discusses iron technology in India. S.C. MALIK, *Indian Civilization: The Formative Period* (1968), raises the question of relating archaeological evidence to formulations on early Indian society. V. GORDON CHILDE, *The Aryans* (1926, reprinted 1970), remains a standard work on the origin and dispersal of the Aryans. Early views on this period are contained in VINCENT A. SMITH, *The Early History of India,* 4th ed. rev. (1924, reprinted 1957), but this is now considered somewhat out of date. kingship is discussed in N.N. LAW, *Aspects of Ancient Indian Polity* (1921). A more detailed discussion on political theories is found in U.N. GHOSHAL, *History of Hindu Political Theories,* 2nd ed. (1927). (*c. 600–150 BC*): Standard works on the Buddhist evidence are T.W. RHYS DAVIDS, *Buddhist India* (1903); B.C. LAW, *Some Kṣatriya Tribes of Ancient India* (1924) and *Geography of Early Buddhism* (1932); more recent studies include DEV RAJ CHANANA, *Slavery in Ancient India, as Depicted in Pali and Sanskrit Texts* (1960); and NARENDRA WAGLE, *Society at the Time of the Buddha* (1966). The interrelation between society and political ideas is discussed by CHARLES DREKMEIER in *Kingship and Community in Early India* (1962). On the Mauryan period, general works include K.A. NILAKANTA SASTRI (ed.), *The Age of Nandas and Mauryas* (1952); VINCENT A. SMITH, *Aśoka, the Buddhist Emperor of India,* 3rd ed. rev. (1920); and ROMILA THAPAR, *Aśoka and the Decline of the Mauryas* (1961). (*c. 150 BC–AD 300*): A.K. NARAIN, *The Indo-Greeks* (1957); J.E. VAN LOHUIZEN DE LEEUW, *The Scythian Period . . .* (1949); and ROMAN GIRSHMAN, *Bégram: Recherches archéologiques et historiques sur les Kouchans* (1946), concern the political history of this period in northern India. An in-depth and relatively recent monograph on the date of Kaniṣka is ARTHUR L. BASHAM (ed.), *Papers on the Date of Kaniska* (1968). GULAM YAZDANI (ed.), *The Early History of the Deccan,* 2 vol. (1960); P.T.S. AIYANGAR, *History of the Tamils, from the Earliest Times to 600 A.D.* (1929); N. SUBRAHMANIAN, *Sangam Polity* (1967); and K.A. NILAKANTA SASTRI (ed.), *A History of South India from Prehistoric Times to the Fall of Vijayanagar,* 3rd ed. (1966), deal with the southern parts of the subcontinent. (*AD 300–750*): Political histories include R.C. MAJUMDAR and A.S. ALTEKAR (eds.), *The Vākāṭaka-Gupta Age, Circa 200–500 A.D.* (1967); R.N. DANDEKAR, *A History of the Guptas* (1941); R.K. MOOKERJI, *The Gupta Empire,* 4th ed. (1969); EDWARD A. PIRES, *The Maukharis* (1934); B.P. SINHA, *The Decline of the Kingdom of Magadha (c. 455–1000 A.D.)* (1954); R.S. TRIPATHI, *History of Kanauj to the Moslem Conquest* (1937); D. DEVAHUTI, *Harsha* (1970); and D.C. SIRCAR, *The Successors of the Sātavāhanas in Lower Deccan* (1939). For discussions on administration and revenue, see D.N. JHA, *Revenue System in Post-Maurya and Gupta Times* (1967); V.R.R. DIKSHITAR, *The Gupta Polity* (1952); and U.N. GHOSHAL, *Contributions to the History of the Hindu Revenue System* (1929). Social and economic history is the main theme of R.N. SALETORE, *Life in the Gupta Age* (1943); S.K. MAITY, *Economic Life of Northern India in the Gupta Period, c. A.D. 300–550,* rev. 2nd ed. (1970); D.R. DAS, *Economic History of the Deccan, from the First to the Sixth Century A.D.* (1969). (*AD 750–1200*): There are a number of studies of political history relating to this period, some on regions and others on individual dynasties. Among them are A.S. ALTEKAR, *The Rāshṭrakūṭas and Their Times* (1934); A.C. BANERJEE, *Rajput Studies* (1944); H.C. RAY, *The Dynastic History of Northern India (Early Mediaeval Period)* (1931); R.C. MAJUMDAR, *The History of Bengal,* 2 vol. (1943–48); S.K. MITRA, *The Early Rulers of Khajuráho* (1958); N.S. BOSE, *History of the Chandellas of Jejakabhukti* (1956); DASHARATHA SHARMA, *Early Chauhán Dynasties* (1959); PRATIPAL BHATIA, *The Paramaras, c. 800–1305 A.D.* (1970); K.A. NILAKANTA SASTRI, *The Cōlas,* 2nd ed. rev. (1955); and J. DUNCAN DERRETT, *The Hoysalas, a Medieval Indian Royal Family* (1957). Related to political institutions is T.V. MAHALINGAM, *South Indian Polity,* 2nd rev. ed. (1967). Two major sources have been translated and make interesting reading: M.A. STEIN (trans.), *Kalhana's Chronicle of the Kings of Kashmir* (1892); and EDWARD C. SACHAU, *Alberuni's India,* 2 vol. (1888; reprinted in 1 vol., 1964). Trade and commerce in the context of Arab trade is discussed by GEORGE F. HOURANI in *Arab Seafaring in the Indian Ocean in Ancient and Early Medieval Times* (1951); the economic background is treated in LALLANJI GOPAL, *The Economic Life of Northern India, c. A.D. 700–1200* (1965); and feudalism in India is discussed in R.S. SHARMA, *Indian Feudalism, c. 300–1200* (1965).

The early Islāmic period and the Mughal Empire: The most detailed treatment is R.C. MAJUMDAR (ed.), *The History and Culture of the Indian People,* 2nd ed., vol. 5–6 (1966–67). Although the articles in this work offer scholarly treatments, they usually tend to present an Indian (as opposed to a Pakistani), and sometimes a Hindu nationalist point of view, and thus tend to emphasize the accomplishments of Hindu culture and of Hindu states, while they sometimes minimize those of the Muslims. For a very short treatment of the period that is slanted in the opposite direction, see I.H. QURESHI, *The Muslim Community of the Indo-Pakistan Sub-Continent, 610–1947* (1962). *The Cambridge History of India,* vol. 3 (1928, reprinted 1965); and *The Oxford History of India,* 4th ed. (1981), offer accounts that were written during the period of British rule and that reflect to some extent the British imperial point of view, although they are generally factually accurate. For a relatively short and good factual account written in India since independence, see R.C. MAJUMDAR, H.C. RAYCHAUDHURI, and K.K. DATTA, *An Advanced History of India,* 3rd ed. (1967); and *The Gazeteer of India,* vol. 2: *History and Culture* (1973), issued by India's MINISTRY OF INFORMATION AND BROADCASTING. Another history of the Muslims in India is S.M. IKRAM, *History of Muslim Civilisation in India and Pakistan* (1961). Good scholarly treatments of many aspects of the period that have to do with Muslims can be found in the *Encyclopaedia of Islam,* new ed. (1960–). A short but stimulating treatment from a Marxist point of view is contained in D.D. KOSAMBI, *An Introduction to the Study of Indian History* (1956). For a sample of the Persian chronicles that are the basic source for most of the history of the period, the reader may consult the translations in the appropriate volumes of HENRY M. ELLIOT and JOHN DOWSON, *The History of India, as Told by Its Own Historians,* 8 vol. (1867–77, reprinted 1966). (*The Delhi Sultanate*): Monographs of varying quality exist for each subperiod. The best is A.B.M. HABIBULLAH, *The Foundation of Muslim Rule in India,* 2nd rev. ed. (1961), which covers the pre-Khaljī period. K.S. LAL, *History of the Khaljis, A.D. 1290–1320,* rev. ed. (1967), provides an adequate political history and an interesting account of Alluddin Khaljī's economic reforms. AGHA MAHDI HUSAIN, *The Tughluq Dynasty* (1963); and K.S. LAL, *Twilight of the Sultanate* (1963), are less interesting although they provide detailed political narratives. PETER HARDY, *Historians of Medieval India* (1960), is a good historiographical treatment; WILLIAM H. MORELAND, *The Agrarian System of Moslem India* (1929), is one of the best accounts of the economic history. Political and administrative institutions are covered by R.P. TRIPATHI, *Some Aspects of Muslim Administration* (1936); S.P. NIGAM, *Nobility Under the Sultans of Delhi, A.D. 1206–1398* (1967); I.H. QURESHI, *The Administration of the Sultanate of Delhi* (1942); and K.A. NIZAMI, *Some Aspects of Religion and Politics During the Thirteenth Century in India* (1961). Nizami is also useful for social and religious history, and in many ways is one of the most absorbing books on the Sultanate. Finally, a most interesting and almost unique attempt to provide a social history of the common people is K.M. ASHRAF, *Life and Conditions of the People of Hindustan,* 2nd ed. (1970). (*The Bahmanīs and successor states*): Practically the only monographic work done on the Bahmanīs has been by H.K. SHERWANI. His writings provide a basic political history, and give a different view of the Bahmanī–Vijayanagar wars than the view presented in *History and Culture of the Indian People.* Otherwise, consult Sastri's *History of South India* or *The Cambridge History of India.* (*Vijayanagar*): Sastri's account again is useful, as are three monographs by T.V. MAHALINGAM: *Administration and Social Life Under Vijayanagar* (1940); *Economic Life in the Vijayanagar Empire* (1951); and *South Indian Polity,* 2nd rev. ed. (1967), although the last work provides a generalized account of a much longer period, and is difficult to disentangle. A very engrossing edition of the accounts of two European travellers who visited Vijayanagar is contained in ROBERT SEWELL, *A Forgotten Empire—Vijayanagar* (1900, reprinted 1924). (*The Mughal period*): In addition to the relevant sections of the Cambridge and Oxford histories mentioned above, the following works deal with various aspects of Mughal history: STEPHEN M. EDWARDES and H.L.D. GARRETT, *Mughal Rule in India* (1930, reprinted 1962); IRFAN HABIB, *The Agrarian System of Mughal India, 1556–1707* (1963); IBN HASAN, *The Central Structure of the Mughal Empire and Its Practical Working up to the Year 1657* (1936, reprinted 1970); ISHWARI PRASAD, *The Life and Times of Humāyūn* (1955); J.N. SARKAR, *History of Aurangzib,* 5 vol. (1912–24; 2nd ed. rev., 1925); *The Fall of the Mughal Empire,* 4 vol. (1932–50); and *Mughal Administration,* 4th ed. rev. (1952).

The Marāṭhās: The first major work on Marāṭhā history was JAMES C. DUFF, *History of the Mahrattas,* 3 vol. (1826). J.N. SARKAR made substantial contributions both to the translation and editing of a large number of Persian documents in English and producing first-rate original writings on the subject. Among them are: *Delhi During the Anarchy, 1749–1788, as Told in Contemporary Records* (1921); *Historical Records Relating to Northern India, 1700–1817* (1925); *Delhi Affairs, 1761–1788,* in *Persian Records of Maratha History,* ed. by P.M. JOSHI (1953); *Sindhia as Regent of Delhi, 1789–1791,* in

Persian Records of Maratha History (1954); *House of Shivaji: Studies and Documents on Maratha History,* 3rd ed. (1955); and *Daulat Rao Sindhia and North Indian Affairs, 1810–1818* (1951). P.M. JOSHI, a Marāṭhā historian, besides editing some of the works translated by J.N. Sarkar, also published translations in English of several Marāṭhā documents, including *Expansion of Maratha Power, 1707–1761: Selections from the Peshwa Daftar* (1956), and *Revival of Maratha Power, 1761–1772: Selections from the Peshwa Daftar* (1962). JAMES H. GENSE and D.R. BANAJI (eds.), *The Gaikwads of Baroda: English Documents,* 10 vol. (1937–45), give a great deal of information on the history of the princely state of Baroda (now Vadodara). RUSTOM D. CHOKSEY (ed.), *The Last Phase: Selections from the Deccan Commissioner's Files—Peshwa Daftar, 1815–1818* (1948), throws valuable light on the last few years of the Marāṭhā rule in India. The first scientific study of the forces—social, economic, and cultural, as well as political—that led to the building up of Marāṭhā power was M.G. RANADE, *Rise of the Maratha Power, and Other Essays* (1900, reprinted 1961), which was followed by a critical analysis by G.S. SARDESAI in his *Main Currents of Maratha History* (1926). Other studies include B.K. APTE, *A History of the Maratha Navy and Merchantships* (1973); V.G. DIGHE, *Bajirao I and Expansion of Maratha Power* (1944); P.C. GUPTA, *Bajirao II and the East India Company, 1796–1818,* rev. ed. (1964); V.V. JOSHI, *Clash of Three Empires* (1941); G.S. SARDESAI, *New History of the Marathas,* 3 vol. (1946–48); J.N. SARKAR, *Shivaji and His Times,* 6th ed. (1961); S.N. SEN, *Administrative System of the Marathas,* rev. ed. (1925) and *Military System of the Marathas* (1958); H.N. SINHA, *Rise of the Peshwas,* vol. 1 (1931); and S.P. VARMA, *A Study in Maratha Diplomacy: Anglo–Maratha Relations, 1772–1783* (1956).

India and European expansion, c. 1500–1858: (*Europeans in India, 1498–1760*): WILLIAM W. HUNTER, *A History of British India,* 2 vol. (1899–1900), vol. 1 is valuable for the Portuguese, and vol. 2 for the East India Company up to 1708; CHARLES R. BOXER, *The Portuguese Seaborne Empire, 1415–1825* (1969), an authoritative modern study, placing the Portuguese venture into India into its world perspective; and *The Dutch Seaborne Empire, 1600–1800* (1965), performs the same function for the Dutch connection with India; H.H. DODWELL, *Dupleix and Clive* (1920, reprinted 1967), a detailed and acute study of the Anglo-French struggle in South India and the British acquisition of Bengal. (*The growth of British India, 1757–1818*): ABDUL MAJEED KHAN, *The Transition in Bengal, 1756–1775* (1969), a detailed study of the transfer of administration from Indian to British hands in Bengal; KEITH FEILING, *Warren Hastings* (1954), a thorough and shrewd study of this enigmatic man; LUCY S. SUTHERLAND, *The East India Company in Eighteenth-Century Politics* (1952), a masterly survey of Indian politics at the British base in the Namier style; SIR C.P. ILBERT, *The Government of India,* 3rd ed. (1915), a constitutional lawyer's classic exposition of the development of the forms of the Government of India; ARTHUR ASPINALL, *Cornwallis in Bengal* (1931), a lucid description of the Cornwallis administrative revolution in Bengal; PAUL E. ROBERTS, *India Under Wellesley* (1929), a not wholly satisfactory but the best available study of this empire builder; EDWARD J. THOMPSON, *The Making of the Indian Princes* (1943), a lively account of the eclipse of the independent Indian powers by an imaginative writer turned historian. (*Organization and policy, 1818–57*): T.H. BEAGLEHOLE, *Thomas Munro and the Development of Administrative Policy in Madras, 1792–1818* (1966), a succinct study of the formulation of British policy in south India by a young scholar; KENNETH A. BALLHATCHET, *Social Policy and Social Change in Western India, 1817–1830* (1957), an authoritative study of Mountstuart Elphinstone's work of reconciling the Marāṭhās; PERCIVAL SPEAR, *The Twilight of the Mughals* (1951), deals with Charles Metcalfe's policy in the Delhi Territory and British relations with the Mughals; ERIC STOKES, *The English Utilitarians and India* (1959), an important work on the influence of British ideas in India that enlarges historical perspective; SOPHIA D. COLLET, *The Life and Letters of Raja Rammohun Roy,* 3rd ed. by D.K. BISWAS and P.C. GANGULI (1962), the standard biography of the Indian reformer updated by two scholarly disciples; KHUSHWANT SINGH, *A History of the Sikhs,* 2 vol. (1964–66), the best modern work on the Sikhs as a whole, by a Sikh; WILLIAM LEE-WARNER, *The Life of the Marquis of Dalhousie,* 2 vol. (1904), still the standard work on this key figure. (*The Indian Mutiny*): S.N. SEN, *Eighteen-Fifty-Seven* (1957), the best modern work on the Indian Mutiny by the most distinguished of recent Indian historians; MICHAEL MACLAGAN, "*Clemency*" *Canning* (1962), uses Canning's papers to describe his part in the Mutiny crisis.

British India, 1858–1920: The most comprehensive surveys of this period are in *The Cambridge History of India,* vol. 6; *The Oxford History of India;* SPEAR (1958); and *The History and Culture of the Indian People,* vol. 9, *British Paramountcy and Indian Renaissance,* 2 pt. (1963–65) and vol. 11, *Struggle for Freedom* (1969). The first two works relate the history of this period primarily from the British point of view, the last tells the same story from the vantage point of Indian nationalism. For primary sources of this period, see CYRIL H. PHILIPS (ed.), *The Evolution of India and Pakistan, 1858 to 1947* (1962). For the decade immediately after the Indian Mutiny of 1857–59, see THOMAS R. METCALF, *The Aftermath of Revolt: India, 1857–1870* (1964). A useful survey of most of this period is SARVEPALLI GOPAL, *British Policy in India, 1858–1905* (1965); and the best economic study is D.R. GADGIL, *The Industrial Evolution of India in Recent Times,* 4th ed. (1942). An excellent monograph on the Indigo revolt is BLAIR KLING, *The Blue Mutiny* (1966). For social history and Imperial "racial" problems, the best work is still E.M. FORSTER, *A Passage to India* (1924, reprinted 1981). More recent monographic studies are FRANCIS ROBINSON, *Separation Among Indian Muslims* (1975), a history of Muslim politics in the United Provinces from 1860 to 1923; ARVIND DAS, *Agrarian Unrest and Socio-economic Change: 1900–1980* (1983); CHRISTINE BOLT, *Victorian Attitudes to Race* (1971); and FRANCIS G. HUTCHINS, *The Illusion of Permanence* (1967). For the genesis of the Indian National Congress and its first meeting, see BRITON MARTIN, *New India, 1885* (1969). The official history of the Congress is B. PATTABHI SITARAMAYYA, *The History of the Indian National Congress,* 2 vol. (1946–47). A good early account of Indian nationalism by one of its leaders is SURENDRANATH BANERJEA, *A Nation in Making* (1925, reprinted 1963). For a study of early Congress factionalism, see STANLEY WOLPERT, *Tilak and Gokhale: Revolution and Reform in the Making of Modern India* (1962). A new study of Lord Curzon's viceroyalty is DAVID DILKS, *Curzon in India,* 2 vol. (1969–70). For Morley and Minto, see JOHN MORLEY, *Recollections,* 2 vol. (1917); MARY MINTO, *India, Minto and Morley, 1905–1910* (1934); and STANLEY WOLPERT, *Morley and India, 1906–1910* (1967). For early Muslim nationalism, see KHALID B. SAYEED, *Pakistan: The Formative Phase, 1857–1948,* 2nd ed. (1968); HAFEEZ MALIK, *Moslem Nationalism in India and Pakistan* (1963); and WILFRED C. SMITH, *Modern Islām in India,* rev. ed. (1946). An excellent new anthology on the "grandfather" of Pakistan is *Iqbal, Poet-Philosopher of Pakistan,* ed. by HAFEEZ MALIK (1971). The best biography of Pakistan's "Quaid-i-Azam" is still HECTOR BOLITHO, *Jinnah: Creator of Pakistan* (1954). For Gandhi, the best introduction remains his autobiography, *The Story of My Experiments with Truth,* 2 vol. (1927–29). A provocative psycho-historical study is ERIK ERIKSON, *Gandhi's Truth: On the Origins of Militant Nonviolence* (1969). For Motilal and Jawaharlal Nehru, see B.R. NANDA, *The Nehrus, Motilal and Jawaharlal* (1962). Nehru has also written a brilliant autobiography, *Toward Freedom: The Autobiography of Jawaharlal Nehru* (1941). LORD HARDINGE wrote a brief memoir of his viceroyalty, *My Indian Years, 1910–1916* (1948). Montagu has left a vivid account of his visit to India as secretary of state: EDWIN S. MONTAGU, *An Indian Diary* (1930). The Government of India published a short official account on *India's Contribution to the Great War* (1923). The Montagu–Chelmsford *Report on Indian Constitutional Reforms* (1918), is also worth reading. For the postwar massacre at Jallianwala Bāgh in Amritsar, see STANLEY WOLPERT, *An Error of Judgment* (1970).

Emerging nationalism, partition, and independence: PERCIVAL SPEAR, *India, Pakistan, and the West,* 4th ed. (1967), a broad survey of all aspects of modern life and thought; HUGH TINKER, *India and Pakistan: A Political Analysis,* rev. ed. (1967), covers political development at all levels. (*Pre-Independence, 1920–47*): L.S.S. O'MALLEY (ed.), *Modern India and the West* (1941, reprinted 1968), studies of many governmental and cultural trends—still very relevant; V.P. MENON, *The Transfer of Power in India* (1957), a detailed study of events leading to partition, drawing extensively upon confidential papers; HUGH TINKER, *Experiment with Freedom: India and Pakistan, 1947* (1967), a concentrated account of the last phase, 1945–47; R.C. GUPTA and J.C. JOHARI, *Indian Freedom Movement and Thought: 1930–1947* (1983); HENRY V. HODSON, *The Great Divide: Britain, India, Pakistan* (1969), the definitive account of events from 1930 onward, especially with reference to Mountbatten's viceroyalty. (*Post-Independence in India*): W.H. MORRIS-JONES, *Parliament in India* (1957), a scholarly analysis of the political power structure in Nehru's heyday; SELIG S. HARRISON, *India: The Most Dangerous Decades* (1960), a prophetic work outlining the pattern of politics beyond the Nehru system; GENE D. OVERSTREET and MARSHALL WINDMILLER, *Communism in India* (1959), an account of the rise of Communism to the threshold of power; DONALD EUGENE SMITH, *India As a Secular State* (1963), an extremely valuable study of the interaction of religious and political pressures;

MYRON WEINER (ed.), *State Politics in India* (1968), a survey of regionalism in eight states in the 1960s; RAJNI KOTHARI, *Politics in India* (1970), a sophisticated political science textbook; DILIP HIRO, *Inside India Today,* 2nd ed. (1978), a journalist's investigation into Mrs. Gandhi's emergency; DOM F. MORAES, *Indira Gandhi* (U.K. title, *Mrs. Gandhi,* both 1980), a sympathetic political biography; MAYANTARA SAHGAL, *Indira Gandhi: Her Road to Power* (1982), a less complimentary biography and an analysis of contemporary Indian politics; VED MEHTA, *A Family Affair: India Under Three Prime Ministers* (1982), an account of Indian politics from 1975 to the early 1980s.

Andhra Pradesh. The GOVERNMENT OF INDIA, *Census of India—1961,* vol. 2 (1966), contains information on population, urbanization, and industrialization trends in the state. The GOVERNMENT OF ANDHRA PRADESH, *Fourth Five-Year Plan* (1965), provides information on the state's economy. V. JAGANNADHAM, *Social Welfare Organisation* (1967), and the Central Social Welfare Board's monthly organ, *Social Welfare,* contain material on the organization and work of the social welfare programs in the state. The *Andhra Pradesh Almanac* (annual) includes very useful information to the general reader on various aspects of the state. Other useful sources include A. RAMESH CHOUDHARY, *Andhra Pradesh* (1979); V. HANUMANT RAO (ed.), *Andhra Pradesh: Changing Horizons* (1980); T.D. JAGADESAN, *Andhra Pradesh* (1969); and SATYAPRAKASH (ed.), *Andhra Pradesh: A Bibliography* (1969).

Assam. *Census of India, 1961,* vol. 3, *Assam,* part 1-A, *General Report* (1964), a general collection of data on various geographical, economic, cultural, and social aspects of the state; H.P. DAS, *Geography of Assam* (1970), an illustrated geographical account of the state; JOHN P. WADE, *An Account of Assam,* 2 parts, ed. by BENUDHAR SHARMA (1927), an important reference book including the historical geography of important places; P.C. CHOUDHURY, *The History of Civilization of the People of Assam to the 12th Century, A.D.,* 2nd ed. (1966), a study of prehistoric and ancient Assam; S.K. BHUYAN, *Anglo-Assamese Relations, 1771–1826* (1949), a detailed analysis of the late Ahom and early British periods; see also *Ahom-Buranji* (1930), a chronicle of Ahom rule in the Ahom language with a parallel English translation; D.N.D. GOSWAMI, *Geology of Assam* (1960), a systematic treatment of the subject; P.D. STRACEY and H.P. DAS, *Assam's Economy and Forest* (1949); and M.C. JACOB, *The Forest Resources of Assam* (1940), classifications and evaluations of forest resources; DEPARTMENT OF ECONOMICS AND STATISTICS, ASSAM, *Statistical Handbook, Assam* (annual), a pocketbook collection of important data; P.C. GOSWAMI, *The Economic Development of Assam* (1963), an analysis of various resources and their utilization; P.B. BARTHAKUR, *Assam* (1971); and SATYAPRAKASH (ed.), *Assam: A Bibliography* (1976).

Bihār. P. DAYAL, *Bihar in Maps* (1953), a source of reference for Bihār geography: "The Bihar Plain: A Regional Study," *Trans. Indian Council of Geographers,* vol. 5 (1968); R.R. DIWAKAR (ed.), *Bihar Through the Ages* (1959), an authoritative cultural and political history up to the post-independence period with introductory chapters on geology and geography; SIR JOHN HOULTON, *Bihar: The Heart of India* (1949), a popular account of history and culture; NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of Bihar,* vol. 1 (1959), a guide to economic resources and potentialities; *Census of India, 1961,* vol. 4 *Bihar,* pt. 1-A, *General Report on the Census* (1969), and pt. 9, *Census Atlas of Bihar* (1968); E. AHMAD, *Bihar: A Physical, Economic and Regional Geography* (1965), sections on population and settlement; DIRECTORATE OF STATISTICS AND EVALUATION, BIHAR, *Bihar Through Figures . . .* (annual). Also see P.C. CHAUDHURY, *Bihar* (1974); and SATYAPRAKASH (ed.), *Bihar: A Bibliography* (1976).

Gujarāt. GUJARAT, INDIA, DIRECTORATE OF INFORMATION, *Gujarat: A Reference Annual,* a yearly survey with statistical details of various state activities; H.D. SANKALIA, *Investigations in Prehistoric Archaeology of Gujarat* (1946), an account of the riverside settlements of prehistoric times; A.K. MAJUMDAR, *Chaulukyas of Gujarat* (1956), a history of the Solaṅki Vāghelā period, with a discussion of the administrative system and the social conditions of the time; M.S. COMMISSARIAT, *A History of Gujarat,* 2 vol. (1938–57), dealing with the history of the state from the Muslim sultanate to the break-up of the Mughal Empire, including an account of the socio-economic and cultural life; K.R. DIKSHIT, *Geography of Gujarat* (1970), a brief outline of all relevant topics; W.T. SAXTON and L.J. SEDGWICK, "Plants of Northern Gujarat," *Rec. Bot. Surv. India,* vol. 6 (1918), pp. 207–323; NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of Gujarat;* GUJARAT CHAMBER OF COMMERCE, *Gujarat at a Glance—Statistical Data*; GOVERNMENT OF INDIA, *Census of India, 1961,* vol. 5, pt. 9 (1965), a census atlas of Gujarāt, showing physical aspects and demographic trends; K.M. MUNSHI, *Gujarāt and Its Literature, from Early Times to 1852,* 3rd ed. (1967), tracing the evolution of Gujarati literature; T.N. DAVE, *Language of Gujarat* (1964), a broad outline of the evolution of Gujarātī and Indo-European through Old and Middle Indo-Aryan down to old and modern Gujarātī. Also useful are N.K. BHATT, *Gujarat* (1972); GOVERNMENT OF GUJARAT, *Handbook of Basic Statistics: Gujarat State, 1969 to 1976,* vol. 1 (1978); GOVERNMENT OF GUJARAT, *Handbook of Basic Statistics, Gujarat State, 1969 to 1976,* vol. 2 (1978); and SATYAPRAKASH (ed.), *Gujarat: A Select Bibliography* (1977).

Haryāna. R.S. SHARMA (ed.), *Haryana Directory & Who's Who: 1967–68* (1968), gives the background of the formation of Haryāna state and various details of the state's economy. The *Haryana Review* (quarterly), gives useful information about the state; *Facts about Haryana* (1970), is a small statistical handbook. Both are published by the Public Relations Department, Haryāna. See also S.B. SINGHAI, *Haryana* (1971); and JASBIR SINGH, *An Agricultural Geography of Haryana* (1976).

Himāchal Pradesh. Y.S. PARMAR, *Himachal Pradesh: Its Proper Shape and Status* (1965) and *Himachal Pradesh: Case for Statehood* (1968), present the historical perspective of the movement for a separate state of Himāchal Pradesh, by the first chief minister of the state. R.K. KAUSHAL, *Himachal Pradesh* (1965), is a historical study that deals with developments leading to the formation of the state. H.C. SARASWAT, *Himachal Pradesh* (1970), provides a synoptical account of the state, its people, social customs, and economy. The *Census of India,* Paper 1 of 1971 (1971), gives the population and other vital statistics. The *Himachal Pradesh: Fourth Five Year Plan* (1969), embodies the programs formulated by the state government for development in agriculture, power and irrigation, transport and communications, social services, and other fields. Other useful publications include H.C. SARASWAT, *Himachal Pradesh* (1970); A.P. AGARWALA (ed.), *Holiday Resorts of Himachal Pradesh* (1977); S.S. SHASHI, *Himachal: Nature's Peaceful Paradise* (1971).

Jammu and Kashmir. A good introduction to the physical landscapes of the western Himalayas is provided by HELMUT DE TERRA and T.T. PATERSON, *Studies on the Ice Age in India and Associated Human Cultures* (1939); and by KENNETH MASON, "The Glaciers of the Karakoram and Neighbourhood," *Rec. Geol. Surv. India,* 63:214–278 (1930). General geographical accounts are contained in KAZI S. AHMAD, *A Geography of Pakistan* (1964); D.N. WILBER *et al., Pakistan: Its People, Its Society, Its Culture* (1964); RAM NANDAN, "Jammu and Kashmir," *Focus,* 13:1–6 (1962); and HARALD UHLIG, "Hill Tribes and Rice Farmers in the Himalayas and South-East Asia." *Trans. Inst. Br. Geogr.,* 47:1–23 (1969)—the latter develops an interesting account of the complementary relationship between hill folk and valley folk in the mountainous environment of Kashmir and of interlocking systems of transhumance. P.N.K. BAMZAI, *A History of Kashmir* (1962); and K.M. PANIKKAR, *The Founding of the Kashmir State: A Biography of Maharajah Gulab Singh, 1792–1858* (1953), provide sound accounts of the formation of the state of Jammu and Kashmir under British political protection. The development of political conflicts in the postwar period is dealt with by a number of authors. ALASTAIR LAMB, *Crisis in Kashmir, 1947–1966* (1966), analyzes the recent history leading to partition and contains a comprehensive bibliography; the recurrent discussions between India and Pakistan are chronicled in the *United Nations Official Records of the General Assembly* and *. . . of the Security Council;* the historical background to the Chinese occupation of parts of Ladākh is presented in MARGARET W. FISHER, LEO E. ROSE, and ROBERT A. HUTTENBACK, *Himalayan Battleground: Sino-Indian Rivalry in Ladakh* (1963). The influence of the diverse physical environments on the political problems is dealt with in R.C. MAYFIELD, "A Geographic Study of the Kashmir Issue," *Geogrl. Rev.,* 45:181–196 (1955); ALI TAYYEB, *Pakistan: A Political Geography* (1966), and by NANDAN (*op. cit.*); while the economic issues and consequences are indicated in P.N. DHAR, "The Kashmir Problem: Political and Economic Background," *India Quarterly,* 7:143–162 (1951); and in INDIA, MINISTRY OF INFORMATION AND BROADCASTING, *India: A Reference Manual* (annual). Other important books are P.N.K. BAMZAI, *Jammu and Kashmir* (1972); STATE STATISTICS BUREAU, GOVERNMENT OF JAMMU AND KASHMIR, *Digest of Statistics: Jammu and Kashmir, 1958–59;* A.P. AGARWALA (ed.), *Holiday Resorts of Jammu and Kashmir* (1977) A.N. RAINA, *Geography of Jammu and Kashmir* (1971); PANDIT ANAND KOUL, *Geography of the Jammu and Kashmir State* (1925); BALRAJ PURI, *Jammu and Kashmir: Triumph and Tragedy of Indian Federalism* (1981); and J. WARIKOO (ed.)

Jammu, Kashmir & Ladakh: A Classified & Comprehensive Bibliography (1977).

Karnātaka. LEWIS RICE, *Mysore: A Gazetteer,* 2 vol. (1897), is an authoritative work on the historical aspects of Mysore. See also the multivolume *Mysore Gazetteer* (1927–30), ed. by C. HAYAVADANA RAO, and his *History of Mysore, 1399–1799 A.D.,* new ed., 3 vol. (1943–46). Research papers on social conditions by G.K. GHORI are collected in *Man and Food in Mysore* (1957). L. DUDLEY STAMP, in *Asia,* 12th ed. (1967), provides a discussion of the economic and geographic situation. For contemporary statistics and description, see B.N. SRI SATHYAN, *Mysore State Gazetteer* (1965); GOVERNMENT OF KARNATAKA, *Karnataka State Gazetteer* (1973) and *Karnataka, Draft Five Year Plan 1978–1983;* and R.P. MISRA, *Geography of Mysore* (1973). For general and statistical information see R.P. MISRA, *Geography of Mysore* (1973); N.S. RAMACHANDRIAH, *Mysore* (1972); BUREAU OF ECONOMICS AND STATISTICS, GOVERNMENT OF KARNATAKA, *Karnataka Today: Basic Facts, 1974–75* (1976); DEPARTMENT OF TOURISM, GOVERNMENT OF INDIA, *Mysore* (1972); SATYAPRAKASH (ed.) *Karnataka: A Select Bibliography* (1978).

Kerala. A. SREEDHARA MENON, *A Survey of Kerala History* (1967), a succinct history based on recent research; L.A. KRISHNA IYER, *Social History of Kerala,* vol. 1, *Pre-Dravidians* (1968), vol. 2, *Dravidians* (1970), an anthropological study; W. LOGAN, *Malabar,* 3 vol. (1887–91, reprinted 1951), a classic; GEORGE WOODCOCK, *Kerala: A Portrait of the Malabar Coast* (1967), an admirable monograph with a contemporary focus; NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of Kerala* (1962); BUREAU OF ECONOMIC STUDIES, KERALA, *Kerala: An Economic Review* (annual); STELLA KRAMRISCH, *The Arts and Crafts of Kerala* (1970), fully illustrated. Also useful are MINISTRY OF INFORMATION AND BROADCASTING, GOVERNMENT OF INDIA, *Kerala* (1977); GEORGE VALLATT, *Discovery of Kerala* (1977); SATYAPRAKASH (ed.), *Kerala: A Select Bibliography* (1979).

Madhya Pradesh. The *Census of India, 1961,* vol. 8, *Madhya Pradesh,* pt. 2 (1963), and *Census of India: Madhya Pradesh, 1971,* pt. II A, are collections of general population tables; while the *Statistical Abstract of Madhya Pradesh* (annual) is a compendium of Madhya Pradesh statistics. See also the *Economic and Statistical Atlas of Madhya Pradesh* (1958). An account of the state's economy is found in the NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of Madhya Pradesh* (1960). O.H.K. SPATE and A.T.A. LEARMONTH, *India and Pakistan,* 3rd ed. rev. (1967), is a general economic and regional geography. D.R. PATIL, *The Cultural Heritage of Madhya Bharat* (1953), deals with the pre-1956 state of Madhya Bharat in its geographical and cultural setting.

Mahārāshtra. The GOVERNMENT PUBLICITY DIRECTORATE, *Maharashtra* (1969), *Maharashtra at a Glance* (1969); and the BUREAU OF ECONOMICS AND STATISTICS publication *Statistical Abstract of Maharashtra* (annual), furnish reliable, up-to-date figures on the state. Much historical material may be found in G. YAZDANI (ed.), *The Early History of the Deccan,* 2 vol. (1960); and a brilliant socio-anthropological analysis of the population is I. KARVE, *Maharashtra Land and Its People* (1968). Resource potentials are discussed in the *Techno-Economic Survey of Maharashtra,* ed. by P.S. LOKANATHAN (1963). A good geographical account is C.B. JOSHI and B. ARUNACHALAM, *Maharashtra: A Regional Study* (1962). S.R. TIKEKAR, *Maharashtra* (1971); and DIRECTORATE GENERAL OF INFORMATION AND PUBLICITY, MAHARASHTRA STATE, *Maharashtra at a Glance: 1975–76* (1976), are also useful.

Meghālaya. *The Tribes of Assam,* a National Book Trust publication (1969), gives in a very concise form detailed information about the various tribes of Assam and among them the Gāros and Khāsis. KAMALESHWAR SINHA, *Meghalaya: Triumph of Tribal Genius* (1970), contains a great deal of history and other information on the state; the *Meghalaya Year Book* provides a variety of statistical and other information about the state. Also useful is D.T. ZIMBRA, *Meghalaya* (1978).

Nāgāland. The *Census of India, 1961, Nagaland,* gives a general account of the state together with important data on geographic, economic, cultural, and social aspects. ANAND V.K. MAJOR, *Nagaland in Transition* (1967), is a traveller's account of Nāgāland. *The Gazetteer of India,* vol. 1, *Country and People* (1965), contains ten chapters written by eminent scholars on the physiography, climate, geology, flora, fauna, people, languages, religion, social structure, and life of all states of the Indian Union. See also ELWIN VERRIER, *Nagaland* (1961); *Statistical Handbook of Nagaland* (1967), an invaluable publication for statistical data; and PRAKASH SINGH, *Nagaland* (1972).

Orissa. General and historical works include R.D. BANERJI, *History of Orissa,* 2 vol. (1930–31); M.M. GANGULY, *Orissa and Her Remains* (1912); H.K. MAHATAB, *History of Orissa,* 2 vol. (1950); S. MISRA, *Economic Survey of Orissa,* 2 vol. (1960); G.C. MOHAPATRA, *The Stone Age Cultures of Orissa* (1962); S.N. RAJGURU, *Inscriptions of Orissa,* 4 vol. (1958–60); and the UTKAL UNIVERSITY, *History of Orissa,* vol. 1 (1964) and *Orissa, Past and Present* (1962). See also the *District Census Handbooks (Orissa),* 13 vol. (1965), which contain detailed statistics relating to economic conditions, agriculture, administration, and education; *Techno-Economic Survey of Orissa* (1961), a publication of the government of India that includes socio-economic data; S.C. DASH, *Orissa* (1970).

Punjab. S.M. LATIF, *History of the Punjab* (1891, reprinted 1964), the first comprehensive account in English by an Indian; INDUBHUSAN BANERJEE, *Evolution of the Khalsa,* 2nd ed., 2 vol. (1963), a modern history of the Sikhs, and thus of Punjab, by an Indian historian using original Punjābī sources; KHUSHWANT SINGH, *A History of the Sikhs,* 2 vol. (1963–66), a popular history by a Sikh scholar; HARBANS SINGH, *The Heritage of the Sikhs* (1964), an account of the Sikhs from 1469 to 1947; L. DUDLEY STAMP, *India, Pakistan, Ceylon and Burma* (1957), with a section on natural regions; R.L. ANAND, B.S. OJHA, and GURDEV SINGH GOSAL, *Punjab Census Atlas* (1966), contains a cartographic portrayal of the relief, demography, and economy of Punjab; GURDEV SINGH GOSAL and GOPAL KRISHNA, "Upper Bari Doab," in R.L. SINGH (ed.), *India: Regional Studies* (1968), deals with physical, social, and economic phenomena in a specific tract—Upper Bari Doab; P.D. STRACEY, *Wild Life in India: Its Conservation and Control* (1963), includes charts on regional distribution; *Report of the Punjab Administrative Reforms Commission* (1966); EDWARD NIRMAL MANGAT RAI, *Civil Administration in the Punjab: An Analysis of State Government in India* (1963); W.G. ARCHER, *Paintings of the Sikhs* (HMSO, 1966), the only work to date on the paintings of the Sikhs by a Western art critic; RICHARD C. TEMPLE, *The Legends of the Punjab,* 3 vol. (1884–1900, reprinted 1962–63), a well-known work on the folklore of the Punjab, consisting of original texts in Roman script with English translation; MOHAN SINGH, *A History of Panjabi Literature, 1100–1932,* 2nd ed. (1956), the first history of Punjābī literature in English by a Punjābī scholar. See also S.S. GUPTA, *Punjab* (1971).

Rājasthān. General works include V.C. MISRA, *Geography of Rajasthan* (1967); DHARAM PAL, *Rajasthan* (1968); and the *Imperial Gazetteer of India, Provincial Series, Rajasthan* (1968). S.P. RAYCHAUDHURI *et al., Soils of India* (1963), contains mineralogical studies and detailed soil profiles for various districts of the state of Rājasthān. For history and social conditions, see J. TOD, *Annals and Antiquities of Rajasthan,* 3 vol. (1829–32; 2 volume edition, 1914); R.E.U. BISHESHWAR, *Glories of Marwar and the Glorious Rathors* (1943); G.N. SHARMA, *Social Life in Mediaeval Rajasthan, 1500–1800 A.D.* (1968); V.P. MENON, *The Story of the Integration of the Indian States* (1956); IQBAL NARAIN, *Panchayati Raj: Planning and Democracy in Pakistan* (1969); and the Rajasthan Planning Department's *Five-Year Plans.* See also A.K. COOMARSWAMI, *Rajput Painting* (1916); M.L. SHARMA, *Rajasthan* (1971); IRIS TURNER, *Rajasthan Travel Guide* (1978); S.S. GAHLOT, *Rural Life in Rajasthan* (1982); and J.S. GAHLOT, *Rajasthan: A Socio-Economic Study* (1981).

Sikkim. C.U. AITCHISON, *A Collection of Treaties, Engagements and Sunnuds Relating to India and Neighbouring Countries,* 5th ed. (1929), includes the texts of three 19th-century treaties with Great Britain; CHARLES A. BELL, *Tibet Past and Present* (1924), contains a discussion of Sikkim's history and politics, based on the author's experience as Political Officer in Sikkim; GEOFFREY GORER, *Himalayan Village: An Account of the Lepchas of Sikkim* (1938), an ethnological study of this society in the village of Lingthem; JOSEPH DALTON HOOKER, *Himalayan Journals,* rev. ed., 2 vol. (1855), a classic work that discusses comprehensively the history, religion, customs, and botany of Sikkim in the 19th century; MARCO PALLIS, *Peaks and Lamas* (1939), a popular, anecdotal account of mountaineering and a form of Mahāyāna Buddhism in Sikkim; THE ROYAL WEDDING COMMITTEE, *Sikkim: Facts and Figures* (1980), a digest of essential data and statistics on Sikkim. Sikkim's political history and the merger with India are discussed in B.S.K. GROVER, *Sikkim and India: Storm and Consolidation* (1974); S.R. SHUKLA, *Sikkim: The Story of Integration* (1974); and A.C. SINHA, *Politics of Sikkim: A Sociological Study* (1975). Geography and modern socioeconomic conditions are covered in SUKHDEV SINGH CHIB, *This Beautiful India: Sikkim* (1977), and P.N. CHOPRA, *Sikkim* (1979). More general recent works are V.H. COELHO, *Sikkim and Bhutan* (1970); C.A. PARKHURST, *Sikkim* (1946); C.Y. SALISBURY, *Sikkim: The Mountain Kingdom* (1974); G. KOTTURAN, *Himalayan Gateway: History and Culture of Sikkim* (1983); and the *Gazetteer of Sikkim* (1973).

The official view about the history of Sikkim is contained in *Sikkim: A Concise Chronicle* (1970), while an official account of the land and people is found in *Sikkim: The Land and People* (1972), published by the government of Sikkim. See also S.C. ROY, *Sikkim* (1980).

Tamil Nādu. K.A. NILAKANTA SASTRI, *The Culture and History of the Tamils* (1964) and *A History of South India from Prehistoric Times to the Fall of Vijayanagar,* 3rd ed. (1966), are eminently readable and factual texts. Important data and an excellent atlas of the state may be found in the *Census of India 1961,* vol. 9, *Madras,* pt. 1–8 (1965–67). The proceedings of the 21st International Geographical Congress, Madras, *This Is Tamil Nadu* (1968), contains a collection of papers on a wide range of topics. See also C.V. RAGHAVAN, *Tamil Nādu* (1973).

Uttar Pradesh. R.L. SINGH (ed.), *India: Regional Studies* (1968), contains some articles on the regional geography of Uttar Pradesh. *India 1970,* a government of India publication (1970), is a reference annual that includes information on various aspects of Uttar Pradesh. For general and statistical information see J.P. CHATURVEDI, *Uttar Pradesh* (1970); and KRIPA SHANKAR, *Uttar Pradesh in Statistics* (1978).

West Bengal. D.D. BASU (ed.), *Shorter Constitution of India,* 4th ed. (1964), deals with the constitutional framework governing the relationship between the State Government and the Union Government of India. Government of India sponsored publications, such as S.C. BOSE, *Geography of West Bengal* (1968); and SUKUMAR SEN, *History of Bengali Literature* (1960), are the main publications in their respective fields. The West Bengal Government has brought out a number of publications on traditional arts and industries, such as *Designs in Traditional Arts of Bengal* (1963). Statistical information on West Bengal is available from Government publications. The MINISTRY OF INFORMATION AND BROADCASTING, *India: A Reference Annual,* contains information on diverse aspects of national life including vital statistics, agriculture, planning, community development, education, and health. More detailed demographic data may be found in *India's Census,* vol. 16, *West Bengal and Sikkim,* pt. 1A (1961). Statistics of mineral production are compiled by the INDIAN BUREAU OF MINES, NAGPUR, NATIONAL COUNCIL OF APPLIED ECONOMIC RESEARCH, *Techno-Economic Survey of West Bengal* (1962), containing useful base statistics on the economic growth of West Bengal between 1951 and 1960. More recent information on agriculture, industry, commerce, and transport is available in the CENTRAL STATISTICAL ORGANISATION, *Statistical Abstract* (annual), which analyzes in detail the growth of infrastructure facilities in different states since 1951. See also GEOGRAPHICAL INSTITUTE, PRESIDENCY COLLEGE, *West Bengal* (1970); S.C. BOSE, *Geography of West Bengal* (1978); S.L. GHOSH, West Bengal (1978).

Indian Ocean Islands

The Indian Ocean is bounded on the north by Iran, Pakistan, India, and Bangladesh; on the west by the Arabian Peninsula and Africa; on the east by the Malay Peninsula, the Sunda Islands of Indonesia, and Australia; and on the south by Antarctica. An arbitrary separation from the Atlantic is made at the longitude of Cape Agulhas (20° E) and from the Pacific at the longitude of Tasmania (147° E).

Madagascar, the largest of the ocean's islands, rises from a submarine plateau extending out from Africa. Sri Lanka and Socotra are continental islands. The Maldives and the Laccadive and Chagos islands rise from the Chagos–Laccative Plateau. The islands extending from the Seychelles to Mauritius rise from the Seychelles–Mauritius Plateau and are separated from Madagascar by the deep water of the Mascarene Basin and, farther south, the Madagascar Basin. The Andaman and Nicobar Islands rise from a ridge extending southward from Cape Negrais. Between the ridge and the Malay Peninsula lies the Andaman Basin opening only to the south. Many oceanic islands, such as Mauritius, Crozet, and St. Paul's, are volcanic. Those in tropical seas are often coral formations (such as Maldives) or possess coral reefs (as do Comoros, Mauritius, and Réunion).

Several of the islands in the Indian Ocean are treated elsewhere in the *Macropædia.* For Christmas Island and the Cocos (Keeling) Islands, see AUSTRALIA. For Lakshadweep and the Andaman and Nicobar Islands, see INDIA. For Socotra, see ARABIA. Sri Lanka is treated separately. Of the islands treated in this article, Madagascar, Maldives, Mauritius, Seychelles, and Comoros are independent states. Réunion is an overseas *département* of France. All except Maldives are represented in the Indian Ocean Commission, established in 1982 to promote political and economic cooperation in the region.

(Ed.)

The article is divided into the following sections:

Madagascar 164
Maldives 174
Seychelles 174
Comoros 176
Mauritius 177
Réunion 178
Bibliography 179

Madagascar

The Democratic Republic of Madagascar (Repoblika Demokratika Malagasy, or République Démocratique de Madagascar) lies off the southeastern coast of Africa. It occupies the fourth largest island in the world, after Greenland, New Guinea, and Borneo, with a surface area of 226,658 square miles (587,041 square kilometres). Located in the southwestern Indian Ocean, it is separated from the African coast by the 500-mile- (800-kilometre-) wide Mozambique Channel.

In spite of Madagascar's proximity to the continent, its population is related not to African peoples but rather to those of Indonesia, more than 3,000 miles to the east. The Malagasy peoples do not consider themselves to be Africans; because of the continuing bond with France that resulted from former colonial rule, the island has developed political, economic, and cultural links with French-speaking countries of West Africa. French, as well as Malagasy, has official status as a language.

Madagascar remains a geographical and historical paradox, linked in practice to Africa but in feeling identified with Indonesia, which is so far away as to have hardly any awareness of Madagascar or to maintain any contemporary ties of substance with it. The Malagasy government has sought a role as a political, economic, and cultural bridge between Africa and Asia, but this desire has not been expressed in any institutional form. The animal life and vegetation of the island are equally anomalous, differing greatly from that of nearby Africa and being, in many respects, unique.

Although the coastlands have been known to Europeans for more than 400 years and for much longer to Arabs, recent development has been more intense and concentrated in the central plateau, which contains the capital city of Antananarivo (former Tananarive). The road network and communications are generally better on the plateau, where the majority of the inhabitants have received some school education and are professing Christians, while in the coastal areas the majority follow traditional religions and generally have not attended school. (Ai.S.)

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Madagascar consists of three parallel longitudinal zones—the central plateau, the narrow coastal strip in the east, and the zone of low plateaus and vast plains in the west.

The plateau

Situated between 2,500 and 4,500 feet (800 and 1,400 metres) above sea level, the plateau has been folded and worn down several times and is tilted to the west, where the slope is less sheer than it is to the east. Three massifs are more than 8,500 feet high. Tsaratanana in the north is separated from the rest of the plateau by the Massif du Tsaratanana, whose summit at an elevation of 9,436 feet is the highest point on the island. Ankaratra in the centre is an enormous volcanic mass whose summit, Tsiafajavona, is 8,671 feet high. They act as a watershed separating three main river basins. Farther south, Andringitra is a vast granite massif north of Faradofay; it rises to 8,720 feet at Pic Boby.

The plateau slopes more regularly toward the extreme southern plain, but its boundaries to the east and west are more abrupt. To the east it descends in a sharp fault, by vertical steps of 1,000 to 2,000 feet. This cliff, which is called the Great Cliff or Cliff of Angavo, is often impassable and is itself bordered by the Betsimisaraka Escarpment, a second and lower cliff to the east, which overhangs the coastal plain.

Behind the scarp face are the remains of ancient lakes, including one called Alaotra. To the south the two steep gradients meet and form the Mahafaly and the Androy plateaus, which overhang the sea in precipitous cliffs. Toward the west the descent is made in a series of steps. In places, however, the central plateau is bordered by an impassable escarpment, one such location being the Cliff of Bongolava. To the extreme north, the plateau is separated by the low belt of Mt. d'Ambre, which includes a series of volcanic craters.

The coasts

The eastern coast has an average width of 30 miles. It is a narrow alluvial plain that terminates at the sea in a low coastline bordered with extensive lagoons linked together by the Canal des Pangalanes, which is 400 miles long. To the south of Farafangana, the coast becomes rocky, and in the southeast there are a number of little bays. The deep Baie d'Antongil (Bay of Antongil) is located to the northeast.

The western zone is between 60 and 125 miles wide. Its sedimentary layers slope toward the Mozambique Channel and produce a succession of hills. The inland (eastern) side of these steep hills dominates the hollows formed in the soft sediments, while the other side descends to the sea in rocky slopes. The coastline is straight, bordered by small dunes and fringed with mangroves. The currents in the Mozambique Channel have favoured the deposit of alluvium and the growth of river deltas. On the northwestern coast there are a number of estuaries and bays. This coast is bordered by coral reefs and volcanic islands, such as Nosy Be (Nossi-Bé), which protects the Baie d'Ampasindava.

(Je.D.)

Drainage. The steep eastern face of the plateau is drained by numerous short, torrential rivers, which discharge either into the coastal lagoons or directly into the sea over waterfalls and rapids. They include the Mandrare, the Mananara, the Faraony, the Ivondro, and the Maningory. The more gently sloping western side of the plateau is crossed by longer and larger rivers, including the Onilahy, the Mangoky, the Tsiribihina, and the Betsiboka, which bring huge deposits of fertile sediments down into the vast plains and many-channelled estuaries; the river mouths, while not completely blocked by this sediment, are full of sandbanks.

The volcanic lakes

There are numerous lakes of volcanic origin, such as Lac Itasy. Alaotra is the last surviving lake of the eastern slope. Lac Tsimanampetsotsa, near the coast south of Toliary (former Tuléar), is a large body of saline water that has no outlet.

Soils. The central plateau and the eastern coast are mainly composed of gneiss, granite, quartz, and other crystalline formations. The gneiss decomposes into red murrum (solid ironstone), laterite, and deeper and more fertile red earths, giving Madagascar its colloquial name of the Great Red Island. Fertile soils in the valleys support intensive cultivation by the island's inhabitants. There are also northern, central, and southeastern volcanic intrusions that produce fertile but easily erodible soils. Lac Alaotra (Lake Alaotra) is a sedimentary pocket in the central plateau, north of Imerina. The western third of the island consists entirely of deposits of sedimentary rock that were raised up from under the sea or that underlay ancient lakes.

Climate. The hot, wet season extends from November to April and the cooler, drier season from May to October. The climate is governed by the combined effects of the southeast trade winds and the northwest monsoon winds as they blow across the central plateau. The trade winds, which blow throughout the year, are strongest from May to October. The east coast is to the windward and receives a high rate of annual precipitation, reaching about 147 inches (3,700 millimetres) at Maroantsetra on the Baie d'Antongil. As the winds cross the plateau, they lose much of their humidity, causing drizzle and mists on the plateau and leaving the west dry.

The monsoon, bringing rain to the northwest coast of Madagascar and the plateau, is most noticeable during the hot, humid season. The wind blows obliquely onto the west coast, which receives a moderate rate of precipitation; the southwest, which is protected, remains arid. Annual rainfall drops from 83 inches on the northwestern island of Nosy Be to 37 inches at Maintirano and 30 inches at Morondava (both on the west coast) and to 14 inches at Toliary in the southwest. The plateau receives moderate rains, with 53 inches falling annually at Antananarivo and 48 inches at Fianarantsoa, which lies about 200 miles farther south.

July is the coolest month, with temperatures ranging between about 50° and 78° F (10° and 26° C), and December the hottest month, with temperatures between 61° and 84° F (about 16° and 29° C). Temperatures are highest on the northwest coast and lowest on the plateau. The annual average temperature at Antsiranana (former Diégo-Suarez), located on the extreme northern coast, is 81° F (27° C) and at Faradofay (former Fort-Dauphin), situated on the extreme southeastern coast, 73° F (23° C). On the central plateau, however, the annual average temperature is 63° F (17° C) at Antananarivo and 65° F at Fianarantsoa.

Cyclones are an important climatic feature. They form far out over the Indian Ocean, especially from December to March, and approach the eastern coast, bringing torrential rains and destructive floods.

Plant and animal life. Much of the island was once covered with evergreen and deciduous forest, but little now remains except on the eastern escarpment and in scattered pockets in the west. The plateau is particularly denuded and suffers seriously from erosion. The forest has been cut in order to clear rice fields, to obtain fuel and building materials, and to export valuable timber, such as ebony, rosewood, and sandalwood. About seven-eighths of the island is covered with prairie grasses and bamboo or small thin trees, including the travellers' tree, which holds water in its cellular trunk and produces the long banana-like leaves that are widely used for thatching. There are also screw pines, palms, and reeds on the coasts. In the arid south of the island grow thorn trees, giant cacti, dwarf baobab trees, pachypodium succulents, and other xerophytes (drought-resistant plants) that are peculiar to the island.

The grassland vegetation

Because of the island's isolation, many zoologically primitive primates have survived and evolved into unique forms. About 40 species of lemurs are indigenous to Madagascar. Several unique hedgehog-like insectivores such as the Centetidae, or tenrecs, have evolved, and there are all sizes of chameleons. Birds are numerous and include guinea fowls, partridges, pigeons, herons, ibis, flamingos, egrets, cuckoos, Asian robins, and several kinds of birds of prey. There are about 800 species of butterfly, many moths, and a variety of spiders. The only large or dangerous animals are the crocodiles, which occupy the rivers. The snakes, including the *do,* which is 10 to 13 feet in length, are harmless.

Inland waters contain tilapia (a perchlike fish), rainbow trout, and black bass. Marine fish and crustaceans abound on the coasts and in the lagoons, estuaries, and even upland streams. They include grouper, gilt-head, tuna,

MAP INDEX

Political subdivisions

Antananarivo....19·00s 47·00e
Antsiranana.....13·30s 49·10e
Fianarantsoa.....22·00s 47·00e
Mahajanga.....17·00s 46·00e
Toamasina.....18·00s 48·40e
Toliary.....24·00s 45·00e

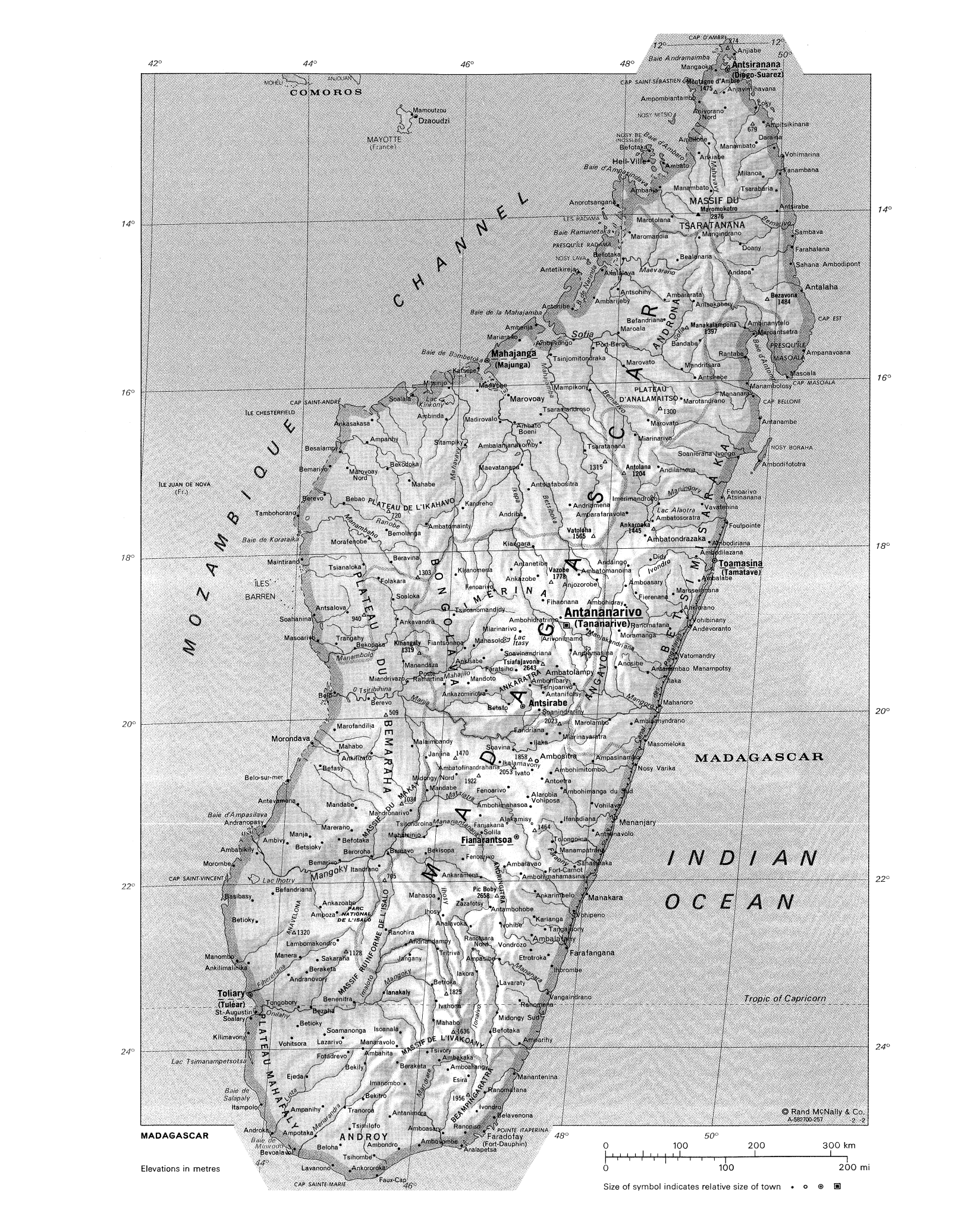
The names of the political subdivisions do not appear on the map because they are the same as the names of their capital cities.

Cities and towns

Alakamisy.....21·19s 47·14e
Alarobia Vohiposa.....20·59s 47·09e
Ambahikily.....21·36s 43·41e
Ambahita.....24·01s 45·16e
Ambakaka.....24·10s 46·17e
Ambalabe.....18·24s 49·10e
Ambalafany.....22·32s 47·36e
Ambalanjana-komby.....16·42s 47·05e
Ambalavao.....21·50s 46·56e
Ambanja.....13·41s 48·27e
Ambararata.....15·03s 48·33e
Ambarijeby.....14·56s 47·41e
Ambato.....13·24s 48·29e
Ambato Boeni...16·28s 46·43e
Ambatofinandra-hana.....20·33s 46·48e
Ambatolampy....19·23s 47·25e
Ambatomainty...17·41s 45·40e
Ambatomanoina..18·18s 47·37e
Ambatondrazaka.17·50s 48·25e
Ambatosoratra...17·34s 48·32e
Ambenja.....15·17s 46·58e
Ambevongo.....15·27s 47·27e
Ambilobe.....13·12s 49·04e
Ambinanyn-drano.....20·20s 48·19e
Ambinanytelo....15·21s 49·35e
Ambinda.....16·25s 45·52e
Ambivy.....21·31s 44·02e
Amboahangy....24·15s 46·22e
Amboasary.....18·26s 48·16e
Amboasary.....25·02s 46·23e
Ambodifototra...16·59s 49·52e
Ambodilazana...18·06s 49·10e
Ambodiriana....17·55s 49·18e
Ambohibary.....19·20s 46·17e
Ambohidratrimo.18·50s 47·26e
Ambohidray.....18·36s 48·18e
Ambohimahama-sina.....21·56s 47·11e
Ambohimahasoa.21·07s 47·13e
Ambohimanga du Sud.....20·52s 47·36e
Ambohimitombo.20·43s 47·26e
Ambondro.....25·13s 45·44e
Ambositra.....20·31s 47·15e
Ambovombe.....25·11s 46·05e
Ampanavoana...15·41s 50·22e
Ampanihy.....24·42s 44·45e
Amparafaravola .17·35s 48·13e
Amparihy.....16·40s 44·49e
Amparihy.....23·57s 47·20e
Ampasibe.....22·56s 46·58e
Ampasinambo...20·31s 48·00e
Ampitsikinana...12·57s 49·49e
Ampombian-tambo.....12·42s 48·57e
Ampotaka.....25·03s 44·41e
Ampoza.....22·20s 44·44e
Analalava.....14·38s 47·45e
Analapetsa.....25·10s 46·42e
Analavoka.....22·33s 46·30e
Andaingo.....18·12s 48·17e
Andapa.....14·39s 49·39e
Andevoranto.....18·57s 49·06e
Andilamena.....17·01s 48·35e
Andramasina....19·11s 47·35e
Andranopasy....21·17s 43·44e
Andranovory.....23·08s 44·10e
Andriamena.....17·26s 47·30e
Andriandampy...22·45s 45·41e
Andriba.....17·36s 46·55e
Androka.....25·02s 44·05e
Anivorano.....18·44s 48·58e
Anivorano Nord..12·44s 49·13e
Anjavimihavana..12·32s 49·16e
Anjiabe.....12·07s 49·20e
Anjozorobe.....18·24s 47·52e
Ankaramena.....21·57s 46·39e
Ankarimbelo.....22·08s 47·20e
Ankasakasa.....16·21s 44·52e
Ankavandra.....18·46s 45·18e
Ankazoabo.....22·18s 44·31e
Ankazobe.....18·21s 47·07e
Ankazomiriotra ..19·38s 46·32e
Ankiabe.....13·13s 48·56e
Ankilimalinika...22·58s 43·45e
Ankilizato.....20·25s 45·01e
Ankisabe.....19·17s 46·29e
Ankororoka.....25·30s 45·11e
Anorotsangana..13·56s 47·55e
Anosibe.....19·26s 48·13e
Antalaha.....14·53s 50·16e
Antambohobe...22·20s 46·47e
Antanambao Manampotsy....19·29s 48·34e
Antanambe.....16·26s 49·52e
Antananarivo (Tananarive)....18·55s 47·31e
Antanetibe.....18·27s 46·42e
Antanifotsy.....19·39s 47·19e
Antanimora.....24·49s 45·40e
Antetikireja.....14·42s 47·29e
Antevamena.....21·02s 44·08e
Antoetra.....20·46s 47·20e
Antonibe.....15·07s 47·24e
Antsakabary.....15·03s 48·56e
Antsalova.....18·40s 44·37e
Antsenavolo.....21·24s 48·03e
Antsiafabositra...17·18s 46·57e
Antsirabe.....14·00s 49·59e
Antsirabe.....15·57s 48·58e
Antsirabe.....19·51s 47·02e
Antsiranana (Diégo-Suarez)..12·16s 49·17e
Antsohihy.....14·52s 47·59e
Arivonimamo....19·01s 47·15e
Bandabe.....15·31s 49·04e
Basibasy.....22·10s 43·40e
Bealanana.....14·33s 48·44e
Bebao.....17·22s 44·33e
Befandriana.....15·16s 48·32e
Befandriana.....22·06s 43·54e
Befasy.....20·33s 44·23e
Befotaka.....13·15s 48·16e
Befotaka.....14·32s 48·01e
Befotaka.....21·29s 44·44e
Befotaka.....23·49s 46·59e
Bekily.....24·13s 45·19e
Bekisopa.....21·40s 45·54e
Bekitro.....24·33s 45·18e
Bekodoka.....16·58s 45·07e
Bekopaka.....19·09s 44·48e
Belavenona.....24·50s 47·04e
Belo.....19·42s 44·33e
Beloha.....25·10s 45·03e
Belo-sur-mer....20·44s 44·00e
Bemarivo.....16·56s 44·21e
Bemarivo.....21·45s 44·45e
Bemavo.....21·37s 45·24e
Bemolanga.....17·44s 45·06e
Benenitra.....23·27s 45·05e
Beraketa.....23·07s 44·25e
Beraketa.....24·11s 45·42e
Beravina.....18·10s 45·14e
Berevo.....17·14s 44·17e
Berevo.....19·44s 44·58e
Beroroha.....21·41s 45·10e
Besalampy.....16·45s 44·30e
Betafo.....19·50s 46·51e
Betioky.....22·27s 43·44e
Betioky.....23·42s 44·22e
Betroka.....23·16s 46·06e
Betsioky.....21·31s 44·28e
Bevoalavo.....25·13s 45·26e
Bezaha.....23·30s 44·31e
Daraina.....13·12s 49·40e
Didy.....18·07s 48·32e
Diégo-Suarez, see Antsiranana
Doany.....14·22s 49·31e
Ejeda.....24·20s 44·31e
Esira.....24·20s 46·42e
Etrotroka.....22·53s 47·36e
Fanambana.....13·34s 50·00e
Fandriana.....20·14s 47·23e
Fanjakana.....21·10s 46·53e
Faradofay (Fort-Dauphin)..25·02s 47·00e
Farafangana.....22·49s 47·50e
Farahalana.....14·26s 50·10e
Faratsiho.....19·24s 46·57e
Faux-Cap.....25·33s 45·32e
Fenoarivo.....18·26s 46·34e
Fenoarivo.....20·52s 46·53e
Fenoarivo.....21·43s 46·24e
Fenoarivo Atsinanana.....17·22s 49·25e
Fianarantsoa.....21·26s 47·05e
Fiantsonana.....19·09s 46·12e
Fierenana.....18·29s 48·24e
Fihaonana.....18·36s 47·12e
Folakara.....18·20s 45·02e
Fort-Carnot.....21·53s 47·28e
Fort-Dauphin, see Faradofay
Fotadrevo.....24·03s 45·01e
Foulpointe.....17·41s 49·31e
Hell-Ville.....13·25s 48·16e
Iakora.....23·06s 46·40e
Ianakaly.....23·21s 45·28e
Ifanadiana.....21·19s 47·39e
Ihorombe.....23·00s 47·33e
Ihosy.....22·24s 46·08e
Ikalamavony.....21·09s 46·35e
Ilaka.....19·33s 48·52e
Ilaka.....20·20s 47·09e
Imanombo.....24·26s 45·49e
Imerimandroso...17·23s 48·38e
Isoanala.....23·50s 45·44e
Itampolo.....24·41s 43·57e
Itandrano.....21·47s 45·17e
Ivahona.....23·27s 46·10e
Ivato.....20·37s 47·12e
Ivohibe.....22·29s 46·52e
Ivondro.....24·47s 46·52e
Jangany.....22·54s 45·41e
Janjina.....20·30s 45·50e
Kandreho.....17·29s 46·06e
Karianga.....22·22s 47·26e
Katsepe.....15·45s 46·15e
Kiangara.....17·58s 47·02e
Kilimavony.....23·48s 43·41e
Kiranomena.....18·17s 46·03e
Lambomakondro..22·41s 44·44e
Lavanono.....25·24s 44·55e
Lavaraty.....23·16s 46·59e
Lazarivo.....23·54s 44·59e
Loky.....12·47s 49·39e
Madirobe.....16·04s 46·15e
Madirovalo.....16·26s 46·30e
Maevatanana....16·56s 46·49e
Mahabe.....17·05s 45·20e
Mahabo.....20·23s 44·40e
Mahabo.....23·40s 46·08e
Mahajanga (Majunga).....15·43s 46·19e
Mahanoro.....19·54s 48·48e
Mahasoa.....22·12s 46·06e
Mahasolo.....19·07s 46·22e
Mahatsinjo.....21·26s 45·51e
Maintirano.....18·03s 44·01e
Majunga, see Mahajanga
Malaimbandy....20·20s 45·36e
Mampikony.....16·06s 47·38e
Manakara.....22·08s 48·01e
Manambato.....13·14s 49·54e
Manambato.....13·43s 49·07e
Manambolosy....16·02s 49·40e
Manampatrana...21·40s 47·35e
Mananara.....16·10s 49·46e
Manandaza.....19·19s 45·23e
Mananjary.....21·13s 48·20e
Manantenina....24·17s 47·19e
Manaravolo.....23·59s 45·39e
Mandabe.....20·55s 45·49e
Mandabe.....21·03s 44·55e
Mandoto.....19·34s 46·17e
Mandritsara.....15·50s 48·49e
Mandronarivo....21·07s 45·38e
Manera.....22·55s 44·20e
Mangaoka.....12·19s 49·07e
Mangindrano....14·17s 48·58e
Manja.....21·26s 44·20e
Manjakandriana..18·55s 47·47e
Manombo.....22·52s 45·44e
Marerano.....21·23s 44·52e
Mariarano.....15·29s 46·42e
Maroala.....15·23s 47·59e
Maroantsetra....15·26s 49·44e
Marofandilia.....20·07s 44·34e
Marolambo.....20·02s 48·07e
Maromandia.....14·13s 48·08e
Maroseranana...18·32s 48·51e
Marotandrano....16·10s 48·50e
Marotolana.....14·01s 48·37e
Marovato.....15·48s 48·05e
Marovato.....16·28s 48·25e
Marovoay.....16·06s 46·39e
Marovoay Nord...16·57s 44·34e
Masoala.....15·59s 50·10e
Masoarivo.....19·03s 44·19e
Masomeloka.....20·17s 48·37e
Miandrivazo.....19·31s 45·28e
Miarinarivo.....16·38s 48·15e
Miarinarivo.....18·57s 46·55e
Miarinavaratra...20·13s 47·31e
Midongy Nord...20·45s 46·13e
Midongy Sud....23·35s 47·01e
Milanoa.....13·35s 49·47e
Mitsinjo.....16·01s 45·52e
Morafenobe.....17·49s 44·55e
Moramanga.....18·56s 48·12e
Morombe.....21·45s 43·22e
Morondava.....20·17s 44·17e
Nosy Varika.....20·35s 48·32e
Port-Bergé.....15·33s 47·40e
Poste Ramartina..19·38s 45·58e
Ranohira.....22·29s 45·24e
Ranomafana.....18·57s 48·50e
Ranomafana.....24·36s 46·58e
Ranomena.....23·25s 47·17e
Ranopiso.....25·03s 46·40e
Ranotsara Nord..22·48s 46·36e
Rantabe.....15·42s 49·39e
Sahana Ambodipont....14·37s 50·11e
Sahasinaka.....21·49s 47·49e
Saint-Augustin...23·33s 43·46e
Sakaraha.....22·55s 44·32e
Sambava.....14·16s 50·10e
Sitampiky.....16·41s 46·06e
Soahanina.....18·42s 44·13e
Soalala.....16·06s 45·20e
Soalary.....23·36s 43·44e
Soaloka.....18·32s 45·15e
Soamanonga....23·52s 44·47e
Soanierana Ivongo.....16·55s 49·35e
Soanindrariny....19·54s 47·14e
Soavina.....20·23s 46·56e
Soavinandriana..19·09s 46·45e
Solila.....21·25s 46·37e
Tamatave, see Toamasina
Tambohorano....17·30s 43·58e
Tananarive, see Antananarivo
Tangainony.....22·42s 47·45e
Toamasina (Tamatave).....18·10s 49·23e
Toliary (Tuléar).....23·21s 43·40e
Tolongoina.....21·33s 47·31e
Tongobory.....23·32s 44·20e
Trangahy.....19·07s 44·43e
Tranoroa.....24·42s 45·04e
Tritriva.....22·46s 46·07e
Tsarabaria.....13·46s 49·58e
Tsaramandroso..16·22s 47·02e
Tsaratanana.....16·47s 47·39e
Tsianaloka.....18·08s 44·50e
Tsihombe.....25·18s 45·29e
Tsimilofo.....24·59s 45·10e
Tsinjoarivo.....19·37s 47·40e
Tsinjomiton-draka.....15·36s 47·08e
Tsiroanoman-didy.....18·46s 46·02e
Tsitondroina.....21·19s 46·00e
Tsivory.....24·04s 46·05e
Tuléar, see Toliary
Vangaindrano....23·21s 47·36e
Vatomandry.....19·20s 48·59e
Vavatenina.....17·28s 49·12e
Vohibinany.....18·49s 49·04e
Vohilava.....21·04s 48·00e
Vohimarina.....13·21s 50·02e
Vohipeno.....22·22s 47·51e
Vohitsora.....23·54s 44·17e
Vondrozo.....22·49s 47·20e
Zazafotsy.....22·13s 46·26e

Physical features and points of interest

Alaotra, Lac, *lake*.17·30s 48·30e
Ambaro, Baie d', *bay*.....13·23s 48·38e
Ambre, Cap d', *cape*.....11·57s 49·17e
Ambre, Montagne d', *mountain*.....12·30s 49·10e
Ampasilava, Baie d', *bay*.....21·16s 43·43e
Ampasindava, Baie d', *bay*.....13·16s 48·43e
Analamaitso, Plateau d', *upland*.....16·15s 48·15e
Anavelona, *mountains*.....22·37s 44·10e
Andramaimba, Baie, *bay*.....12·15s 48·50e
Andringitra, *mountains*.....22·20s 46·55e
Androna, *ridge*...15·15s 48·40e
Androy, *upland*...25·00s 45·30e
Angavo, *cliff*.....19·40s 47·50e
Ankaratra, *mountains*.....19·25s 47·12e
Ankaroaka, *mountain*.....17·48s 48·32e
Antolana, *mountain*.....17·04s 48·09e
Antongil, Baie d', *bay*.....15·45s 49·50e
Barren, Îles, *islands*.....18·25s 43·40e
Be, Nosy (Nossi-Bé), *island*.....13·20s 48·15e
Beampingaratra, *mountains*.....24·30s 46·50e
Bellone, Cap, *cape*.....16·14s 49·51e
Bemaraha, Plateau du.....20·00s 45·15e
Bemarivo, *river*...14·09s 50·09e
Bemarivo, *river*...15·27s 47·40e
Betsiboka, *river*...16·03s 46·36e
Betsimisaraka, *escarpment*.....18·00s 49·00e
Bezavona, *mountain*.....15·02s 49·52e
Boby, Pic, *peak*...22·12s 46·55e
Bombetoka, Baie de, *bay*....15·50s 46·17e
Bongolava, *cliff*..18·45s 46·00e
Boraha, Nosy, *island*.....16·50s 49·55e
Chesterfield, Île, *island*.....16·20s 43·58e


MADAGASCAR
Elevations in metres
Size of symbol indicates relative size of town
0 100 200 300 km
0 100 200 mi
© Rand McNally & Co.
A-582700-257
MOZAMBIQUE CHANNEL
INDIAN OCEAN
MADAGASCAR
COMOROS
MAYOTTE (France)
Mamoutzou
Dzaoudzi
ÎLE JUAN DE NOVA (Fr.)
ÎLE CHESTERFIELD
ÎLES BARREN
CAP D'AMBRE
Antsiranana (Diego-Suarez)
Montagne d'Ambre 1475
Hell-Ville
Befotaka
Mahajanga (Majunga)
Toamasina (Tamatave)
Antananarivo (Tananarive)
Antsirabe
Fianarantsoa
Toliary (Tulear)
Manakara
Farafangana
Morondava
Maintirano
Marovoay
Tsiroanomandidy
Ambatondrazaka
Andilamena
Antalaha
Sambava
Vohemar
Nosy Varika
Ambositra
Ambatolampy
Miandrivazo
Belo-sur-mer
Morombe
Toliary
Ihosy
Ranohira
Isoanala
Ambalavao
Ambohimahasoa
Ambohimanga du Sud
Mananjary
Vohipeno
Farafangana
Ifanadiana
Manakara
Tropic of Capricorn
PLATEAU DE L'IKAHAVO
PLATEAU DU BONGOLAVA
PLATEAU DU BEMARAHA
PLATEAU MAHAFALY
MASSIF DU MAROMOKOTRO 2876
MASSIF DE L'IVAKOANY
MASSIF RUINIFORME DE L'ISALO
PARC NATIONAL DE L'ISALO
TSARATANANA
ANDRONA
SOFIA
BOENY
IMERINA
ANALAMANGA
VAKINANKARATRA
ITASY
BETSILEO
MATSIATRA
ANOSY
ANDROY
MENABE
BETSIMISARAKA
ANTSIHANAKA
MASOALA
PLATEAU D'ANALAMAITSO
PLATEAU DE L'ANKARATRA
Baie d'Antsiranana
Baie de Bombetoka
Baie de la Mahajamba
Baie d'Ampasindava
Baie de Koraitaika
Baie d'Ampasilava
Baie de Salapaly
Lac Kinkony
Lac Alaotra
Lac Itasy
Lac Ihotry
Lac Tsimanampetsotsa
Mahavavy
Betsiboka
Manambolo
Tsiribihina
Mangoky
Onilahy
Ihosy
Matsiatra
Mananjary
Namorona
Fiherenana
Menarandra
Linta
Imaloto
Manambaho
Ranobe
Ikopa
Sofia
Bemarivo
Maevarano
Loky
Manambato
Daraina
Vohimarina
Milanoa
Tsaratanana
Bealanana
Andapa
Antsohihy
Ambararata
Antsakabary
Bezavona 1484
Manakalampona 1597
Ambilanivelo
Maroantsetra
Ampanavoana
Rantabe
Mandritsara
Antsohe
Manambolosy
Mananara
Marotandrano
Antananbe
Nosy Boraha
Ambodifototra
Soanierana Ivongo
Fenoarivo Atsinanana
Vavatenina
Foulpointe
Ambodiriana
Ambodilazana
Didy
Ivondro
Ambalabe
Maroseranana
Vohibinany
Andevoranto
Vatomandry
Antambao Manampotsy
Mahanoro
Ambila-Lemaintso
Masomeloka
Nosy Varika
Vohilava
Antsenavolo
Sahasinaka
Ambohimanga du Sud
Ambohimahasoa
Alakamisy
Fianarantsoa
Ambohimahamasina
Fort-Carnot
Ankaramena
Ambohimanga
Ambohimahamasina
Ankarimbelo
Vohipeno
Karianga
Tanga-Gony
Ambalatany
Vondrozo
Etrotroka
Ihorombe
Vangaindrano
Lavaraty
Betroka
Ranomena
Midongy Sud
Befotaka
Amboasary
Rancpiso
POINTE ITAPERINA
Faradofay (Fort-Dauphin)
Analapetsa
Ambondro
Ambovombe
Tsihombe
Lavanono
Ankororoka
Faux-Cap
CAP SAINTE-MARIE
Ambilobe
Antsohihy
Befandriana
Maroala
Port-Bergé
Bandabe
Mampikony
Marovato
Mahajanga
Kandreho
Andriba
Maevatanana
Ambalamanakomby
Tsaratanana
Miarinarivo
Soalala
Ambinda
Madirovalo
Ambato Boeni
Sitampiky
Besalampy
Ankasakasa
Amparihy
Bekodoka
Marovoay Nord
Bemariyo
Mahabe
Antsalova
Soaloka
Folakara
Tsianaloka
Beravina
Morafenobe
Bemolanga
Ambatomainty
Tambohorano
Berevo
Bebao
Soahanina
Masoarivo
Trangahy
Bekopaka
Kinangaly
Fiantsonana
Mahasoloa
Ankavandra
Ankazomiriotra
Belo
Miandrivazo
Ranohira
Bemaraha
Ankilizato
Befasy
Malaimbandy
Morondava
Mahabo
Andranovory
Beraketa
Bemaraha
Itandrano
Ambohikily
Manja
Ambivy
Befotaka
Marerano
Andranopasy
Antevamena
Mandabe
Mandronarivo
Tsitondroina
Mahatsinjo
Beroroha
Bekisopa
Benavo
Ambohimahasoa
Alakamisy
Fanjakana
Solila
Ambohimanga
Ambohimahamasina
Fenoarivo
Ambohimahasoa
Alarobia Vohiposa
Vohilava
Ifanadiana
Mananjary
Antsenavolo
Tolongoina
Manampatrana
Sahasinaka
Ambohimahamasina
Fort-Carnot
Ankaramena
Ambohimanga
Ankarimbelo
Manakara
Vohipeno
Karianga
Tanga-Gony
Ambalatany
Vondrozo
Etrotroka
Farafangana
Ihorombe
Vangaindrano
Lavaraty
Betroka
Ranomena
Midongy Sud
Befotaka
Amboasary
Rancpiso
POINTE ITAPERINA
Faradofay (Fort-Dauphin)
Analapetsa
Ambondro
Ambovombe
Tsihombe
Lavanono
Ankororoka
Faux-Cap
CAP SAINTE-MARIE
Ambilobe
Antsohihy
Befandriana
Maroala
Port-Bergé
Bandabe
Mampikony
Marovato
Mahajanga
Kandreho
Andriba
Maevatanana
Ambalamanakomby
Tsaratanana
Miarinarivo
Soalala
Ambinda
Madirovalo
Ambato Boeni
Sitampiky
Besalampy
Ankasakasa
Amparihy
Bekodoka
Marovoay Nord
Bemariyo
Mahabe
Antsalova
Soaloka
Folakara
Tsianaloka
Beravina
Morafenobe
Bemolanga
Ambatomainty
Tambohorano
Berevo
Bebao
Soahanina
Masoarivo
Trangahy
Bekopaka
Kinangaly
Fiantsonana
Mahasoloa
Ankavandra
Ankazomiriotra
Belo
Miandrivazo
Ranohira
Bemaraha
Ankilizato
Befasy
Malaimbandy
Morondava
Mahabo
Andranovory
Beraketa
Bemaraha
Itandrano
Ambohikily
Manja
Ambivy
Befotaka
Marerano
Andranopasy
Antevamena
Mandabe
Mandronarivo
Tsitondroina
Mahatsinjo
Beroroha
Bekisopa
Benavo
Ambohimahasoa
Alakamisy
Fianarantsoa
Ambohimahamasina
Fenoarivo
Ambohimahasoa
Alarobia Vohiposa
Vohilava
Ifanadiana
Mananjary
Antsenavolo
Tolongoina
Manampatrana
Sahasinaka
Ambohimahamasina
Fort-Carnot
Ankaramena
Ambohimanga
Ankarimbelo
Manakara
Vohipeno
Karianga
Tanga-Gony
Ambalatany
Vondrozo
Etrotroka
Farafangana
Ihorombe
Vangaindrano
Lavaraty
Betroka
Ranomena
Midongy Sud
Befotaka
Amboasary
Rancpiso
POINTE ITAPERINA
Faradofay (Fort-Dauphin)
Analapetsa
Ambondro
Ambovombe
Tsihombe
Lavanono
Ankororoka
Faux-Cap
CAP SAINTE-MARIE

MAP INDEX (continued)

Est, Cap, *cape*....15·16s 50·29e
Faraony, *river*....21·47s 48·10e
Fiherenana, *river*.23·19s 43·37e
Ihosy, *river*.....21·58s 43·38e
Ihotry, Lac, *lake*..21·56s 43·41e
Ikahavo, Plateau de l'.....17·25s 45·50e
Ikepa, *river*.....17·01s 46·45e
Imaloto, *river*....23·27s 45·13e
Imerina, *physical region*..18·30s 46·50e
Indian Ocean....22·00s 50·00e
Ionaivo, *river*.....22·56s 46·54e
Isalo, Parc, National de l', *national park*....22·45s 45·15e
Itaperina, Pointe, *point*.....24·59s 47·06e
Itasy, Lac, *lake*.....19·04s 46·17e
Ivakoany, Massif de l', *mountains*.23·50s 46·25e
Ivondro, *river*.....18·15s 49·22e
Kinangaly, *mountain*.....19·12s 45·40e
Kinkony, Lac, *lake*.....16·08s 45·50e
Koraraika, Baie de, *bay*.....17·45s 43·57e
Lava, Nosy, *island*.....14·33s 47·36e
Linta, *river*.....25·02s 44·05e
Maevarano, *river*.14·35s 47·58e
Mahafaly, Plateau, *upland*.....24·20s 44·10e
Mahajamba, *river*.....15·33s 47·08e
Mahajamba, Baie de la, *bay*.....15·24s 47·05e
Mahajilo, *river*...19·42s 45·22e
Mahavavy, *river*..13·00s 48·55e
Mahavavy, *river*..15·57s 45·54e
Makay, Massif du, *mountains*...21·15s 45·15e
Manakalampona, *mountain*.....15·23s 48·50e
Manambaho, *river*.....17·42s 44·03e
Manambolo, *river*.....19·18s 44·24e
Mananantanana, *river*.....21·25s 45·33e
Mananara, *river*..23·21s 47·42e
Mandrare, *river*..25·10s 46·27e
Mangoky, *river*...21·29s 43·41e
Mangoky, *river*...23·27s 45·13e
Mangoro, *river*...20·00s 48·45e
Mania, *river*.....19·42s 45·22e
Maningory, *river*..17·13s 49·28e
Maromokotro, *mountain*.....14·01s 48·58e
Masoala, Cap, *cape*.....15·59s 50·13e
Masoala, Presqu'île, *peninsula*...15·40s 50·12e
Matsiatra, *river*...21·25s 45·33e
Menarandra, *river*.....25·17s 44·30e
Minirodo, Baie de, *bay*.....25·10s 44·15e
Mitsio, Nosy, *island*.....12·54s 48·36e
Mozambique Channel.....16·00s 44·00e
Narinda, Baie de, *bay*.....14·55s 47·30e
Nossi-Bé, see Be, Nosy
Onilahy, *river*....23·34s 43·45e
Pangalanes, Canal des, *canal*.....22·40s 47·50e
Radama, Îles, *islands*.....14·00s 47·47e
Radama, Presqu'île, *peninsula*...14·16s 47·53e
Ramanetaka, Baie, *bay*.....14·13s 47·52e
Ranobe, *river*....17·10s 44·08e
Ruinforme de l'Isalo, Massif, *mountains*.....22·45s 45·15e
Saint-André, Cap, *cape*.....16·11s 44·27e
Sainte-Marie, Cap, *cape*.....25·36s 45·08e
Saint-Sébastien, Cap, *cape*.....12·26s 48·44e
Saint-Vincent, Cap, *cape*.....21·57s 43·16e
Salapaly, Baie de, *bay*.....24·40s 43·55e
Sofia, *river*.....15·27s 47·23e
Tsaratanana, Massif du, *mountains*.....14·00s 49·00e
Tsiafajavona, *mountain*.....19·21s 47·15e
Tsimanampetsotsa, Lac, *lake*...24·08s 43·46e
Tsiribihina, *river*.19·42s 44·31e
Vatoloha, *mountain*.....17·52s 47·48e
Vazobe, *mountain*.....18·25s 47·18e

sharks, sardines, whiting, crayfish, crabs, shrimp, mussels, and oysters. The spiny globefish and cofferfish species are peculiar to Madagascar. The coelacanth, referred to as a living fossil and once thought extinct for millions of years, inhabits offshore waters.

Settlement patterns. Despite the importance of intensive rice cultivation, the land is primarily used for pastoral purposes. Cattle are kept in all parts of the island and are more numerous than the human population; they are fewer in the dense forest areas of the eastern escarpment. Elsewhere, pastoralism predominates, although the Malagasy people also cultivate subsistence crops. On the plateau, the valley floors and irrigable slopes are mainly used for growing rice. The forest peoples traditionally grew hill rice, after cutting and burning the forest; this practice continues, although it is discouraged by the government, which promotes the establishment of permanent irrigated rice fields.

The older villages of the Merina and Betsileo were often perched on hilltops and defended by huge ditches. Today, villages have been rebuilt on lower ground, and hamlets and homesteads are scattered over the landscape.

On the plateau, cattle enclosures are built of dry stone walls; the landscape is also dotted with funerary monuments; in parts of the south and west, these take the form of beautifully carved wooden posts.

Small towns began to develop at the capital centres of the island's several kingdoms at least by the 18th century. The most populous cities are Antananarivo, in the central plateau; Mahajanga (formerly Majunga), on the northwest coast; Fianarantsoa, in the southern plateau; Toamasina (formerly Tamatave), on the east coast; Antsiranana, in the north; Toliary, in the southwest; and Antsirabe, south of Antananarivo, mainly a tourist centre.

Racial or economic separation is not noticeable in the older cities such as Antananarivo and Fianarantsoa, but the newer towns are often divided into socioeconomic sectors. Because of recent internal migration, most of the cities are composed of a mixture of ethnic groups.

The people. *Population groups.* Madagascar has been inhabited by human beings for the relatively short period of only about 2,000 years. Language and culture point unequivocally to Indonesian origins, but there is no empirical evidence of how, why, or by what route the first settlers came to the island. Studies of the winds and currents of the Indian Ocean indicate that the voyage from Indonesia could have been made. It is assumed that the original peopling of the island, however sparse, was accomplished by a single cultural group, probably as the result of a single voyage.

Indonesian origins

There is also widespread evidence from linguistics, archaeology, and tradition of influence from Afro-Arab settlers on the coasts before AD 1000. There is slighter evidence of an Indian influence in vocabulary, but there is no trace of Hinduism in Malagasy culture and of the orthodox Islām only in later coastal settlements.

The inhabitants of Madagascar speak Malagasy, which, written in the Latin alphabet, is a standardized version of Merina, an Austronesian language. Although there are numerous local variations of Malagasy, they are all mutually intelligible. French is widely spoken and is also officially recognized. It is used as a medium of instruction, as is Malagasy.

The population is divided into about 20 ethnic groups, the largest and most dominant of which is the Merina people, who are scattered throughout the island. The name Merina (Imerina) is said to mean Elevated People, deriving from the fact that they lived on the plateau. The second largest group is the Betsimisaraka (The Inseparable Multitude), who live generally in the east. The third most numerous group is the Betsileo (The Invincible Multitude), who inhabit the plateau around Fianarantsoa. Other important peoples are the Tsimihety (Those Who Do Not Cut Their Hair); the Sakalava (People of the Long Valley); the Antandroy (People of the Thorn Bush); the Tanala (People of the Forest); the Antaimoro (People of the Banks); and the Bara (a name of uncertain origin). Smaller groups are the Antanosy (People of the Island); the Antaifasy (People of the Sand); the Sihanaka (People of the Lake); the Antakarana (People of the Rocks); the Betanimena (People of the Red Soil), who are now largely absorbed by the Merina; the Bezanozano (Those with Many Braided Hair); and the Mahafaly (the Joyful People). These ethnic names do not stand for clear-cut cultural boundaries, for in many cases one group fades imperceptibly into another. Moreover, the conventional translations are by no means reliable and most of the names themselves are of somewhat recent origin, probably crystallized and rigidified by the exigencies of colonial administration more than by the realities of indigenous culture. In no sense are these groups to be regarded as "tribes," a concept now considered invalid, nor are they composed of clans, but rather, in most cases, of endogamous and often non-unilinear descent groups.

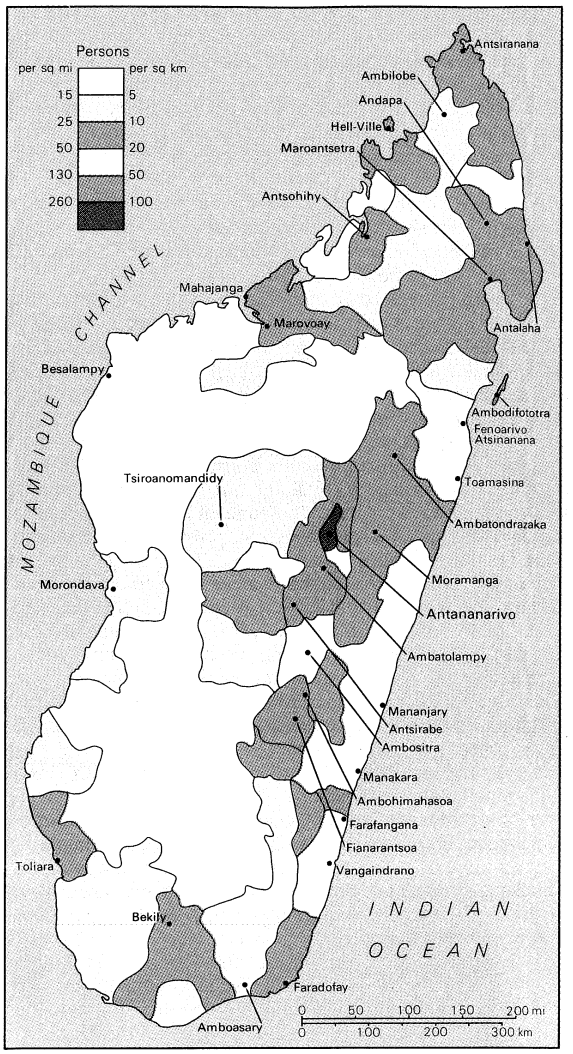
Around half of the population has been converted to Christianity, and the Christian community is about equally divided between Protestantism and Roman Catholicism. There is a sizable community of Muslims in the northwest.

The rest of the people continue to practice their traditional religion, which is based upon ancestor worship. The dead are buried in tombs and are believed to reward or punish the living. There is a supreme being called Zanahary (the Creator) or Andriamanitra (the Fragrant One). There is also a belief in local spirits, and a complex system of taboos constrains Malagasy life.

Traditional religion

Demography. More than 95 percent of the population is Malagasy. The major foreign communities are French, Comorian, Indian, and Chinese. Births greatly outnumber deaths. Registration of births and deaths is, however, incomplete. Government policy opposes any form of population control.

Emigration of the French, Comorian, Indian, and Chinese in the late 20th century has significantly reduced



Population density of Madagascar.

their populations. There has been, however, no significant emigration of local peoples abroad.

Population densities

The eastern part of the central plateau is the region of highest population density, and the eastern coastal plain has the second highest density. The eastern forest zone and the northeastern coast densities vary but rank as the next most densely populated regions. Most of the western two-thirds of the country is sparsely inhabited.

The eastern half of the island contains almost all the major cities and towns. Antananarivo is the most populous city; perched on two precipitous mountain ridges, the old part of the city is dominated by the Manjakamiadana palace and has an extremely picturesque, almost medieval appearance.

The economy. The control of Madagascar's economy by France, which had taken nearly half of the island's exports and supplied more than half of its imports, was drastically reduced following changes in government during 1972–75. Currency, banking, finance, loans, and economic planning had been influenced not only by accords with France, but also by French personnel in government, commerce, and technical assistance. Madagascar, after 1972–75, received aid from new sources, including the Soviet Union, the People's Republic of China, and North Korea, but the greater part of the country's aid continued to come from France, other Western countries, and Japan.

Resources. The Great Red Island disappointed those who hoped to discover precious metals in large quantities. Considerable small-scale gold mining was conducted toward the end of the 19th century, both by French and Malagasy prospectors. There is a wide variety of semiprecious stones, but deposits are not of significant economic importance. Mineral deposits include chromite, which is found north of Antananarivo and near Toamasina; ilmenite, found on the southeast coast and near Fenoarivo Atsinanana (former Fénérive Est); low-grade iron, found in scattered deposits in the southern half of the island; low-grade coal near Antsirabe, north of Toliary and inland from Besalampy; nickel near Fianarantsoa; and copper north of Ampanihy and near Ambilobe. There are also smaller deposits of zircon, monazite, bauxite, lead, graphite, quartzite, jasper, gold, uranothorianite, bentonite, kaolin, and alunite. The various gems include garnet, amethyst, beryl, and columbite.

Principal mineral deposits

The eastern evergreen forest and the remnants of the deciduous forest are valuable for their timber. Except for the fish and crustaceans of the rivers and coasts, the indigenous animal life is of little economic importance. Domestic animals include the humped zebu (cattle of Asiatic derivation, doubtless brought from Africa), of great religious importance; fat-tailed sheep; dogs; poultry; and pigs.

Although there are many magnificent waterfalls, especially on the eastern escarpment, only a small number of them have been developed. Hydroelectric power stations serve the cities of Antananarivo, Toamasina, Antsirabe, and Fianarantsoa and are supplemented by thermal stations, which burn coal. There are also thermal stations serving Mahajanga, Toliary, Antsiranana, Ambositra, Manakara, and Morondava. Bituminous shales were discovered at Bemolanga and oil at Tsimiroro.

Agriculture, forestry, and fishing. Rice occupies the largest share of the total crop acreage. Many varieties of dry, wet, and irrigated rice are grown in the central plateau; dry rice is also grown in the eastern forests and wet rice in the estuaries, mainly by populations migrated from overpopulated parts of the plateau. Costly imports are still required.

Slash-and-burn techniques (the temporary clearance of land for agriculture) are used in the escarpment forest and along the east coast. In the river valleys of the west, cultivation is permanent; irrigation and rainfall are used.

Sugarcane is grown on plantations in the northwest, around Mahajanga, and on the east coast near Toamasina. Manioc is grown all over the island, and potatoes and yams are grown mainly in the highland region of Ankaratra. Bananas are produced commercially on the east coast, and corn (maize) is grown mainly in the central plateau, in the south, and in the west. Fruits include apples, grapefruits, avocado pears, plums, grapes, oranges, litchis, pineapples, guavas, pawpaws, passion fruits, and bananas. Robusta coffee is grown on the east coast and arabica coffee on the plateau. Other significant crops are beans, peanuts (groundnuts), *pois du cap* (lima beans), coconuts, pepper, vanilla, cacao, sisal, raffia, tobacco, copra, cotton, and castor beans.

Agricultural produce

Cattle (mainly zebu) are distributed throughout the island. The large numbers of pigs, sheep, goats, chickens, ducks, geese, and turkeys are found mainly on the plateau. The hoarding of cattle as a sign of wealth and for religious sacrifice has frustrated government efforts to increase the use of cattle for domestic meat consumption and for export.

A significant area of the forest is degraded (*i.e.*, regenerated after repeated burnings, with many original species lost and smaller, fewer, less valuable species becoming dominant); the rest is wet or dry tropical forest. Major reforestation has taken place, with mainly eucalyptus, pine, and cypress trees planted.

The fishing industry

Fisheries are poorly developed and depend mostly on the small traditional fishing communities of Vezo on the west coast; production is badly adjusted to the island's needs and to the potential market. In the late 20th century, however, the industrialization of marine fishing was begun and rivers, lakes, and irrigated rice fields were stocked with breeding fish. The inshore waters are fished, while the distant offshore waters are neglected; inland distribution of the fish caught is poor or nonexistent, except in main towns. The bulk of the catch is composed of fresh fish and crustaceans; some fish are dried. There is also considerable raising of fish in the irrigated rice fields, mainly for home consumption.

Industry. Mining and quarrying are little developed.

The most important products are graphite, iron ore, bauxite, mica, monazite, quartz, garnet, amethyst, salt, and chromite. Less valuable amounts of gold, tourmaline, citrine, beryl, ilmenite, columbite, zircon, and jasper are produced as well. Merina jewellers polish and set semiprecious stones at small workshops in most of the towns of the plateau.

The country's manufacturing consists mainly of rice mills and small industrial establishments. More than half of the manufacturing is located in the province of Antananarivo. Products include wood, paper pulp, cotton cloth, fertilizer, oils, soap, sugar, cigarettes and tobacco, sisal rope and mats, bricks, processed foods, and beverages. A small printing industry and a motor-vehicle assembly plant are also located there.

Electricity is generated by means of thermal and hydroelectric power plants. Most of the hydroelectric plants are located on the central plateau and in the neighbouring eastern escarpment.

Finance. The Malagasy franc replaced the CFA (Communauté Financière Africaine) franc in 1963, and in 1973 Madagascar officially left the franc zone. The Central Bank issues currency.

Trade. The majority of the value of exports derives from vegetable products, of which coffee, cloves and clove oil, vanilla, sisal, raffia, sugar, rice, pepper, tobacco, and peanuts are the most important. Petroleum products are also exported. Although exports are more varied than those of other African countries, a number of them, such as coffee, cloves, vanilla, and sisal, are endangered by world overproduction or by the manufacture of synthetic substitutes.

The most important imports are mineral products. Imports include crude petroleum, chemical products, metal products, machinery, vehicles, electrical equipment, cotton textiles, and essential foods.

Before 1972, the majority of Madagascar's trade was with France. After 1972, trade with France fell sharply, although France remains Madagascar's main trading partner. Other important trading partners include the United States, West Germany, Japan, Italy, Réunion, Malaysia, the United Kingdom, and The Netherlands.

Administration of the economy. The government, promoted and supported by France, kept a free capitalist economy until 1972. After the turbulence of 1972–75 and the emergence of the military regime, a long-term plan was announced to achieve "a socialist paradise under divine protection" by the year 2000. Under this plan the government has the majority of direct control of the economy. The five French Banks were nationalized and transformed into three state banks for agriculture, industry, and trade, and the foreign insurance companies were turned into two state corporations. The great French import-export trading and shipping firms were nationalized, as were the textile, cotton, and power industries and the dockyards at Antsiranana. State corporations were created for petroleum production; agricultural marketing; the extension of irrigated agriculture at Lac Alaotra, in the Mangoky valley, and in Fianarantsoa province; and the regional development of Mondraya.

Economic planning

Before 1972, the government had established producers' cooperatives, which collected and processed most of the rice crop (at prices that were bitterly resented by the peasants); state farms intended to increase the commercial production of rice, cattle, coffee, oil palms, cotton, and silk; a rural development program, the Animation Rurale; and a national consumers' cooperative with retail shops located in most towns. Both external and internal private investment continued to be officially welcomed, however, if reluctantly.

Taxation is mainly indirect and is largely derived from various customs, import, and export duties. The various excise taxes are also important. Direct taxes take the form of taxes on company income, registration fees, stamp duties, and personal taxes; the latter bear heavily on the peasants.

Most union members hold office jobs or work in industry. A popular and extensive trade union is the Union des Syndicats Autonomes de Madagascar. The large membership is held by the Fivondrononam Ben'ny Sendika Kristianina Malagasy (Christian Confederation of Malagasy Trade Unions). The main employers' association is the Union des Syndicats d'Intérêt Économique de Madagascar.

Transportation. Transport facilities primarily serve the plateau and the east coast. Facilities are rudimentary on the western half of the island, although the country's best natural harbours are located there.

The majority of roads are unpaved. Roads down the eastern escarpment and across the western coastal strip, as well as minor roads everywhere, become impassable during the wet season.

The main paved road runs south from Antananarivo to Fianarantsoa, where it branches southwest to Toliary, southeast to Faradofay via Ihosy, and east to Mananjary and Manakara. Paved roads run east from Antananarivo toward Toamasina, west to Analavory, and north to Mahajanga and Antsiranana.

Two railways connect the plateau with the east coast; they run from Antananarivo to Toamasina and from Fianarantsoa to Manakara. Plateau routes run from Antananarivo south to Antsirabe, to Fianarantsoa, and north to Ambatondrazaka. A southwestern railway runs between Soalary and Vohitsara.

The main port is Toamasina, which has a fine deepwater harbour, equipped with quay berths and directly linked to Antananarivo by rail, air service, and road. Mahajanga is second in importance but is only accessible to small ships of shallow draft; it has considerable dhow traffic with the Comoro Islands. Antsiranana has one of the finest natural harbours in the world but is as yet too remote from the main centres of economic activity; it contains the former French naval base, arsenal, and dry dock and also has a small commercial port. Faradofay also has a small port, and there are small wharves and minor facilities at Toliary, Manakara, Nosy Be, Morondava, Analalava, and Vohimavina. The coastal lagoons and swamps of the Canal des Pangalanes on the east coast, which are linked by artificial channels where necessary, provide a waterway that is 434 miles long.

Airfields are found throughout the island. The main international airport is at Ivato, near Antananarivo, and some international flights make secondary landings at Toamasina, Nosy Be, and Mahajanga. Air Madagascar provides internal services; its international flights are supplemented by several foreign airlines.

Administrative and social conditions. *Government.* The Malagasy Republic became independent in 1960, having been an autonomous republic in the French Community from 1958. There was a president and a bicameral legislature consisting of the National Assembly, directly elected by universal adult suffrage, and the Senate, one-third of which was nominated by the president and the rest of which was indirectly elected. Support for Pres. Philibert Tsiranana's Parti Social Démocrate (PSD; Social Democratic Party) was strongly promoted by the administration, but discontent erupted sporadically in local uprisings, and the devastating burning of trees cancelled otherwise promising reforestation. In 1972, mass demonstrations of students and workers in Antananarivo, Toamasina, and other towns led the aging president to hand over power to the army chief of staff, Maj. Gen. Gabriel Ramanantsoa. After the fall of three military governments, Lieut. Comdr. Didier Ratsiraka, then the foreign minister, took power in 1975, determined upon radical political and economic reorganization.

The turbulence of 1972–75

A new constitution was implemented in 1975 for the renamed Repoblika Demokratika Malagasy (République Démocratique de Madagascar), despite suspected plots and violent outbreaks, including the massacre of Comorians in Mahajanga, said to have caused the death of almost 1,400 and the departure of another 16,000 to their overpopulated islands. The new constitution established a presidency; the Supreme Council of the Revolution, a body of members including the president, prime minister, military chiefs, some ministers, and other nominated leaders; a Cabinet, composed of the prime minister and other ministers; the People's National Assembly, composed of deputies directly elected for five years; a Military Development Committee,

which advises the government on defense and economic development; and a Constitutional High Court.

There is a layered local electoral system. In this system, many thousands of local village communities (*fokonolona*) elect delegates to associations of village councils (*firaisampokontany*), which then send delegates to district groupings (*fivondrynampokontany*), which, in turn, send delegates to six provinces (*faritany*). This system puts power into the hands of party delegates at all levels, under supreme military oversight, away from the former bureaucratic officials that existed at all levels before the reorganization.

The constitution of 1975 called for the establishment of the Andrim-Pirenena amin'ny fiarovana ny Tolompiavotana, or the National Front for the Defense of the Revolution. Within the National Front, several parties participate, although one party, the Avant-garde de la Révolution Malgache (Arema), has dominated. Almost all political factions supported this party's political manifesto, known as the "Boky Mena" ("Red Book"), or *Charte de la Révolution Malgache Tous Azimuths.* The old polarization of coastal and plateau peoples has diminished with the changes in the government, although the Merina have continued to dominate the civil service, professions, and education.

Justice. The former Merina state that ruled the island throughout the 19th century had an elaborate system of laws, courts, and justice. The present Malagasy legal system, however, is based upon French codes and practices, and most judges and magistrates have had French training. There is a Constitutional High Court, an appeal court, criminal tribunals, and tribunals of first instance; there is also a criminal code, a code of criminal procedure, and a code of civil procedure. The customary law of the Merina and other ethnic groups is taken into account by state magistrates when judging marriage, family, land, and inheritance cases.

The armed forces. The army played no direct political role in Madagascar until 1972, although the presence of French Army units had bolstered the former government. The French troops were withdrawn in 1973, and the French naval base at Antsiranana was handed over to Madagascar in 1975. The Malagasy armed forces consist of an army, a navy, and an air force, and there is a paramilitary force.

Detachments of police are stationed at the headquarters of each admnistrative division, as well as in Antananarivo. They have general responsibility for keeping the peace, enforcing laws, and apprehending criminals. The whole force is under unified command and falls within the responsibility of the ministry of the interior.

Education. The educational system consists of primary and secondary schools, technical institutes, teacher-training colleges, and a university system. Enrollment at the University of Madagascar and its five regional branches has increased dramatically in the late 20th century. There has been an increased use of the Malagasy language in teaching, although some coastal peoples have objected to that language's close relationship with Merina. Most French professors have been replaced, some by teachers from the Soviet Union.

The role of mission schools

The level both of school attendance and of educational attainment is higher on the plateau than in the coastal areas. Protestant and Roman Catholic missions have been providing education since the 19th century, and the missions still educate a majority of the school children, although the government now maintains official schools at all levels. In the main towns there are also privately run schools, catering to those unable to enter either government or mission schools. The majority of illiterate persons are female.

Health and welfare. Malagasy doctors began to practice Western medicine in 1880; and a medical school was established in Antananarivo. The health system includes principal and secondary hospitals, dispensaries, and medical centres. Medical personnel includes doctors, as well as pharmacists, dentists, midwives, social assistants, visiting nurses, and health assistants.

Hospitals and specialists are mainly in the towns, apart from some rural hospitals run by Christian missions. Health insurance and other social benefits are mainly available to better paid and professional elements among the employed population.

The extension of health services is largely credited for the steady population increase. Infant mortality remains high, but infant deaths from malaria, which is endemic all over the island, have been cut by half. Debilitating parasitic diseases, such as schistosomiasis, an infection of the bladder or intestines, remain serious and are hard to control since their breeding grounds are the irrigated rice fields and the streams that feed them. Veneral disease is also widespread, especially in its incipient form. Rural diet is often deficient in protein and some vitamins.

Housing. Houses are typically rectangular and are usually crowned with steeply angled roofs. In the rural areas, most houses are made of either mud and wattle or woven matting supported by poles. In the eastern forest they are built of interlaced split bamboo and are thatched with palm, while in the south, overlapping upright wooden planks are used for the walls. In the plateau, rural housing is constructed of earth blocks and thatched roofing, while upper income and most urban housing consists of two- or three-story homes—typically with kitchen at the top, living quarters in the middle, and storage below—all surrounded by wide balconies supported by brick columns and crowned with steep tiled roofs. This is the lofty Indonesian style of architecture, transformed by new techniques contributed by the missionaries. The original style survives in the house of Andrianampoinimerina (reigned 1787–1810) at Ambohimanga and reaches its apotheosis in the Queen's palace built by Jean Laborde in the 19th century.

The French colonial government provided housing for the more senior civil servants, and the Malagasy government has continued the practice. The government-sponsored housing authority conducts research into design, materials, and methods of production and is endeavouring to promote inexpensive urban housing, but the problem of overcrowding is expected to increase with continued urban growth. The existence of a well-established craft of house construction, however, may successfully alleviate housing pressures without resorting to the use of imported materials or relying upon foreign enterprise.

Three traditional castes

Social and economic divisions. Traditionally, society was divided into three castes—the nobles, the freemen, and the former slaves and their descendants. These social distinctions are no longer strict and are manifest only on ceremonial occasions, such as weddings and funerals. They do, however, form the basis of other economic and social distinctions. During the 19th century, the Merina elite conquered the island, established themselves as rulers, and adopted Protestant Christianity; in the late 1800s, some became Roman Catholics. Under French rule in the 20th century, the Merina retained their supremacy in education, business, and the professions, while the remainder of the population retained its sense of "difference" from the dominant peoples and some adopted Roman Catholicism.

A further distinction is made between the peoples of the plateau and those of the coast, who are called *cotiers.* The coastal peoples feel deprived of the education, power, and wealth that is concentrated on the plateau. Since independence, the government has been composed of *cotiers,* and a conscious effort has been made to keep the Merina elite of the plateau from power.

Cultural life. *The cultural milieu.* The culture is basically Indonesian. Arabic and Islāmic contributions include an intricate system of divination, or *sikidy,* and calendrical features, such as the names of the days of the week, which are particularly significant in their application to markets. The coastal areas of the west, north, and south might be expected to show African cultural elements, but, apart from some Bantu words, these are often difficult to identify conclusively.

The state of the arts. The conquest of the plateau peoples by the French and their assimilation of Western values have deprived them of most of their traditional institutions. In music, however, Western dance and musical instruments have been adapted to Malagasy rhythms. The tube zither, the conch, and the cone drum are of Indone-

sian origin, while other types of drums and animal horns suggest African influence. Folk music has been retained, but much of the singing is of Western church hymns and chants adapted to the distinctive Malagasy musical style.

Social and religious life on the plateau centres upon the church congregation, and the cultural emphasis on ancestral tombs is now largely expressed in Christian terms. More time, money, and care are spent on building tombs than houses. The dead are always brought back to their ancestral tombs, however long or far away they have spent their lives. Tombs are opened every few years, the corpses taken out and carried in procession with much ceremony, then replaced after being rewrapped in new shrouds, which are still woven from locally produced silk, coloured with natural, herbal dyes. The male peasants wear cloth trousers with tunics reaching to the knees. Women wear cloth dresses but wrap a silk cloth under one arm and over the other shoulder, even when wearing Western fashions.

The cult of the dead

The coastal peoples have retained more of their traditional customs. Funeral practices are similar to those of the plateau, with local variations of detail. In the eastern forests men wear shorts rather than trousers, and many still wear the short tunic that is woven from raffia fibres. In the far south some older men wear a homespun silk cloth that is wrapped around the waist and between the legs, but most have adopted imported cotton clothes.

The Mahafaly have a remarkable wood-carving industry, and their tombs of coloured stones and carved wooden posts are the most beautiful on the island. The Betsileo also have a thriving wood-carving industry, making inlaid furniture of valuable hardwoods. They also produce ornamental cloths of very finely woven raffia and have become specialists in the production of coloured straw hats. Betsileo and Merina women have become experts in French-style embroidery, sewing, and dressmaking.

Handicrafts

The Malagasy language is very rich in proverbs, and there is now an extensive written literature including poetry, legend, history, and scholarly works, as well as contemporary themes. Literary production is aided by an excellent printing industry, for which the Merina have shown a flair since learning it from the London Missionary Society in the 1820s. The peoples of the southeast still preserve their manuscripts in Arabic script with great reverence; few can be more than 200 years old, although some may be copies of much earlier manuscripts.

Cultural institutions. The ministries of cultural affairs and of traditional arts encourage the blending of old and new cultural expressions, and a number of new seasonal festivals have been promoted, including the Festival of Rice, the Festival of the Trees, the Festival of the Party, and Independence Day. Towns, churches, schools, and private groups hold concerts or dances, and in the cities there are cultural associations based on the members' home districts.

The main libraries and museums, located in Antananarivo, include the Bibliothèque Nationale, the Bibliothèque Municipale, and the Archive Nationale. There are also the library of the Académie Malgache, the university library, and the university museum. There are museum collections of Malagasy culture and archaeology; natural science collections include a zoo with lemurs and other animals specific to Madagascar.

The press and broadcasting. There are daily and other newspapers published in French and Malagasy, and a government gazette is also published. The island receives radio, television, and telephone service.

For statistical data on the land and people of Madagascar, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Ai.S.)

HISTORY

Although the Malagasy peoples are of mixed African and Indonesian ancestry, their language is basically Indonesian, with a few words of Bantu. The Indonesian element probably migrated to Madagascar during the 1st millennium AD; one hypothesis is that they came by way of the African coast and reached the island after mingling with Africans. Archaeological evidence indicates that these people were living on the island in the 10th century.

Madagascar to 1810. After the 14th century, small groups of Muslims, coming perhaps from the eastern coast of Africa, set up trading colonies in the north; others, mingling with Africans, founded in the southeast the kingdoms of Zafi-Raminia and Antemoro. The settlement of the Indonesian Merina (Imerina) in the central-plateau region probably occurred at the same time.

Early European contacts. Madagascar is mentioned in the writings of Marco Polo, but the first European known to have visited the island is Diogo Dias, a Portuguese navigator, in 1500. It was called the Isle of St. Lawrence by the Portuguese, who frequently raided Madagascar during the 16th century, attempting to destroy the Muslim settlements. Other European nations also moved in; in 1643 the French established Fort-Dauphin in the southeast and maintained it until 1674. One of their governors, Étienne de Flacourt, wrote the first substantial description of the island. In the late 17th and early 18th centuries, Madagascar was frequented by European pirates (among them Capt. William Kidd) who preyed upon shipping in the Indian Ocean.

In the 18th century, the Mascareigne Islands to the east were colonized by the French with the help of Malagasy slaves. Two attempts at fortified settlements failed, one at Fort-Dauphin by the Comte de Modave, the other at Antongil Bay by Baron Benyowski; however, French trading settlements prospered, notably at Tamatave.

The kingdoms. In Madagascar as on the African mainland, new states arose between 1500 and 1800, and many of them engaged in the traffic in slaves and firearms with the Europeans. The traditional clan organization of the Malagasy gave way to a number of autocratic kingdoms. In the 17th century, the southern Sakalava state spread northward into the sparsely populated plains of the far west. Antemoro chieftains founded kingdoms in the interior, among the Tanala and the Betsileo peoples. In the 18th century, the Betsimisaraka kingdom was founded on the east coast by Ratsimilaho, a half-English Malagasy.

The Merina kingdom was founded toward the end of the 16th century in the swampy Ikopa Valley in the centre of the island. Antananarivo (Tananarive) became its capital. In the 18th century Merina was divided among four warring kings. One of them, Andrianampoinimerina, who reigned 1787–1810, reunited the kingdom about 1797. He gave it uniform laws and administration, and sold slaves to the French on the coast, using the guns he got in return to conquer his neighbours, the Betsileo. Under Andrianampoinimerina, Merina society was divided into a ruling noble class (*Andriana*), a class of freemen (*Hova*), and a slave group (*Andevo*). The name Hova was sometimes used to refer to all Merinas. At Andrianampoinimerina's death, he left to his son Radama a single political ambition: "The sea will be the boundary of my rice field" (*i.e.*, of his kingdom).

The kingdom of Madagascar. *Formation of the kingdom (1810–61).* Radama I (1810–28) allied himself with Sir Robert Farquhar, the British governor of the nearby island of Mauritius. In order to prevent reoccupation of the east coast by the French, Farquhar supported Radama's annexation of the area by supplying him with weapons and advisers and giving him the title "king of Madagascar." At the same time, Radama agreed to cooperate with Britain's new campaign to end the slave trade. In 1817 he took Tamatave and then launched yearly expeditions against the coastal populations. He thus conquered almost the entire east coast, the northern part of the island, and most of the two large Sakalava kingdoms. Only the south and a part of the west remained independent. The French retained no more than the small island of Sainte-Marie. In addition, Radama invited in European workmen; the London Missionary Society spread Christianity and influenced the adoption of a Latin alphabet for the Malagasy language. Radama died prematurely in 1828; he was succeeded by his widow, Ranavalona I, who reversed his policy of Europeanization. She expelled Christian missionaries and persecuted Malagasy converts. A very few Europeans, notably the Frenchman Jean Laborde, maintained external trade and local manufacture, but eventually they also were expelled. The British and French launched an expedition

Merina expansion

against Ranavalona, but were repulsed at Tamatave in 1845. By the time of her death (1861), Madagascar was isolated from European influence.

Outside influences (1861–95). Ranavalona was succeeded by her son, Radama II, who readmitted the foreigners. English Protestants and French Roman Catholics vied for supremacy; businessmen obtained excessive concessions. This policy led to Radama's overthrow by the Merina oligarchy (1863). The head of the army, Rainilaiarivony, a Hova, became prime minister and remained in power by marrying three queens in succession: Rasoherina, Ranavalona II, and Ranavalona III. He embarked on a program of modernization; in 1869 he caused Protestantism to be adopted and suppressed the Malagasy religion. European-style ministries were created and governors set up in the provinces. Villages were supervised by former soldiers. Education was declared obligatory under the direction of the Christian missions. A code of laws was worked out that combined ancient customs with Western practices such as monogamy.

The French began to extend their influence over the Sakalava, and the first "Franco-Merina" War (1883–85) ended with an ambiguous treaty: France was given a settlement at Diégo-Suarez and a resident at Antananarivo, but the institution of a protectorate was temporarily avoided. The succeeding period was marked by disorder and internal strife. In 1890 the British recognized Madagascar as a French protectorate, but Rainilaiarivony refused to submit. In January 1895, French troops landed at Majunga and formed a mobile column which, on September 30, 1895, after the only significant battle of the war, occupied Antananarivo. The prime minister was exiled. The queen signed a treaty recognizing the protectorate and was maintained on the throne as a figurehead.

The French invasion

(H.J.D.)

The French period. *The colonial period (1896–1945).* The occupation soon extended to the entire part of the island conquered by the Merina. But in Imerina itself, armed bands (the Menalamba, or "red togas") resisted modernization and French rule. The French Parliament voted the annexation of the island on August 6, 1896, and sent Gen. J.S. Gallieni first as military commander, then as governor general. Slavery was abolished. Gallieni put down the insurrection, subdued the oligarchy, and sent the queen into exile on February 27, 1897. In 1898 the old Merina kingdom was pacified; Gallieni then undertook the difficult task of subjugating the independent peoples. Two insurrections, in the northwest (1898) and in the southeast (1904), were quickly put down and, when he left the island in 1905, unification had been achieved. The Merina governors had been replaced by French administrators with leaders taken from local peoples. The teaching of French in the schools was made compulsory. Customs duties favoured French products, though Malagasy enterprise was also encouraged. The Tamatave–Antananarivo railroad was begun, roads were built, and a modern health service was inaugurated.

The economic development of the island continued under Gallieni's successors. The railroad and its branch lines were completed in 1913. A second line, the Fianarantsoa–Manakara, was finished in 1935. Automobile roads increased after 1920, airlines after 1936. The cities and seaports were built up and equipped, and loans were contracted in France. Exports were confined to agricultural products and raw materials for industry. Rice, manioc, rubber, raffia, meat, and graphite predominated at first. Between World Wars I and II, coffee, vanilla, cloves, and tobacco, introduced by the Europeans and then taken up by native planters, became more important. Three-quarters of all trade was with France. From 1900 to 1940 the population rose from 2,500,000 to 4,000,000. Material aspects of life became westernized, especially in the cities. Half the population became Christianized.

Malagasy nationalism

In 1915 a nationalist secret society, the Vy Vato Sakelika (VVS), was outlawed. In 1920 a teacher, Jean Ralaimongo, launched a campaign in the press to give the Malagasy "subjects" French citizenship and to make Madagascar a French *département*. When France failed to respond to the demand for assimilation, the movement turned toward nationalism. In 1940 Madagascar, though hesitant at first, rallied to the Vichy government. Then came a blockade, occupation by the British and South Africans (1942), and finally a return to Free France.

The French Union (1946–58). In the elections of 1945, two Malagasy nationalists were elected to the French parliament. The constitution of 1946, creating the French Union, made Madagascar an Overseas Territory of the French Republic, with representatives to the Paris assemblies and a local assembly at Antananarivo. Six provincial assemblies were created later. The political struggle erupted in violence on March 30, 1947, with an insurrection in eastern Madagascar. The leaders of the Mouvement Démocratique de la Rénovation Malgache (MDRM), including the three representatives to the French national assembly, were outlawed. More than 11,000 lives were lost before the revolt was ended.

A period of political inactivity followed until the 1950s. After the "Overseas Territories Law" of 1956 gave Madagascar an executive elected by the local assembly, Vice Premier Philibert Tsiranana founded the Parti Social Démocrate (PSD), which, though most of its members were non-Merina from the coastal areas, offered to cooperate with the Merina. In 1958 France agreed to let the overseas territories decide their own fate. In a referendum on September 28, Madagascar voted for autonomy within the French Community. On October 14, 1958, the autonomous Malagasy Republic was proclaimed; Tsiranana headed the provisional government.

The Malagasy Republic. The opposition regrouped under the name of Ankoton'ny Kongresy ny-Fahaleovantena Malagasy (AKFM), which included Protestant Merina dissidents and Communists. Antananarivo was this party's stronghold; it also had some support in the provinces but, owing to the electoral system established by the PSD, held only three seats in the legislature.

The PSD also settled the provincial question: executive power in the local assemblies was vested in a minister delegated by the central government. Tsiranana was elected president of the republic. He obtained independence for the republic on June 26, 1960. Tsiranana and the PSD remained in power until 1972. Under this regime, successive development plans were inspired, according to Tsiranana, by a "grass roots socialism," and were aimed at improving the lot of the peasantry. In foreign policy, the bond with France remained strong, and close relations were established with the United States, West Germany, Taiwan, South Africa, and other anti-Communist powers.

The Tsiranana regime

Tsiranana was reelected in January 1972, but political and labour unrest, and his own poor health, led him to appoint Maj. Gen. Gabriel Ramanantsoa as prime minister with full powers of government. A plebiscite on October 8, 1972, confirmed Ramanantsoa as head of government; there was no longer a president after Tsiranana resigned on October 11. The new head of government initiated radically different foreign and domestic policies. Under new agreements with France, French military and naval forces were removed from the island between 1973 and 1975, and all French citizens were to be treated as aliens. Ties were established with the Soviet Union, the People's Republic of China, and other Communist nations, and the country was withdrawn from the franc zone. In 1973, a rural reorganization program—in which elected committees would sell produce to state-owned companies—was initiated, and the government began to take control of joint French-Malagasy organizations.

In the wake of political and social unrest, on February 5, 1975, Ramanantsoa handed power over to a former minister of the interior, Col. Richard Ratsimandrava. He assumed the titles of president and prime minister but was assassinated six days later. A military directorate was then established; it dissolved on June 15, after naming Lieut. Comdr. Didier Ratsiraka president and head of the Revolutionary Council. A referendum on December 21, 1975, approved Ratsiraka as president under a new constitution that set up the Democratic Republic of Madagascar. Ratsiraka, sworn in as president on January 4, 1976, continued the policies begun by Ramanantsoa. He nationalized the banks, insurance companies, and the nation's mineral

The Ratsiraka regime

resources and solidified his nation's ties with the Communist powers. The formation in 1976 of the National Front for the Defense of the Revolution—a coalition of formerly banned political parties headed by Ratsiraka—further increased the president's control of the government and of Madagascar's political life. (H.J.D./Ed.)

The Avant-garde de la Révolution Malgache (AREMA) party, which formed the nucleus of the National Front, won control of the government in elections held from March to June of 1977. The Merina of the high plateau remained opposed to the government, and in May 1978 student protests in Antananarivo precipitated riots, which also involved many unemployed youths. Security was tightened in 1979 as fears of intrigue and plots against the government led to increased arrests, expulsion of foreigners, and repression of strikes. Ratsiraka has maintained close relations with the Eastern bloc countries and with the People's Republic of China, but most of the country's trade is with Western nations. Relations with France improved, despite a territorial dispute over the sovereignty of several islands in the Mozambique Channel.

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 943, 96/11, and 978. (Ed.)

Maldives

The Maldive Islands are a chain of some 1,300 small and low-lying coral islands, grouped in clusters or atolls (the word atoll is Maldivian) between latitudes 7°6′ N and 0°42′ S, about 400 miles (643 kilometres) southwest of Sri Lanka. Their land area is approximately 115 square miles (298 square kilometres).

PHYSICAL AND HUMAN GEOGRAPHY

One and a Half Degree Channel separates the main chain of atolls from Huvadu (Suvadiva) and the southern islands. Politically, the islands form an independent republic. The island of Minikoy (Maliku in Maldivian) geographically and historically formed part of the Maldives but was captured by a raja of Kannanur in 1752 and now belongs to the Lakshadweep group, although a form of the Maldivian language is still spoken there. Of the islands, some 200 are inhabited. The climate is hot and humid. The capital island is Male. The Maldivians are a mixed people, speaking an Indo-European language called Divehi, akin to Sinhalese. They are primarily Muslim.

The islands are covered with coconut palms, and millet, tropical fruit, and vegetables are grown. The principal industry is fishing, primarily for skipjack tuna, and the first factory for canning and processing fish was opened in 1978. Shipping, tourism, garment manufacturing, and coconut and coconut oil production are other significant industries. Traditional activities include coir and lace making, mat and cadjan weaving, and collection of cowries and other brightly coloured shells.

For administrative purposes the islands are divided into 19 atolls, each known by a letter of the Maldivian alphabet, in addition to its geographical name. The Maldives are governed by a popularly elected president and a Majlis (Parliament), to which all atolls elect members.

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

HISTORY

The first settlers probably arrived in the Maldives from Ceylon (Sri Lanka) not later than AD 500 and were Buddhists; some Buddhist remains survive. According to tradition, Islām was adopted in 1153, and there is a written *tarikh,* in Arabic, from that date, supplemented by other histories written in Maldivian. Ibn Baṭṭūṭah, the North African traveller, resided there during the mid-1340s and described conditions at that time, remarking disapprovingly on the freedom of the women—a feature that has been noticeable throughout Maldivian history.

The Portuguese forcibly established themselves in Male from 1558 to 1573. In 1602 a Frenchman, François Pyrard de Laval, was shipwrecked in the Maldives and remained there until 1607; he afterward published a lengthy account of the place, much of which remained valid until modern times. In the 17th century the islands came under the protection of the Dutch rulers of Ceylon, and protection was extended by the British when they took possession of Ceylon, though not formally recorded until the agreement of 1887. Owing to their remoteness and unhealthiness, the islands never came under colonial administration. Bohrās (Ismaʿīlī) merchants from Bombay established themselves in Male during the 19th century, acquiring a control of foreign trade that led to some trouble; after 1960 they were all expelled.

The first constitution

In 1932 a constitution was imposed upon the King by the nobles, and the King was subsequently deposed. The next king, who was elected, resigned during World War II, and the country remained a regency until 1953. In that year a republic was proclaimed under Amin Didi, but, after nine months, the President was deposed and later killed in a riot. A new king was then elected and reigned until November 1968, when, after a referendum, a new republic was inaugurated.

In 1948, when Ceylon became independent, an agreement provided that Great Britain control the foreign affairs of the islands but not interfere internally. The Maldivians undertook to provide facilities to British forces for the defense of the islands or of the Commonwealth. In 1956–57 Britain secured an agreement permitting reestablishment of a wartime airfield on Gan Island in Addu Atoll as a staging post and undertook resettlement of the Gan islanders. A new government then asked for a review of the agreements. This was followed by an insurrection against the Maldivian government in the southern atolls, but in 1960 the Maldivian government presented the British government with the use of Gan and other facilities in Addu Atoll for 30 years, with effect from December 1956, receiving in return a direct grant of £100,000 and a grant of £750,000 to finance specific development projects over a period of years. This agreement reaffirmed Britain's defense obligations, but in July 1965 Britain recognized the Maldive Islands as a completely sovereign and independent state and ceased to be responsible for their defense. The British government, however, retained the base and continued to make payments to the Maldivian government until 1976.

By the early 1980s a national monetary authority and a statistics bureau were established. Tourism remained the major economic hope for the republic. The Maldives became a member of the Commonwealth in 1982.

For later developments in the history of Maldives, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL (C.H.B.R.)

Seychelles

The Republic of Seychelles is an archipelago situated between 4° and 11° S latitude and 46° and 56° E longitude. Mahé, the main island, is about 680 miles (1,100 kilometres) northeast of Madagascar. The country is composed of 40 central, mountainous, granitic islands of the Mahé group and 52 outer, flat, coralline islands, all spread over an area of about 150,000 square miles (400,000 square kilometres). The archipelago has a land area of 171 square miles, of which Mahé Island takes up 57 square miles and is the home of Victoria, the capital. Seychelles is a member of the United Nations, the Commonwealth, and the Organization of African Unity.

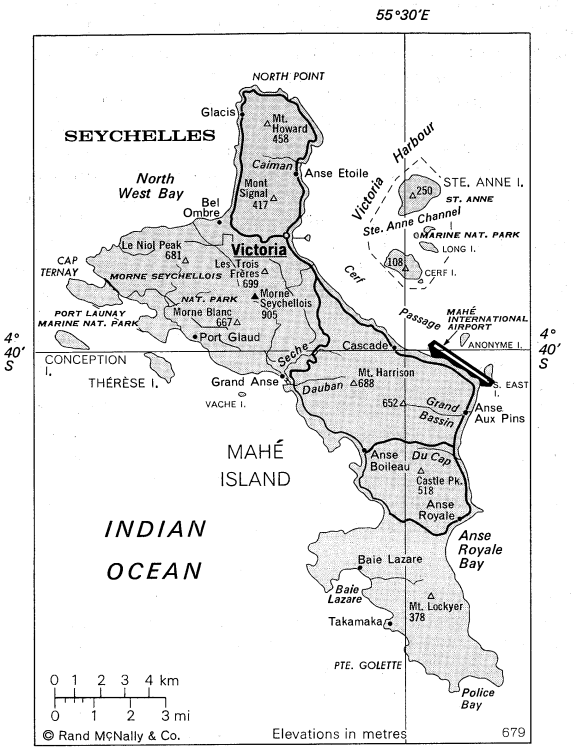
PHYSICAL AND HUMAN GEOGRAPHY

The land. The islands of the Mahé group rise from a large, shallow, crescent-shaped submarine plateau to 3,000 feet (900 metres) above sea level in the Morne Seychellois on Mahé. The overall aspect of these islands, with their lush tropical vegetation, is that of high, hanging gardens overlooking silver-white beaches and clear lagoons. In contrast, the outer, coralline islands rise only a few feet above sea level.

Flora and fauna

The climate is tropical-oceanic, the average annual temperature varying from 24° C (75° F) to some 30° C (86° F) at sea level. Two monsoons blow over the islands, the southeast ("Suette") from May to November and the northwest ("Vents d'Nord") from November to May. Rainfall varies from island to island, tending to decline from the granitic islands to the southernmost coralline islands. On Mahé, the annual rainfall at sea level averages 92 inches (2,300 millimetres). The granitic islands have only small streams that run swiftly into the sea, hence, even short droughts produce water shortages.

The only remaining primary forests are of coco-de-mer (or sea coconut; *Lodoicea seychellarum*), which grows in



SEYCHELLES

protected reserves on Praslin and Curieuse islands. Aldabra Island is famed for its birdlife. The giant tortoise, once widely distributed, is now almost extinct and is protected. Hunting of the green sea turtle also is regulated. There are few insects and no animals dangerous to man on the islands, though there are sharks in the seas around.

The people. The original French colonists on the previously uninhabited islands, and their black slaves, were joined in the 19th century by deportees from France and by a few British. Asians from China, India, and Malaya arrived later in smaller numbers. Widespread intermarriage has resulted in a population of mixed descent.

All along the coasts of the main granitic islands are small villages, many of them populated by fishermen and their families, made up of thatched cottages clustered around a church and facing the sea. On the small islands there are only isolated cottages near the beaches. Whale watching is a major tourist attraction.

Creole, English, and French are the official languages. Creole, a patois, is a mixture of clipped French (much of it archaic) with words from African languages and a few mutilated English words. About 90 percent of the population is Roman Catholic.

The economy. The so-called tertiary economic sector (commerce, transport, banking, insurance, tourism, administration, and domestic services) is the most important in Seychelles' economic life, followed by agriculture, forestry, fishing, and the building industry. There is a small but growing number of industrial plants, including a brewery and manufacturers of fabricated steel products, furniture, plastics, and soap. Tourism is of major importance in the economy, providing the bulk of foreign exchange earnings. The traditional export products (cinnamon, copra, essential oils, and guano) have gradually lost importance. A modern fishing industry supplies both home and foreign markets.

There is a good system of roads on Mahé, Praslin, and La Digue islands. The harbour of Victoria has been enlarged and modernized. The international airport, near Victoria, was opened in 1971, the first modern hotel in 1972. Inter-island transport is provided by sea and air services.

Administrative and social conditions. The constitution of the Seychelles, adopted in 1979, establishes the Seychelles People's Progressive Front (SPPF) as the sole legitimate party. The SPPF nominates all candidates for political office. Executive power is vested in the president, who is head of state and head of government. The president nominates and dismisses members of the Council of Ministers as well as holders of certain other public offices and the judiciary. The unicameral National Assembly is composed of 23 members elected to five-year terms and two appointed members. There are a Supreme Court, a Court of Appeal, and Magistrates' Courts for both criminal and civil matters. In 1977 the government began to enlarge the country's security forces.

Though education is not compulsory, 95 percent of children of primary-school age attend school. Six-year primary school is free. Children over 12 may attend secondary grammar school (admission by examination) or junior secondary school followed by technical and vocational courses. There is a teacher-training college, and scholarships are awarded for further education overseas.

Seychelles has few tropical diseases. Tuberculosis, leprosy, and filariasis, once common, are declining. The main islands have adequate hospitals and clinics.

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

HISTORY

The Arabs were probably the first to see the Seychelles, but they did not settle there, and the Portuguese sighted the archipelago in the early 16th century. The first recorded landing was made in 1609 by the English East India Company, which found the islands uninhabited and abounding in water, fish, birds, and fruits. During the early 18th century they proved to be ideal hideouts for pirates.

In 1742 Bertrand-François Mahé, comte de La Bourdonnais, the French governor of Mauritius, sent Capt. Lazare Picault to explore the islands. Picault eventually anchored in a bay on the southwest coast of Mahé, which he named. The French, however, made no real attempt to annex Mahé and its surrounding islands till 1756, when a frigate was sent to take possession and forestall the British. The islands were named Séchelles, later altered by the British to Seychelles. In 1770 the first colonists began to arrive from Réunion. They so depleted the forests and the giant tortoises for exports that in 1789 they were confined to fixed areas and forbidden to trade. By the early 19th century the population had grown to more than 2,000. War between France and Great Britain did not affect the islands till 1794, when a British squadron appeared and demanded unconditional surrender. The Seychelles were finally ceded to Britain in 1814, by the Treaty of Paris, and were incorporated as a dependency of Mauritius, which by that time the British also had taken from France.

Under British rule the Seychelles entered a long period of peaceful obscurity. The abolition of slavery in the 1830s produced a fundamental change in the agricultural pattern. The settlers, finding themselves without a labour force, changed from raising labour-intensive food crops and cotton to crops that required less labour, such as coconut, vanilla, and cinnamon. In 1903 the Seychelles became a crown colony. Previous constitutional advances had occurred in 1872 with the appointment of a Board of Civil Commissioners and in 1888 with the appointment of the first nominated Legislative and Executive councils. The Seychelles, however, had to wait till 1948 to have their first Legislative Council with elected members. In 1970 they obtained a new constitution, universal adult suffrage, and a governing council with an elected majority.

The year 1964 had seen the formation of the Seychelles People's United Party (SPUP), led by France-Albert René, and the Seychelles Democratic Party (SDP), led by James R. Mancham. The first was leftist, the second centre–conservative, but both came to espouse independence. The elections of November 1970 brought the majority of votes and 10 seats in the Legislative Council to the SDP, and Mancham became chief minister. At the next election (April 1974) the SDP increased its majority.

On October 1, 1975, the archipelago obtained the absolute right of self-government, and at midnight on June 28–29, 1976, Seychelles became independent as the 36th member of the Commonwealth of Nations. At the same time the outer islands Aldabra, Desroches, and Farquhar—which, together with Diego García in the Chagos Archipelago, for military reasons had received in 1965 an exceptional position as the British Indian Ocean Territory—were returned to Seychelles. (W.R.M.)

In 1975 the two political parties agreed to form a coalition government, with Mancham as president and René as prime minister. On June 5, 1977, while Mancham was in London, rebels seized control of the police headquarters and other public buildings and invited René to assume the presidency. René's socialist government survived numerous attempted coups, which were launched by opposition groups with the help of foreign mercenaries. By the early 1980s five-year plans abolishing private medicine, redistributing land, and expanding public housing were under way.

For later developments in the history of Seychelles, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

Comoros

The Comoros are a group of volcanic islands at the northern end of the Mozambique Channel, lying roughly halfway between Madagascar and the African mainland in latitude 11° to 13° S, with a combined area of about 863 square miles (2,235 square kilometres). Three of the islands—Grande Comore (Ngazidja), Mohéli (Moili), and Anjouan (Ndzouani)—comprise the independent state of Comoros. The fourth island of the archipelago, Mayotte (Maore), is a de facto French dependency.

PHYSICAL AND HUMAN GEOGRAPHY

The land. From southeast to northwest the islands are: (1) Mayotte (144 square miles), which has a basalt plateau rising to 2,165 feet (660 metres) and a coastal plain on the west. Mayotte is partly encircled by an irregular coral reef, within which are the islets of M'Zambourou and Pamanzi. On Pamanzi is Dzaoudzi, the capital of the group. Opposite it on the main island is the port of M'Sapéré. (2) Anjouan (164 square miles) is a triangular island rising centrally in a volcanic massif (M'Tingui) reaching 5,177 feet. Mutsamudu on the northwest coast is its chief town. (3) Mohéli (112 square miles), the smallest of the group, is a basalt plateau about 1,000 feet high, ending in the west in an escarped ridge reaching 2,625 feet. Its chief towns are Fomboni on the north coast and Nioumachoua in the southwest. (4) Grande Comore (443 square miles), the largest and loftiest island, rises near its southern end in the dome-shaped active volcano of Kartala (7,746 feet). In the southwest is Moroni, the capital and largest city of Comoros, and on the north coast the town of Mitsamiouli.

Climate

The tropical climate has two clearly marked seasons: a fresh, dry period between May and October and a hot, humid season between November and April. In November the onset of the monsoon (*cascassi*) brings the highest temperatures. The highest rainfall occurs in January, and the summer also brings cyclones, often accompanied by waterspouts and damaging tidal waves. During the winter the temperature falls with the southeast trade wind (*coussi*), reaching a minimum in July (Dzaoudzi, 73° F [23° C]).

The hydrology of the islands is of importance, especially on arid Grande Comore. Rain sinks so deeply into the lavas and porous rocks that wells (*foumbous*) deep enough to reach the water table can be sunk only in the coastal lowlands, where the water obtained is brackish. Most of the supplies for Grande Comore and Dzaoudzi are drawn from reservoirs filled in the rainy season. In the other islands it is possible for streams to form on the decomposite basalt lavas, and wells can be sunk.

The character of the vegetation varies with altitude and degree of exposure to the wind. Only about 16 percent of the land remains covered with the primitive forest. A coastal zone of mangroves is followed by one of coconut palms, mangoes, and bananas up to about 1,300 feet, above which the forest zone (mahogany, rosewood, and tacamahac) rises to about 5,900 feet. On the highest peaks only broom, heather, and lichens grow. The animal life is allied to that of Malagasy rather than to the mainland of Africa; it includes some land birds and a lemur peculiar to the islands. Turtles abound along the coasts and are exported. In the 1950s specimens of the primeval coelacanth were caught in the surrounding waters.

The people. The people are of diverse origins. Malay immigrants and Arab traders have mixed with Sakalavas and other Malagasy tribes and with various African peoples brought in as slaves. The most important community are the Antalaotra, who are of mixed origin and are most numerous in Anjouan (hence their incorrect name Anjouanais) and Grande Comore. Most of the people are Shafi'i Muslims, and most speak Comorian, which is closely allied to Swahili; French and Arabic are also spoken. Of the population, about half live on Grande Comore and about two-fifths on Anjouan. The residents of Mayotte also are primarily Muslim, and although the official language is French, most speak Comorian.

The economy. The Comoros are heavily dependent on foreign aid, contributed mainly by France and Arab countries. The lack of pasture and of flat cultivable land calls for a large importation of foodstuffs. Subsistence agriculture on small plots tilled on a family basis yields cassava, sweet potatoes, and other vegetables; some mountain rice is grown. Livestock raising is concentrated in Grande Comore, and there are goats, cattle, and sheep. European plantations dating from the early 19th century first cultivated sugarcane; this was supplanted in the 20th century by vanilla, to which were added perfume plants (ylang-ylang, citronella, bitter basil). Later coconuts, coffee, cocoa, and (in Anjouan and Grande Comore) sisal were cultivated. The plantations, which cover about 20 percent of the total area, are mostly managed by companies but are being rapidly transferred to small farmers. Forests on Grande Comore provide timber for construction, and a modern fishing industry is being developed, although its potential is limited. Industrial development is slight; it includes production of soft drinks, vanilla processing, distillation of essential oils, and woodworking. There are three hydroelectric power plants.

The plantations, while providing a living for part of the population, do not succeed in balancing the economy, and the value of imports substantially exceeds that of exports. Emigration is extensive. Regular shipping services call at the islands, and Grande Comore, Mohéli, and Anjouan are linked by regular air service. The international airport at Moroni-Hahaya on Grande Comore opened in 1975 and provides service mainly to countries in Europe and Africa. (Je.D./Ed.)

Administrative and social conditions. According to the constitution of 1978, the government is headed by an elected president, who nominates a council of ministers and the governors of individual islands. The national legislature is chosen by direct popular vote.

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

HISTORY

The Comoro Islands were shown on Diego Ribero's world map of 1527, but the first European to visit them appears to have been the British mariner James Lancaster in about 1591. The dominant influence and civilization in the islands was then and for long afterward Arab. In 1843 France took possession of Mayotte, and in 1886 the sultans of the other three islands were placed under French protection. In 1914 the group was subordinated

to the governor-general of Madagascar and in 1925 was given its own local administration; in 1947 it became an overseas territory of France.

In a referendum held in 1958 the Comoro Islands voted to retain overseas territory status. In 1960 they were granted internal autonomy, and in 1975 they became independent, except Mayotte, which remained a de facto French dependency. A constitution approved in 1978 proclaimed the Federal Islāmic Republic of the Comoros. Mayotte refused to join the republic, seeking status instead as a French *département*. France announced its intention in 1984 to reintegrate Mayotte and Comoros, but the residents of Mayotte declined to hold the referendum necessary to decide the issue. (Je.D.)

For later developments in the history of Comoros, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

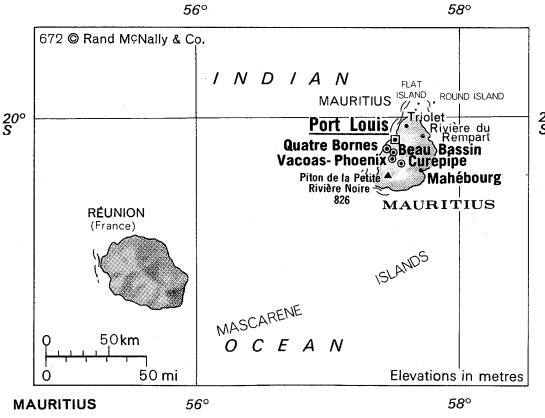
Mauritius

Mauritius is a constitutional monarchy and member of the Commonwealth. It comprises the islands of Mauritius (the central island of the Mascarene group) and Rodrigues and the island dependencies of Agalega and the Cargados Carajos Shoals. The island of Mauritius is situated about 500 miles (800 kilometres) east of Madagascar. It is 38 miles long by 29 miles wide and has an area of 720 square miles (1,865 square kilometres). The country's total area is 788 square miles and its capital is Port Louis, on the northwestern coast of Mauritius.

PHYSICAL AND HUMAN GEOGRAPHY

The natural setting

The land. The island of Mauritius is fringed by coral reefs. The northern part is a plain; the centre is a plateau rising to a height of about 2,200 feet (670 metres) and bordered by small mountains, which are remnants of a huge volcano; the south is largely mountainous. The highest peak on the island is the Piton de la Petite Rivière-Noire (Little Black River Peak) in the southwest, which is 2,711 feet high. There are many streams, generally flowing through deep ravines. During heavy rains these streams become torrents. On the plateau are two lakes, Grand Bassin and Mare aux Vacoas; the latter is the chief source of water supply. Four other large reservoirs are used for irrigation. Rodrigues is a volcanic island surrounded by coral reefs. It lies 344 miles east of Mauritius and has an area of about 40 square miles.



Climate

The climate is maritime subtropical, with only two seasons—hot from November to April and cool from May to October. The southeast trade winds blow most of the year, and the climate is generally humid. The average temperature at Port Louis, on the coast, during the hot summer months is about 84° F (31° C), and in the cooler winter months it is about 77° F (25° C); at Curepipe, on the inland plateau, the temperature averages about 75° F (24° C) in summer and about 68° F (20° C) in winter. Heavy rains and cyclones occur during the hot season. Annual coastal rainfall on Mauritius varies between 35 inches (890 millimetres) on the west coast to 60 inches on the southeast coast. The central plateau receives about 200 inches of rain. The vegetation includes approximately 600 indigenous species, even though little original forest is now left. Many plants have been introduced since the 18th century, and in 1770 a botanical garden, which still exists, was formed at Pamplemousses. The original fauna was composed of insects, birds, reptiles (including tortoises), and one mammal, the fruit-eating bat. Of particular interest are a number of wingless birds, such as the dodo and the *Aphanapteryx* (a large long-billed rail), all now extinct. Among imported mammals are the sambar (long-tailed dark brown deer), tenrec (spiny insectivore), and mongoose. There are no venomous animals, but in about 1865 the *Anopheles* mosquito, which carries malaria, was introduced and spread the disease.

Extinct species

The island of Mauritius is divided into seven coastal and two inland districts. Most of the rural population still lives in the coastal districts, chiefly in the north and east, which were settled first. The original slave population lived on plantations. After emancipation in 1835 the freed slaves left to settle in coastal fishing villages. The Indian labourers who replaced them lived in estate camps until, in the 1860s, they spread out over the island, forming straggling villages. The first towns were Port Louis, laid out on the French gridiron pattern, on the leeward (westward) side of the island, then Mahébourg, on the windward (eastward) side. After a malaria epidemic that lasted from 1866 to 1868, the wealthier inhabitants of the towns migrated inland. Port Louis, nevertheless, remained the capital, as well as the business and administrative centre, as it is today. It achieved city status in 1966. Some inland towns grew up in the 1870s on the Plaines Wilhems. These are Beau Bassin-Rose Hill, Quatre Bornes, Vacoas-Phoenix, and Curepipe. All are residential and have few commercial activities. Mahébourg is now little more than a village.

The people. Mauritius differs from many countries in that as late as the 17th century it had no indigenous population. Its current population is composed entirely of recent immigrants from four principal ethnic and religious groups—French, Creole (generally descendants of the slaves), Indian, and Chinese. Most French, Creoles, and Chinese are Christian. Nearly two-thirds of the Indians are Hindu; the remaining third are Muslim. A small portion of the Indian community is Christian. The first successful settlers on the island were French. They in turn imported slaves from Madagascar and East Africa.

Languages

The languages then spoken were French and Creole, a French patois. French remains the predominant European language, and Creole has become the lingua franca of the island. English, the official language, was introduced after 1810, as were some Asian languages. The chief Indian dialect is Bhojpuri. Many Chinese speak Cantonese.

Importance of sugar

The economy. Mauritius' economy is primarily agricultural and is dominated by sugar, which accounts for about two-thirds of the export trade. The secondary crop is tea. Some food crops and tobacco are grown, and there is a meat and dairy industry. Tourism is a major source of foreign exchange earnings. The industrial sector is small but growing. The establishment of duty-free zones in the early 1970s has promoted the development of new industries, particularly textiles and clothing, electronic components and precision engineering, and diamonds and diamond cutting. Mauritius' currency is the rupee.

Mauritius has about 1,100 miles of roads, of which most are tarred. Railways have been discarded, but there is an extensive bus transport system. Port Louis harbour has deepwater quays and dry-dock facilities and usually handles over a million tons of shipping a year. There is an international airport at Plaisance. The Mauritius Broadcasting Corporation runs television and radio programs.

Administrative and social conditions. The British monarch, as head of the Commonwealth, is represented by a governor-general. The Legislative Assembly consists of the speaker, 62 elected members, eight additional members, and the attorney general. The Cabinet consists of the prime minister and 20 other ministers. For electoral purposes, Mauritius is divided into three-member constituencies, of which there are 20. Rodrigues elects two

members. Local government is exercised at the village level by councils whose activities are coordinated by district councils. In the urban areas, administration is vested in municipalities, with 24 elected members each. Port Louis has a fully elective council. There are many social welfare centres, mainly in the rural areas.

Cultural life. British, French, and Indian culture are all promoted by agencies of the respective governments. The standard of secondary education is relatively high, and a university college, providing technical education, has been founded.

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

HISTORY

Known to Arab or Swahili seamen since the 10th century or earlier, Mauritius was visited by the Portuguese in about 1510. Finding Mauritius still uninhabited in 1598, the Dutch named it after their *stadhouder* (head of state) Maurice (Maurits) of Nassau. After two attempts to settle the island, made in 1638–58 and again in 1664–1710, the Dutch abandoned it, and it was then haunted by pirates. In 1721 the French East Indian Company, after annexing Réunion, sent a party to occupy Mauritius, which they renamed Île de France. Under the rule of the company, which lasted from 1721 to 1767, the progress of settlement was slow. Port Louis was founded and was used as a base for attacking the British in India. During the Seven Years' War, from 1756 to 1763, a further attempt to use it as a base against the British again failed. In 1767 the French Ministry of Marine took over administration and a new era began. During the American Revolution, French naval units and privateers, based at Port Louis, gained many successes. Port Louis then became a vital way station, being visited by many French, Danish, and American ships. When, after the French Revolution, the government in Paris tried to abolish slavery in 1796, the island's settlers broke away from France. Even without French governmental support, however, the settlers managed to maintain themselves.

First settlement of the island

In 1810 the British captured Mauritius, restored its Dutch name, and occupied it with its dependencies as a crown colony. From a commercial and military outpost it was then transformed into a plantation colony, with sugar as its sole product. (Au.T./Ed.)

In 1835 the British abolished slavery, and Mauritian planters began to import indentured labourers from India. Between 1835 and 1907 more than 450,000 Indians were brought to Mauritius to work on the sugar estates. The abuses to which these labourers were subjected were investigated occasionally during both the 19th and 20th centuries, but the conditions on the estates remained cruel. Most of the labourers were Hindu, but Muslim and Chinese traders also settled on Mauritius in the 19th century.

In 1937 the island's poor, Creoles and Indians alike, joined in riots demanding better economic conditions and participation in government. Labour unions were formed, and a new constitution gave the vote to a large number of the Indians and Creoles in 1948 and provided for an important role for elected representatives in the much enlarged legislature. Between 1948 and the granting of independence, political alliances based on class gave way to alliances based on race and culture. Franco-Mauritians, Creoles, and many Muslims and Chinese feared that the Indian-dominated Labour Party would bring economic problems to the island. In 1967 the Labour Party alliance won a startling victory and led the country to independence. The government was dominated by the founders of the Labour Party, who had first agitated for constitutional reform in the 1930s and 1940s. By the 1980s many of these leaders were old and were blamed for the country's economic difficulties. (Ad.S.)

In 1983 the Labour Party formed a coalition with the Mauritian Socialist Movement (which incorporated the former Mauritian Socialist Party) and the Social Democratic Party to block the rise of the Mauritian Militant Movement. The coalition sought status as a republic for Mauritius.

For later developments in the history of Mauritius, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

Réunion

Réunion, one of the Mascarene Islands and a French overseas *département*, lies about 420 miles (680 kilometres) east of Madagascar and about 110 miles southwest of Mauritius. Elliptical in form, it has an area of 970 square miles (2,512 square kilometres). The capital is Saint-Denis.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Réunion is entirely volcanic. The main body is a volcanic mass that ceased activity in the Pliocene and is in an advanced state of dissection by short torrential rivers, the chief of which are the Rivières du Mât and des Galets, Bras de Cilaos, and Ravine Sèche. In the west central area is a mountain massif with three summits exceeding 9,000 feet (2,740 metres), the Piton des Neiges (10,069 feet), Le Gros Morne (9,816 feet), and Grand Bénard (9,501 feet). Around them are scooped out three wide basins, the Cirque de Cilaos on the south and the Cirque de Salazie and Cirque Mafatte on the north. Around these amphitheatres are a series of plateaus—tablelike facets of the uneroded slopes of the former cones. In the east is an area of about 200 square miles in which vulcanism has been more recently active, and in the extreme east is Le Volcan, one of whose craters, Piton de la Fournaise, has been active several times since 1925. Le Volcan, exceeding 8,600 ft., and its eastern lava slope, Grand Brûlé, are entirely enclosed within a parabolic mountain wall about 30 miles long, the whole basin being a huge caldera refilled by recent vulcanism. The massif descends steeply on all sides, but everywhere except on the east it is bordered by a narrow coastal shelf or plain one-half to three miles in width. The coast consists mainly of cliffs and has no natural harbours.

The climate is dominated by the southeast trade winds during April–October, when the northern and western sides are almost rainless and heavy rain falls in the south and east. Temperatures are cool, especially in the highlands, and snow falls frequently on the Piton des Neiges. In summer Réunion lies within the doldrums, and the lowlands are oppressively humid. Rain falls in all parts, and the island is liable to destructive tropical cyclones. Above 2,500 feet tropical conditions give way to subtropical and temperate conditions. Above 5,000 feet the climate is cool, with average temperatures around and below 60° F (15.5° C), and with cold nights.

Climate

The lowlands formerly carried tropical rain forest, though with a palm savanna on the leeward side. Higher up is a mixed forest with much bamboo, *Casuarina* and *Terminalia*. Above 5,000 feet a single species, the tamarind, covers a large part of the slopes. Conifers occur above 6,500 feet, and then a subalpine heath appears. Only about one-fifth of Réunion remains under forest.

The indigenous animal life is not rich in species, many having disappeared. Several bats, wildcats (the tenrec from Madagascar), and rats are the chief mammals. Common birds include the Bourbon flycatcher (*Tchitrea bourbonnensis*), the Mascarene bulbul, the Bourbon manioc bird (*Malacirops borbonicus*), and the introduced cardinal (*Foudia madagascariensis*). There is only one snake (*Lycodon aulicus*).

The people. Réunion has no indigenous population. When discovered by Europeans it was uninhabited, and no trace of earlier settlement was ever found. Present inhabitants include Creoles, blacks, mulattoes, Chinese, and Indians. Most of the population is descended from slaves and is of African origin, mixed with other elements. A high proportion of the whites live in the highlands, whereas the lowland population is overwhelmingly dark-skinned. Fertility is highest in the lowlands, especially in the capital. Mortality is highest on the windward coastlands, especially in the towns, and lower in the highlands. There is serious overpopulation, and massive emigration has long characterized Réunion. From the early 1950s the main direction of this movement was to Madagascar.

The Creole patois is French modified by Malagasy and Bantu pronunciation. French is the official language. Most of the population is Roman Catholic, exceptions being small numbers of Hindus, Muslims (Indian and Arab), and Chinese.

The economy. The sugar industry dominates the economy, production being concentrated on the coastal lowlands. It was prosperous up to the mid-19th century but thereafter suffered heavily through competition from beet sugar in France; improvement was slow. Despite persistent archaic land laws and inefficient harvesting techniques and transportation systems, a modernization program has improved both production levels and yields. Rum is distilled.

Besides the sugar Réunion exports, the only other important export crops are geranium, vetiver (from which base oils for perfumes are extracted), and vanilla. The vanilla industry developed on the wettest part of the windward coast during the worst of the sugar crisis in the 1870s. Geranium and vetiver production, and to a lesser extent the perfume plants ylang-ylang and patchouli, are concentrated in the highlands, where they developed after the decline of coffee following an invasion of a coffee-leaf disease in the 1880s. Tourism is important.

Imports

The great majority of Réunion's exports (mainly sugar and rum) go to France. Imports are mainly manufactured goods (cotton piece goods) from France, foodstuffs (rice), machinery, and petroleum products. The currency is the CFA franc. There are two banks operating in the island.

Communications are good on the coast. A road encircles the island; lateral roads give access to the interior, especially to the hill resorts of Cilaos and Salazie. Open roadsteads in front of the main towns (principally Saint-Pierre) were used until the artifical harbour at Pointe des Galets was completed in 1886. All overseas trade is now handled there. A regular shipping service runs to Madagascar and also to France via Suez, and there is regular air service to Paris from Gillot airfield near Saint-Denis. Réunion is connected by telephone and telegraph with Mauritius and, via Madagascar, with France.

Administrative and social conditions. Réunion, as an overseas *département,* sends three deputies and two senators to the French legislature. Locally, it is governed by a commissioner, an elected general council, and an elected regional council. Technical services are provided by metropolitan ministries, and French legislation and justice extend to the island. Réunion is divided into four *arrondissements,* which are further subdivided into communes, each governed by an elected municipal council.

The mass of the population is poor, though wage labourers are generally better off than the peasantry. Commerce is largely controlled by a small number of sugar companies. Housing is poor, with much overcrowding in the towns. There are two technical schools, an agricultural school, and a teachers' training school. Public and private health facilities include hospitals and dispensaries.

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

HISTORY

Réunion is usually said to have been discovered on February 9, 1513, St. Apollonia's Day, by the Portuguese navigator Pedro de Mascarenhas. The island was at first called Santa Apollonia; and the Mascarene Islands, of which Réunion forms a part, were named for the navigator. François Cauche first hoisted the French flag in 1638, followed five years later by Jacques Pronis, who again took possession of the island in the name of France. Pronis deported from Madagascar to the island 12 mutineers, who, after three years, took back idyllic descriptions of it. Occupation was decided on by Étienne de Flacourt, governor of Madagascar, who gave it the name of Île de Bourbon. In 1662 Louis Payen settled the first colonists, and two years later the Compagnie des Indes Orientales initiated a regular colonization plan. By 1717 there were about 900 free inhabitants and 1,100 Malagasy and Kaffir slaves.

Coffee and sugar

Every colonist was obliged to own at least 200 coffee trees for every black working for him. Mocha coffee soon made the island prosperous. Bertrand François Mahé, comte de La Bourdonnais, named governor in 1735, developed the fields of rice, grain, and manioc and established the first plantations of sugarcane, indigo, and cotton; sugar soon became the principal crop. In 1764 the company's activities were brought to an end, and the administration of the island was entrusted to a governor, or lieutenant general of the king, and an intendant.

In 1793, with the proclamation of the Republic, Bourbon became the island of Réunion. Gen. Charles M.I. Decaen landed in 1803 and began preparations for war in that part of the Indian Ocean, but large numbers of English troops disembarked and in July 1810 the French capitulated. The island was occupied by the English until April 1815, when France again took possession. In 1848 the slaves (more than 60,000) were liberated. In 1946 Réunion became an overseas *département* of France, and in 1973 it was made the headquarters of the French military forces in the Indian Ocean. (R.Co.)

Réunion has generally remained content as a *département* and has benefitted from France's economic subsidies.

For later developments in the history of Réunion, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (Ed.)

BIBLIOGRAPHY

Madagascar: MALAGASY REPUBLIC, INSTITUT NATIONAL DE LA STATISTIQUE ET DE LA RECHERCHE ECONOMIQUE, *Inventaire socio-économique de Madagascar, 1960–1965,* 2 vol. (1966), information on social and economic conditions; *Recensements urbains: Ambositra, Ambalavao, Mananjary, Manakara, Farafangana* (1966), special surveys carried out in the main cities in addition to the general census; *Recensement Industriel* (1967), a survey of industries; SERVICE DE STATISTIQUE ET DES ETUDES SOCIO-ECONOMIQUES, *Population de Madagascar au 1er janvier,* a detailed annual compendium of demographic information; *Enquête agricole* (1966), details of agricultural production by crops and regions; MADAGASCAR, SERVICE GEOGRAPHIQUE, *Madagascar et Archipel des Comores,* 1:500,000 (1947–48). *The Antananarivo Annual and Madagascar Magazine,* 24 vol. (1875–78, 1881–1900), a journal in English produced by members of the London Missionary Society in Madagascar, containing unique articles on history, custom, belief, fauna, and flora; *Bulletin de l'Académie Malgache* (1905–), the main scholarly journal of Madagascar; MAURICE BLOCH, *Placing the Dead: Tombs, Ancestral Villages and Kinship Organization in Madagascar* (1971), a detailed ethnographic account of the numerically and culturally dominant Merina people; ANDRE DANDOUAU, *Manuel de geógraphie de Madagascar,* 10th ed. (1960), a useful general geography; HUBERT J. DESCHAMPS, *Histoire de Madagascar,* 4th ed. (1972), *Les Migrations intérieures passées et présentes à Madagascar* (1959), the only detailed study of internal population movements, and with SUZANNE VIANES, *Les Malgaches du Sud-Est* (1959), an ethnographic account of the southeastern peoples through whom Arabic scripts and elements of Islāmic belief and practice influenced Madagascar; GUILLAUME GRANDIDIER, *Bibliographie de Madagascar,* 4 vol. (1905–57), an authoritative, cumulative bibliography, and with ALFRED GRANDIDIER, *Histoire physique, naturelle et politique de Madagascar* and *Ethnographie de Madagascar* (1908–28), two works that represent the most monumental attempt to present a complete account of the natural history and culture of Madagascar and its peoples; NIGEL HESELTINE, *Madagascar* (1971), a recent general study of the land and people; ROBERT ARCHER, *Madagascar depuis 1972* (1976), an account of the events of the 1970s; MERVYN BROWN, *Madagascar Rediscovered* (1978); *Madagascar* ("Les Guides Bleus," 1968), information on all towns and places of interest; AIDAN SOUTHALL (ed.), "Kinship, Descent, and Residence in Madagascar," *Am. Anthrop.,* 73:144–208 (1971), a comparative analysis of variations in Malagasy kinship and family structure; MARCELLE URBAIN-FAUBLEE, *L'Art Malgache* (1963), a publication devoted to Malagasy art. General histories include HUBERT DESCHAMPS, *Histoire de Madagascar,* 3rd ed. (1965). For the Traditional Period, see ALFRED GRANDIDIER (ed.), *Collection des ouvrages anciens concernant Madagascar,* 9 vol. (1903–20; see esp. vol. 8 and 9 for the works of Flacourt); GUILLAUME GRANDIDIER, *Histoire politique et coloniale de Madagascar,* 3 vol., the last in collaboration with RAYMOND DECARY (1947–58); and R.K. KENT, *Early Kingdoms in Madagascar, 1500–1700* (1970). Information on the 19th century may be found in WILLIAM ELLIS, *History of Madagascar,* 2 vol. (1838), *Three Visits to Madagascar* (1858), *Madagascar Revisited* (1867); and R.P. MALZAC, *Histoire du royaume Hova* (1912); on the 20th century, HUBERT DESCHAMPS and P. CHAUVET, *Gallieni pacificateur* (1949); and VIRGINIA THOMPSON and RICHARD ADLOFF, *The Malagasy Republic, Madagascar Today* (1965). Modern political developments are summarized

in JOHN M. OSTHEIMER (ed.), *The Politics of the Western Indian Ocean Islands* (1975).

Maldives: CLARENCE MALONEY, *People of the Maldive Islands* (1980). H.C.P. BELL, *The Maldive Islands* (1940); T.W. HOCKLY, *The Two Thousand Isles* (1935); A. GRAY (ed.), assisted by H.C.P. BELL, *The Voyage of François Pyrard of Laval to the East Indies, the Maldives . . .*, Hakluyt Society, first series, vol. 76, 77, 80 (1887–90).

Seychelles: Summary discussion is contained in MARCUS FRANDA, *The Seychelles: Unquiet Islands* (1982). JONATHAN D. SAUER, *Plants and Man on the Seychelles Coast: A Study in Historical Biogeography* (1967); BURTON BENEDICT, *People of the Seychelles* (1966); GUY LIONNET, *The Seychelles* (1972); and WILHELM MARQUARDT, *Seychellen, Komoren und Maskarenen: Handbuch der ostafrikanischen Inselwelt* (1976). On the history, see J.T. BRADLEY, *The History of the Seychelles,* 2 vol. (1940), and LOUIS DAYER, *Les Îles Seychelles: Esquisse historique* (1974). On the economy, see J.W.F. ROWE, *Report on the Economy of the Seychelles and Its Future Development* (1959); OFFICE OF THE PRESIDENT, *The Seychelles Handbook* (1976). On the colonial period see also the biennial *Colonial Reports* on the Seychelles (HMSO).

Comoros: C. ROBEQUAIN, *Madagascar et les bases dispersées de l'Union Française* (1958); H. ISNARD, "L'Archipel des Comores," *Cah. d'Outre-mer*, vol. 6, pp. 5 ff. (1953).

Mauritius: PATRICK BARNWELL and AUGUSTE TOUSSAINT, *A Short History of Mauritius* (1949); BURTON BENEDICT, *Mauritius: Problems of a Plural Society* (1965); JAMES E. MEADE, *The Economic and Social Structure of Mauritius,* 2nd ed. (1968); RICHARD TITMUSS, *Social Policies and Population Growth in Mauritius,* 2nd ed. (1968); *Une Cité tropicale: Port-Louis de l'île Maurice* (1966; *Port-Louis: A Tropical City,* 1973); *Historie de l'île Maurice* (1971); ADELE SMITH SIMMONS, *Modern Mauritius* (1982).

Réunion: CATHERINE LAVAUX, *La Réunion* (1975); JOHN M. OSTHEIMER (ed.), *The Politics of the Western Indian Ocean Islands* (1975); CLARISSE DESILES (ed.), *À la Réunion, à l'île Maurice, aux Seychelles* (1982).

Indian Philosophy

Indian philosophy, which includes both orthodox (*āstika*) systems, namely, the Nyāya, Vaiśeṣika, Sāṃkhya, Yoga, Pūrva-mīmāṃsā, and Vedānta schools of philosophy, and unorthodox (*nāstika*) systems, such as Buddhism and Jainism, has been concerned with various philosophical problems. Significant among these concerns have been the nature of the world (cosmology), the nature of reality (metaphysics), logic, the nature of knowledge (epistemology), ethics, and religion.

The article is divided into the following sections:

General considerations 180
- Significance of Indian philosophies in the history of philosophy 180
- General characteristics of Indian philosophy 181
  - Common concerns
  - Forms of argument and presentation
- Roles of sacred texts, mythology, and theism 181
- A general history of development and cultural background 182
  - The prelogical period
  - The logical period
  - The ultralogical period

Historical development of Indian philosophy 183
- Presystematic philosophy 183
- Early Buddhist developments 184
- The philosophical portions of the *Mahābhārata* 184
- Doctrines and ideas of the Buddhist *Tipiṭaka* 185
- Early system building 186
  - The history of the *sūtra* style
  - The *Pūrva-mīmāṃsā-sūtra*s and Śabara's commentary
  - The *Vedānta-sūtra*s
  - The *Sāṃkhya-kārikā*s
  - The *Yoga-sūtra*s
  - The *Vaiśeṣika-sūtra*s
  - The *Nyāya-sūtra*s
  - The beginnings of Mahāyāna Buddhist philosophy
  - The worldview of the *Arthaśāstra*
  - Fragments from the Ajīvikas and the Cārvākas
- Further developments of the system 190
  - Developments in Mahāyāna
  - Pūrva-mīmāṃsā: the Bhāṭṭa and Prābhākara schools
  - The linguistic philosophies: Bhartṛhari and Maṇḍana-Miśra
  - Nyāya-Vaiśeṣika
  - Sāṃkhya and Yoga
  - Vedānta
  - Vaiṣṇava schools
  - Śaiva schools
- Jaina philosophy 199
- Mughal philosophy 199
- 19th- and 20th-century philosophy in India and Pakistan 199

Bibliography 201

General considerations

SIGNIFICANCE OF INDIAN PHILOSOPHIES IN THE HISTORY OF PHILOSOPHY

Relations to Western thought

In relation to Western philosophical thought, Indian philosophy offers both surprising points of affinity and illuminating differences. The differences highlight certain fundamentally new questions that the Indian philosophers asked. The similarities reveal that, even when philosophers in India and the West were grappling with the same problems and sometimes even suggesting similar theories, Indian thinkers were advancing novel formulations and argumentations. Problems that the Indian philosophers raised for consideration, but that their Western counterparts never did, include such matters as the origin (*utpatti*) and apprehension (*jñapti*) of truth (*prāmāṇya*). Problems that the Indian philosophers for the most part ignored but that helped shape Western philosophy include the question of whether knowledge arises from experience or from reason and distinctions such as that between analytic and synthetic judgments or between contingent and necessary truths. Indian thought, therefore, provides the historian of Western philosophy with a point of view that may supplement that gained from Western thought. A study of Indian thought, then, reveals certain inadequacies of Western philosophical thought and makes clear that some concepts and distinctions may not be as inevitable as they may otherwise seem. In a similar manner, knowledge of Western thought gained by Indian philosophers has also been advantageous to them.

Vedic hymns, Hindu scriptures dating from the 2nd millennium BC, are the oldest extant record from India of the process by which the human mind makes its gods and of the deep psychological processes of mythmaking leading to profound cosmological concepts. The *Upaniṣad*s (Hindu philosophical treatises) contain one of the first conceptions of a universal, all-pervading, spiritual reality leading to a radical monism (absolute nondualism, or the essential unity of matter and spirit). The *Upaniṣad*s also contain early speculations by Indian philosophers about nature, life, mind, and the human body, not to speak of ethics and social philosophy. The classical, or orthodox, systems (*darśana*s) debate, sometimes with penetrating insight and often with a degree of repetition that can become tiresome to some, such matters as the status of the finite individual; the distinction as well as the relation between the body, mind, and the self; the nature of knowledge and the types of valid knowledge; the nature and origin of truth; the types of entities that may be said to exist; the relation of realism to idealism; the problem of whether universals or relations are basic; and the very important problem of *mokṣa*, or salvation—its nature and the paths leading up to it.

GENERAL CHARACTERISTICS OF INDIAN PHILOSOPHY

Common concerns. The various Indian philosophies contain such a diversity of views, theories, and systems that it is almost impossible to single out characteristics that are common to all of them. Acceptance of the authority of the Vedas characterizes all the orthodox (*āstika*) systems, but not the unorthodox (*nāstika*) systems, such as Cārvāka (radical materialism), Buddhism, and Jainism. Moreover, even when philosophers professed allegiance to the Vedas, their allegiance did little to fetter the freedom of their speculative ventures. On the contrary, the acceptance of the authority of the Vedas was a convenient way for a philosopher's views to become acceptable to the orthodox, even if a thinker introduced a wholly new idea. Thus, the Vedas could be cited to corroborate a wide diversity of views; they were used by the Vaiśeṣika thinkers (*i.e.,* those who believe in ultimate particulars, both individual souls and atoms) as much as by the Advaita (monist) philosophers.

Authority of the Vedas and the ideal of *mokṣa*

In most Indian philosophical systems, the acceptance of the ideal of *mokṣa,* like allegiance to the authority of the scriptures, was only remotely connected with the systematic doctrines that were being propounded. Many epistemological, logical, and even metaphysical doctrines were debated and decided on purely rational grounds that did not directly bear upon the ideal of *mokṣa.* Only the Vedānta ("end of the Vedas") philosophy and the Sāṃkhya (a system that accepts a real matter and a plurality of the individual souls) philosophy may be said to have a close relationship to the ideal of *mokṣa.* The logical systems—Nyāya, Vaiśeṣika, and Pūrva-mīmāṃsā—are only very remotely related. Also, both the philosophies and other scientific treatises, including even the *Kāma-sūtra* ("Aphorisms on Love") and the *Arthaśāstra* ("Treatise on Material Gain"), recognized the same ideal and professed their efficacy for achieving it.

When Indian philosophers speak of intuitive knowledge, they are concerned with making room for it and demonstrating its possibility, with the help of logic—and there, as far as they are concerned, the task of philosophy ends. Indian philosophers do not seek to justify religious faith; philosophic wisdom itself is accorded the dignity of religious truth. Theory is not subordinated to practice, but theory itself, as theory, is regarded as being supremely worthy and efficacious.

Three basic concepts form the cornerstone of Indian philosophical thought: the self, or soul (*ātman*), works (*karma,* or *karman*), and salvation (*mokṣa*). Leaving the Cārvākas aside, all Indian philosophies concern themselves with these three concepts and their interrelations, though this is not to say that they accept the objective validity of these concepts in precisely the same manner. Of these, the concept of *karma,* signifying moral efficacy of human actions, seems to be the most typically Indian. The concept of *ātman,* not altogether absent in Western thought, corresponds, in a certain sense, to the Western concept of a transcendental or absolute spirit self—important differences notwithstanding. The concept of *mokṣa* as the concept of the highest ideal has likewise been one of the concerns of Western thought, especially during the Christian Era, though it probably has never been as important as for the Hindu mind. Most Indian philosophies assume that *mokṣa* is possible, and the "impossibility of *mokṣa*" (*anirmokṣa*) is regarded as a material fallacy likely to vitiate a philosophical theory.

In addition to *karma,* the lack of two other concerns further differentiates Indian philosophical thought from Western thought in general. Since the time of the Greeks, Western thought has been concerned with mathematics, and, in the Christian Era, with history. Neither mathematics nor history has ever raised philosophical problems for the Indian. In the lists of *pramāṇa*s, or ways of knowing accepted by the different schools, there is none that includes mathematical knowledge or historical knowledge. Possibly connected with their indifference toward mathematics is the significant fact that Indian philosophers have not developed formal logic. The theory of the syllogism (a valid deductive argument having two premises and a conclusion) is, however, developed, and much sophistication has been achieved in logical theory. Indian logic offers an instructive example of a logic of cognitions (*jñānāni*) rather than of abstract propositions—a logic not sundered and kept isolated from psychology and epistemology, because it is meant to be the logic of man's actual striving to know what is true of the world.

Indifference to mathematics and history as philosophical problems

Forms of argument and presentation. There is, in relation to Western thought, a striking difference in the manner in which Indian philosophical thinking is presented as well as in the mode in which it historically develops. Out of the presystematic age of the Vedic hymns and the *Upaniṣad*s and many diverse philosophical ideas current in the pre-Buddhistic era, there emerged with the rise of the age of the *sūtra*s (aphoristic summaries of the main points of a system) a neat classification of systems (*darśana*s), a classification that was never to be contradicted and to which no further systems are added. No new school was founded, no new *darśana* came into existence. But this conformism, like conformism to the Vedas, did not check the rise of independent thinking, new innovations, or original insights. There is, apparently, an underlying assumption in the Indian tradition that no individual can claim to have seen the truth for the first time and, therefore, that an individual can only explicate, state, and defend in a new form a truth that had been seen, stated, and defended by countless others before him: hence the tradition of expounding one's thoughts by affiliating oneself to one of the *darśana*s.

If one is to be counted as a great master (*ācārya*), one has to write a commentary (*bhāṣya*) on the *sūtra*s of the *darśana* concerned, or one must comment on one of the *bhāṣya*s and write a *ṭīkā* (subcommentary). The usual order is *sūtra–bhāṣya–vārttika* (collection of critical notes)–*ṭīkā*. At any stage, a person may introduce a new and original point of view, but at no stage can he claim originality for himself. Not even an author of the *sūtra*s could do that, for he was only systematizing the thoughts and insights of countless predecessors. The development of Indian philosophical thought has thus been able to combine, in an almost unique manner, conformity to tradition and adventure in thinking.

ROLES OF SACRED TEXTS, MYTHOLOGY, AND THEISM

The role of the sacred texts in the growth of Indian philosophy is different in each of the different systems. In those systems that may be called *adhyātmavidyā,* or sciences of spirituality, the sacred texts play a much greater role than they do in the logical systems (*ānvīkṣikīvidyā*). In the case of the former, Śaṅkara, a leading Advaita Vedānta philosopher (*c.* 788–820), perhaps best laid down the principles: reasoning should be allowed freedom only as long as it does not conflict with the scriptures. In matters regarding supersensible reality, reasoning left to itself cannot deliver certainty, for, according to Śaṅkara, every thesis established by reasoning may be countered by an opposite thesis supported by equally strong, if not stronger, reasoning. The sacred scriptures, embodying as they do the results of intuitive experiences of seers, therefore, should be accepted as authoritative, and reasoning should be made subordinate to them.

Whereas the sacred texts thus continued to exercise some influence on philosophical thinking, the influence of mythology declined considerably with the rise of the systems. The myths of creation and dissolution of the universe persisted in the theistic systems but were transformed into metaphors and models. With the Nyāya (problem of knowledge)–Vaiśeṣika (analysis of nature) systems, for example, the model of a potter making pots determined much philosophical thinking, as did that of a magician conjuring up tricks in the Advaita (nondualist) Vedānta. The *nirukta* (etymology) of Yāska, a 5th-century-BC Sanskrit scholar, tells of various attempts to interpret difficult Vedic mythologies: the *adhidaivata* (pertaining to the deities), the *aitihāsika* (pertaining to the tradition), the *adhiyajña* (pertaining to the sacrifices), and the *ādhyātmika* (pertaining to the spirit). Such interpretations apparently prevailed in the *Upaniṣad*s; the myths were turned into symbols, though some of them persisted as models and metaphors.

The issues of theism and atheism

The issue of theism vis-à-vis atheism, in the ordinary senses of the English words, played an important role in Indian thought. The ancient Indian tradition, however, classified the classical systems (*darśanas*) into orthodox (*āstika*) and unorthodox (*nāstika*). *Āstika* does not mean "theistic," nor does *nāstika* mean "atheistic." Pāṇini, a 5th-century-BC grammarian, stated that the former is one who believes in a transcendent world (*asti paralokah*) and the latter is one who does not believe in it (*nasti paralokah*). *Āstika* may also mean one who accepts the authority of the Vedas; *nāstika* then means one who does not accept that authority. Not all among the *āstika* philosophers, however, were theists, and even if they were, they did not all accord the same importance to the concept of God in their systems. The Sāṃkhya system did not involve belief in the existence of God, without ceasing to be *āstika,* and Yoga (a mental–psychological–physical meditation system) made room for God not on theoretical grounds but only on practical considerations. The Pūrva-Mīmāṃsā of Jaimini, the greatest philosopher of the Mīmāṃsā school, posits various deities to account for the significance of Vedic rituals but ignores, without denying, the question of the existence of God. The Advaita Vedānta of Śaṅkara rejects atheism in order to prove that the world had its origin in a conscious, spiritual being called Īśvara, or God, but in the long run regards the concept of Īśvara as a concept of lower order that becomes negated by a metaphysical knowledge of Brahman, the absolute, nondual reality. Only the non-Advaita schools of Vedānta and the Nyāya-Vaiśeṣika remain zealous theists, and of these schools, the god of the Nyāya-Vaiśeṣika school does not create the eternal atoms, universals, or individual souls. For a truly theistic conception of God, one has to look to the non-Advaita schools of Vedānta, the Vaiṣṇava, and the Śaiva philosophical systems. Whereas Hindu religious life continues to be dominated by these last-mentioned theistic systems, the philosophies went their own ways, far removed from that religious demand.

A GENERAL HISTORY OF DEVELOPMENT AND CULTURAL BACKGROUND

The three periods of Indian philosophy

S.N. Dasgupta, a 20th-century Indian philosopher, has divided the history of Indian philosophy into three periods: the prelogical (up to the beginning of the Christian Era), the logical (from the beginning of the Christian Era up to the 11th century AD), and the ultralogical (from the 11th century to the 18th century). What Dasgupta calls the prelogical stage covers the pre-Mauryan and the Mauryan periods (*c.* 321–185 BC) in Indian history. The logical period begins roughly with the Kuṣāṇas (1st–2nd centuries AD) and finds its highest development during the Gupta era (3rd–5th centuries AD) and the age of imperial Kanauj (7th century AD).

The prelogical period. In its early prelogical phase, Indian thought, freshly developing in the Indian subcontinent, actively confronted and assimilated the diverse currents of pre-Aryan and non-Aryan elements in the native culture that the Aryans sought to conquer and appropriate. The marks of this confrontation are to be noted in every facet of Indian religion and thought: in the Vedic hymns in the form of conflicts, with varying fortunes, between the Aryans and the non-Aryans; in the conflict between a positive attitude toward life that is interested in making life fuller and richer and a negative attitude emphasizing asceticism and renunciation; in the great variety of skeptics, naturalists, determinists, indeterminists, accidentalists, and no-soul theorists that filled the Ganges Plain; in the rise of the heretical, unorthodox schools of Jainism and Buddhism protesting against the Vedic religion and the Upaniṣadic theory of *ātman;* and in the continuing confrontation, mutually enriching and nourishing, that occurred between the Brahmanic (Hindu priestly) and Buddhist logicians, epistemologists, and dialecticians. The Aryans, however, were soon followed by a host of foreign invaders, Greeks, Śakas and Hūṇas from Central Asia, Pushtans (Pathans), Mongols, and Mughals (Muslims). Both religious thought and philosophical discussion received continuous challenges and confrontations. The resulting responses have a dialectical character: sometimes new ideas have been absorbed and orthodoxy has been modified; sometimes orthodoxy has been strengthened and codified in order to be preserved in the face of the dangers of such confrontation; sometimes, as in the religious life of the Christian Middle Ages, bold attempts at synthesis of ideas have been made. Nevertheless, through all the vicissitudes of social and cultural life, Brahmanical thought has been able to maintain a fairly strong current of continuity.

In the chaotic intellectual climate of the pre-Mauryan era, there were skeptics (*ajñānikah*) who questioned the possibility of knowledge. There were also materialists, the chief of which were the Ājīvikas (deterministic ascetics) and the Lokāyatas (the name by which Cārvāka doctrines—denying the authority of the Vedas and the soul—are generally known). Furthermore, there existed the two unorthodox schools of *yadṛchhāvāda* (accidentalists) and *svabhāvavāda* (naturalists), who rejected the supernatural. Kapila, the legendary founder of the Sāṃkhya school, supposedly flourished during the 7th century BC. Pre-Mahāvīra Jaina ideas were already in existence when Mahāvīra (flourished 6th century BC), the founder of Jainism, initiated his reform. Gautama the Buddha (flourished 6th–5th centuries BC) apparently was familiar with all of these intellectual ideas and was as dissatisfied with them as with the Vedic orthodoxy. He sought to forge a new path—though not new in all respects—that was to assure blessedness to man. Orthodoxy, however, sought to preserve itself in a vast *Kalpa-* (ritual) *sūtra* literature—with three parts: the *Śrauta-*, based on *śruti* (revelation); the *Gṛhya-*, based on *smṛti* (tradition); and the *Dharma-*, or rules of religious law, *sūtra*s—whereas the philosophers tried to codify their doctrines in systematic form, leading to the rise of the philosophical *sūtra*s. Though the writing of the *sūtra*s continued over a long period, the *sūtra*s of most of the various *darśana*s probably were completed between the 6th and 3rd centuries BC. Two of the *sūtra*s appear to have been composed in the pre-Maurya period, but after the rise of Buddhism; these works are the *Mīmāṃsāsūtra*s of Jaimini (*c.* 400 BC) and the *Vedānta-sūtra*s of Bādarāyaṇa (*c.* 500–200 BC).

Development of philosophical–religious literature and philosophical systems

The Maurya period brought, for the first time, a strong centralized state. The Greeks had been ousted, and a new self-confidence characterized the beginning of the period. This seems to have been the period in which the epics *Mahābhārata* and *Rāmāyaṇa* were initiated, though their composition went on through several centuries before they took the forms they now have. Manu, a legendary lawgiver, codified the *Dharma-śāstra;* Kauṭilya, a minister of King Candragupta Maurya, systematized the science of political economy (*Arthaśāstra*); and Patañjali, an ancient author or authors, composed the *Yoga-sūtra*s. Brahmanism tried to adjust itself to the new communities and cultures that were admitted into its fold: new gods—or rather, old Vedic gods that had been rejuvenated—were worshipped; the Hindu trinity of Brahmā (the creator), Vishnu (the preserver), and Śiva (the destroyer) came into being; and the Pāśupata (Śaivite), Bhāgavata (Vaiṣṇavite), and the Tantra (esoteric meditative) systems were initiated. The *Bhagavadgītā*—the most famous work of this period—symbolized the spirit of the creative synthesis of the age. A new ideal of *karma* as opposed to the more ancient one of renunciation was emphasized. Orthodox notions were reinterpreted and given a new symbolic meaning, as, for example, the *Gītā* does with the notion of *yajña* ("sacrifice"). Already in the pre-Christian era, Buddhism had split up into several major sects, and the foundations for the rise of Mahāyāna ("Greater Vehicle") Buddhism had been laid.

The logical period. The logical period of Indian thought began with the Kusanas (1st–2nd centuries). Gautama (author of the *Nyāya-sūtra*s; probably flourished at the beginning of the Christian Era) and his 5th-century commentator Vātsyāyana established the foundations of the Nyāya as a school almost exclusively preoccupied with logical and epistemological issues. The Mādhyamika ("Middle Way"), or Śūnyavāda ("Voidist") school of Buddhism, arose and the thought of Nāgārjuna (*c.* 200), the great propounder of Śūnyavāda (dialectical thinking), reached great heights.

Though Buddhist logic in the strict sense of the term had not yet come into being, a logical style of philosophizing was in existence in such schools of thought.

During the reign of the Guptas, there was a revival of Brahmanism of a gentler and more refined form. Vaiṣṇavism of the Vāsudeva cult, centring on the prince-god Krishna and advocating renunciation by action, and Śaivism prospered, along with Buddhism and Jainism. Both the Mahāyāna and the Hinayāna ("Lesser Vehicle"), or Theravāda ("Way of the Elders"), schools flourished. The most notable feature, however, was the rise of the Buddhist Yogācāra school, of which Asaṅga (4th century AD) and his brother Vasubandhu were the great pioneers. Toward the end of the 5th century, Dignāga, a Buddhist logician, wrote the *Pramāṇasamuccaya* ("Compendium of the Means of True Knowledge"), a work that laid the foundations of Buddhist logic.

The greatest names of Indian philosophy belong to the post-Gupta period from the 7th to the 10th century. At that time Buddhism was on the decline and the Tantric cults were rising, a situation that led to the development of the tantric forms of Buddhism. Śaivism was thriving in Kashmir, and Vaiṣṇavism in the southern part of India. The great philosophers Mīmāṃsākas Kumārila (7th century), Prabhākara (7th–8th centuries), Maṇḍana Miśra (8th century), Śālikanātha (9th century), and Pārthasārathi Miśra (10th century) belong to this age. The greatest Indian philosopher of the period, however, was Śaṅkara. All of these men defended Brahmanism against the "unorthodox" schools, especially against the criticisms of Buddhism. The debate between Brahmanism and Buddhism was continued, on a logical level, by philosophers of the Nyāya school—Uddyotakara, Vācaspati Miśra, and Udayana (Udayanācārya).

The ultralogical period. Muslim rule in India had consolidated itself by the 11th century, by which time Buddhism, for all practical purposes, had disappeared from the country. Hinduism had absorbed Buddhist ideas and practices and reasserted itself, with the Buddha appearing in Hindu writings as an incarnation of Vishnu. The Muslim conquest created a need for orthodoxy to readjust itself to a new situation. In this period the great works on Hindu law were written. Jainism, of all the "unorthodox" schools, retained its purity, and great Jaina works, such as Devasūri's *Pramāṇanayatattvālokālaṃkāra* ("The Ornament of the Light of Truth of the Different Points of View Regarding the Means of True Knowledge," 12th century AD) and Prabhāchandra's *Prameyakamalamārtaṇḍa* ("The Sun of the Lotus of the Objects of True Knowledge," 11th century AD), were written during this period. Under the Cōla (Chola) kings (*c.* 850–1279) and later in the Vijayanagara kingdom (which, along with Mithilā in the north, remained strongholds of Hinduism until the middle of the 16th century), Vaiṣṇavism flourished. The philosopher Yamunācārya (flourished AD 1050) taught the path of *prapatti,* or complete surrender to God. The philosophers Rāmānuja (11th century), Madhva, and Nimbārka (*c.* 12th century) developed theistic systems of Vedānta and severely criticized Śaṅkara's Advaita Vedānta.

Creativity in logic, epistemology, and religious life

Toward the end of the 12th century, creative work of the highest order began to take place in the fields of logic and epistemology in Mithilā and Bengal. The 12th–13th-century philosopher Gaṅgesa's *Tattvacintāmaṇi* ("The Jewel of Thought on the Nature of Things") laid the foundations of the school of Navya-Nyāya ("New-Nyāya"). Four great members of this school were Pakṣadhara Miśra of Mithilā, Vāsudeva Sārvabhauma (16th century), his disciple Raghunātha Śiromaṇi (both of Bengal), and Gadādhara Bhaṭṭācāryya.

Religious life was marked by the rise of great mystic saints, chief of which are Rāmānanda, Kabīr, Caitanya, and Gurū Nānak, who emphasized the path of *bhakti,* or devotion, a wide sense of humanity, freedom of thought, and a sense of unity of all religions. Somewhat earlier than these were the great Muslim Ṣūfī (mystic) saints, including Khwāja Mu'in-ud-Din Ḥasan, who emphasized asceticism and taught a philosophy that included both love of God and love of humanity.

The British period in Indian history was primarily a period of discovery of the ancient tradition (*e.g.,* the two histories by Radhakrishnan, scholar and president of India from 1962 to 1967, and S.N. Dasgupta) and of comparison and synthesis of Indian philosophy with the philosophical ideas from the West. Among modern creative thinkers have been Mahatma Gandhi, who espoused new ideas in the fields of social, political, and educational philosophy; Sri Aurobindo, an exponent of a new school of Vedānta that he calls Integral Advaita; and K.C. Bhattacharyya, who developed a phenomenologically oriented philosophy of subjectivity that is conceived as freedom from object.

Historical development of Indian philosophy

PRESYSTEMATIC PHILOSOPHY

Śruti and the nature of authority. All "orthodox" philosophies can trace their basic principles back to some statement or other in the Vedas. The Vedānta schools, especially, had an affiliation with the authority of *śruti,* and the school of Mīmāṃsā concerned itself chiefly with the questions of interpreting the sacred texts. The Hindu tradition regards the Vedas as being *apauruṣeya—i.e.,* as not composed by any person. Sāyana, a famous Vedic commentator, said that this means an absence of a human author. For Sāyana, the eternality of the Vedas is like that of space and time; man does not experience their beginning or end. But they are, in fact, created by Brahmā, the supreme creator. For the Advaita Vedānta, because no author of the Vedas is mentioned, an unbroken chain of Vedic teachers is quite conceivable, so that the scriptures bear testimony to their own eternality. The authoritative character of *śruti* may then be deduced from the fact that it is free from any fault (*doṣa*), or limitation, which characterizes human words. Furthermore, the Vedas give knowledge about things—whether *dharma* (what ought to be done) or Brahman (the absolute reality)—which cannot be known by any other empirical means of knowledge. The authority of the Vedas cannot, therefore, be contradicted by any empirical evidence. Later logicians of the "orthodox" schools sought to give these arguments precision and logical rigour.

Significance of the Vedic hymns

The Vedic hymns (*mantras*) seem to be addressed to gods and goddesses (*deva,* one who gives knowledge or light), who are personifications of natural forces and phenomena (Agni, the fire god; Indra, the rain god; Vāyu, the wind god). But there are gods not identifiable with such phenomena (*e.g.,* Aditi, the infinite mother of all gods; Mitra, the friend; Varuṇa, the guardian of truth and righteousness; Viśvakarman, the all-maker; *śraddhā,* faith). Also, the hymns show an awareness of the unity of these deities, of the fact that it is one God who is called by different names. The famed conception of *ṛta*—meaning at once natural law, cosmic order, moral law, and the law of truth—made the transition to a monistic view of the universe as being but a manifestation of one reality about which the later hymns continue to raise fundamental questions in a poignant manner, without, however, suggesting any dogmatic answer.

Development of the notion of transmigration. The hymns may, in general, be said to express a positive attitude toward human life and to show interest in the full enjoyment of life here and hereafter rather than an anxiety to escape from it. The idea of transmigration and the conception of the different paths and worlds traversed by good men and those who are not good—*i.e.,* the world of Vishnu and the realm of Yama—are found in the Vedas. The chain of rebirth as a product of ignorance and the conception of release from this chain as the greatest good of the spiritual life are markedly absent in the hymns.

Meanings of Brahman–*ātman*

Origin of the concept of Brahman and ātman. The *Upaniṣads* answer the question "Who is that one Being?" by establishing the equation Brahman = *ātman.* Brahman—meaning now that which is the greatest, than which there is nothing greater, and also that which bursts forth into the manifested world, the one Being of which the hymn of creation spoke—is viewed as nothing but *ātman,* identifiable as the innermost self in man but also, in reality, the innermost self in all beings. Both the words gain a new, extended, and spiritual significance through this identifica-

tion. *Ātman* was originally used to mean breath, the vital essence, and even the body. Later etymologizing brought out several strands in its meaning: that which pervades (*yad āpnoti*), that which gives (*yadādatte*), that which eats (*yad atti*), and that which constantly accompanies (*yacca asya santato bhavam*). Distinctions were made between the bodily self, the vital self, the thinking self, and the innermost self, whose nature is bliss (*ānanda*), the earlier ones being sheaths (*kośas*) covering the innermost being. Distinctions were sometimes drawn between the waking (*jāgrat*), dreaming (*svapna*), and dreamless-sleep (*suṣupti*) states of the self, and these three are contrasted with the fourth, or transcendent (*turīya*), state that both transcends and includes them all. The identification of the absolute reality underlying the universe with the innermost being within the human person resulted in a spiritualization of the former concept and a universalization of the latter. This final conception of Brahman or *ātman* received many different explications from different teachers in the *Upaniṣads*, some of which were negative in character (*neti neti,* "not this, not this") while others positively affirmed the all-pervasiveness of Brahman. But there were still others who insisted on both the transcendence and immanence of Brahman in the universe. Brahman is also characterized as infinite, truth, and knowledge and as existence, consciousness, and bliss.

The principles underlying macrocosm and microcosm. Though the objective and the subjective, the macrocosm (universal) and the microcosm (individual), came to be identified according to their true essences, attempts were made to correlate different macrocosmic principles with corresponding microcosmic principles. The manifested cosmos was correlated with the bodily self; the soul of the world, or Hiraṇyagarbha, with the vital self; and Īśvara, or God as a self-conscious being, with the thinking self. The transcendent self and the Brahman as bliss are not correlates but rather are identical.

EARLY BUDDHIST DEVELOPMENTS

Background. Buddhism was not a completely new phenomenon in the religious history of India; it was built upon the basis of ideas that were already current, both Brahmanic and non-Aryan. Protests against the Brahmanic doctrines of *ātman, karma,* and *mokṣa* were being voiced in the 6th century BC, prior to the Buddha, by various schools of thought: by naturalists, such as Pūraṇa ("The Old One") Kassapa, who denied both virtue and vice (*dharma* and *adharma*) and thus all moral efficacy of human deeds; by determinists, such as the Ājīvika Makkhali Gosāla, who denied sin and freedom of will; and by materialists, such as Ajita Keśakambalin, who, besides denying virtue, vice, and afterlife, resolved man's being into material elements, Nigantha Nātaputta, who believed in salvation by an ascetic life of self-discipline and hence in the efficacy of deeds and the possibility of omniscience, and, finally, Sanjaya Belathiputta, the skeptic, who, in reply to the question "Is there an afterlife?" would not say "It is so" or "It is otherwise," nor would he say "It is not so" or "It is not not so."

Of these six, the Jaina tradition identifies Nigantha with Mahāvīra; the designation "Ājīvika" is applied, in a narrow sense, to the followers of Makkhali and in a loose sense to all nonorthodox sects other than the Jainas—the skeptics and the Lokāyatas.

Buddhism, Jainism, and the Ājīvikas rejected, in common, the sacrificial polytheism of the *Brāhmaṇas* and the monistic mysticism of the *Upaniṣads*. All three of them recognized the rule of natural law in the universe. Buddhism, however, retained the Vedic notions of *karma* and *mokṣa,* though rejecting the other fundamental concept of *ātman.*

The four noble truths and the nature of suffering. In such an intellectual climate Gautama the Buddha taught his four noble truths: (1) *duḥkha* (generally but misleadingly translated as "suffering"); (2) the origination of *duḥkha* (*duḥkhasamudāya*); (3) the cessation of *duḥkha;* and finally (4) the way leading to the cessation.

Although the word *duḥkha* in common parlance means suffering, its use by Gautama was meant to include both pleasure and pain, both happiness and suffering. There are three aspects of this conception: *duḥkha* as suffering in the ordinary sense; *duḥkha* arising out of the impermanence of things, even of a state of pleasure; and *duḥkha* in the sense of five aggregates meaning that the "I" constituted by any individual is nothing but a totality of five aggregates—*i.e.,* form, feeling, conception, disposition, and consciousness. In brief, whatever is noneternal—*i.e.,* whatever is subject to the law of causality—is characterized by *duḥkha;* for Gautama, this is the human situation. One who recognizes the nature of *duḥkha* also knows its causes. *Duḥkha* arises out of craving (*tṛṣṇā*), craving arises out of sensation (*vedanā*), and sensation arises out of contact (*sparśa*), so that man is faced with a series of conditions leading back to ignorance (*avidyā*)—a series in which the rise of each succeeding member depends upon the preceding one (*pratītyasamutpāda*).

The concept of *duḥkha*

The path of liberation: methods of eightfold path. The four noble truths follow the golden mean between the two extremes of sensual indulgence and ascetic self-torture, both of which Gautama rejected as spiritually useless. Only the middle path consisting in the eight steps—called the eightfold path—leads to enlightenment and to Nirvāṇa. The eight steps are (1) right views, (2) right intention, (3) right speech, (4) right action, (5) right livelihood, (6) right effort, (7) right mindfulness, and (8) right concentration. Of these eight, steps 3, 4, and 5 are grouped under right morality (*śīla*); steps 6, 7, and 8 under right concentration (*samādhi*); and steps 1 and 2 under right wisdom (*prajñā*).

The concepts of selflessness and Nirvāṇa. Two key notions, even in early Buddhism, are those of *anātman* (Sanskrit: "no-self"; Palī *anattā*) and Nirvāṇa. The Buddha apparently wanted his famed doctrine of *anātman* to be a phenomenological account of how things are rather than a theory. In his discourse to the wandering monk Vacchagotta, he rejected the theories of both eternalism (*śāśvatavāda*) and annihilationism (*ucchedavāda*). The former, he stated, would be incompatible with his thesis that all laws (*dharmas*; Palī *dhammas*) are selfless (*sabbe dhammā anattā*); the latter would be significant only if one had a self that is no more in existence. Thus, by not taking sides with the metaphysicians, the Buddha described how the consciousness "I am" comes to constitute itself in the stream of consciousness out of the five aggregates of form, feeling, conception, disposition, and consciousness. The doctrine of "no-self" actually has two aspects: as applied to *pudgala,* or the individual person, and as applied to the *dhammas*, or the elements of being. In its former aspect, it asserts the fact that an individual is constituted out of five aggregates; in its latter aspect it means the utter insubstantiality of all elements. Intuitive realization of the former truth leads to the disappearance of passions and desires, realization of the latter removes all misconceptions about the nature of things in general. The former removes the "covering of the passions" (*kleśāvaraṇa*); the latter removes "the concealment of things" (*jñeyāvaraṇa*). Together, they result in Nirvāṇa.

Both negative and positive accounts of Nirvāṇa are to be found in the Buddha's teachings and in early Buddhist writings. Nirvāṇa is a state of utter extinction, not of existence, but of passions and suffering; it is a state beyond the chain of causation, a state of freedom and spontaneity. It is in addition a state of bliss. Nirvāṇa is not the result of a process; were it so, it would be but another perishing state. It is the truth—not, however, an eternal, everlasting substance like the *ātman* of the *Upaniṣads*, but the truth of utter selflessness and insubstantiality of things, of the emptiness of the ego, and of the impermanence of all things. With the realization of this truth, ignorance is destroyed, and, consequently, all craving, suffering, and hatred is destroyed with it (see also the article BUDDHISM).

The concept of Nirvāṇa

THE PHILOSOPHICAL PORTIONS OF THE "MAHĀBHĀRATA"

The great epic *Mahābhārata* represents the attempt of Vedic Brahmanism to adjust itself to the new circumstances reflected in the process of the aryanization (integration of Aryan beliefs, practices, and institutions) of the various non-Aryan communities. Many diverse trends of

religious and philosophical thought have thus been synthesized in this work (see also HINDUISM).

"Mokṣadharma." *Proto-Sāṃkhyan texts.* In its philosophical views, the epic contains an early version of Sāṃkhya (a belief in real matter and the plurality of individual souls), which is prior to the classical Sāṃkhya of Īśvarakṛṣṇa, a 3rd-century philosopher. The chapter on "Mokṣadharma" in Book 12 of the *Mahābhārata* is full of such proto-Sāṃkhya texts. Mention is made of four main philosophical schools: Sāṃkhya-Yoga, taught by Kapila (a sage living before the 6th century BC); Pāñcarātra, taught by Vishnu; the Vedas; and Pāśupata ("Lord of Creatures"), taught by Śiva. Belonging to the Pāñcarātra school, the epic basically attempts to accommodate certain presystematic Sāṃkhya ideas into the Bhāgavata faith. Sāṃkhya and Yoga are sometimes put together, sometimes distinguished. Several different schemata of the 25 principles (*tattvas*) of the Sāṃkhya are recorded. One common arrangement is that of eight productive forms of *prakṛti* (the unmanifest, intellect, egoism, and five fine elements: sound, smell, form or colour, taste, and touch) and 16 modifications (five organs of perception, five organs of action, mind, and five gross elements: ether, earth, fire, water, and air), and *puruṣa* (man). An un-Sāṃkhyan element is the 26th principle: Īśvara, or the supreme Lord. One notable result is the identification of the four living forms (*vyūhas*) of the Pāñcarātra school with four Sāṃkhya principles: Vāsudeva with spirit, Saṃkarṣaṇa with individual soul, Pradyumna with mind, and Aniruddha with the ego-sense.

Non-Sāṃkhyan texts. Besides the Sāṃkhya-Yoga, which is in the foreground of the epic's philosophical portions, there are Vedānta texts emphasizing the unity of spirits and theistic texts emphasizing not only a personal deity but also the doctrine of avatar (*avatāra*), or incarnation. The Vāsudeva-Krishna cult characterizes the theistic part of the epic.

Early theories of kingship and state. In the *Śānti Parvan* ("Book of Consolation," 12th book) of the *Mahābhārata,* there is also a notable account of the origin of kingship and of *rājadharma,* or the *dharma* (law) of the king as king. Bhīṣma, who is discoursing, refers with approval to two different theories of the origin of kingship, both of which speak of a prior period in which there were no kings. According to one account, this age was a time characterized by insecurity for the weak and unlimited power for the strong; the other regards it as an age of peace and tranquillity. The latter account contains a theory of the fall of mankind from this ideal state, which led to a need for institutionalized power, or kingship; the former account leads directly from the insecurity of the prekingship era to the installation of king by the divine ruler for the protection and the security of mankind. Kingship is thus recognized as having a historical origin. The primary function of the king is that of protection, and *daṇḍanīti,* or the art of punishment, is subordinated to *rājadharma,* or *dharma* of the king. Though it recognizes a quasi-divinity of the king, the *Mahābhārata* makes the *dharma,* the moral law, superior to the king.

The "Bhagavadgītā." The *Bhagavadgītā* ("Divine Song" or "Song of the Lord") forms a part of *Mahābhārata* and deserves separate consideration by virtue of its great importance in the religious life and thought of the Hindus. Not itself a *śruti,* it has, however, been accorded the status of an authoritative text and is regarded as one of the sources of the Vedānta philosophy. At a theoretical level, it brings together Sāṃkhya metaphysics, Upaniṣadic monism, and a devotional theism of the Krishna-Vāsudeva cult. In its practical teaching, it steers a middle course between the "path of action" of the Vedic ritualism and the "path of renunciation" of the Upaniṣadic mysticism, and it accommodates all the three major "paths" to *mokṣa:* the paths of action (*karma*), devotion (*bhakti*), and knowledge (*jñāna*). This synthetic character of the work accounts for its great hold on the Hindu mind. The Hindu tradition treats it as one homogenous work, having the status of an *Upaniṣad.*

The paths of *mokṣa*

Neither performance of the duties prescribed in the scriptures nor renunciation of all action is conducive to the attainment of *mokṣa.* If the goal is freedom, then the best path to the goal is to perform one's duties with a spirit of nonattachment without caring for the fruits of one's actions and without the thought of pleasure or pain, profit or loss, or victory or failure, with a sense of equanimity and equality. The Kantian ethic of "duty for duty's sake" seems to be the nearest Western parallel to Krishna's (Kṛṣṇa's) teaching at this stage. But Krishna soon went beyond it, by pointing out that performance of action with complete nonattachment requires knowledge (*jñāna*) of the true nature of the self, its distinction from *prakṛti,* or Matter (the primeval stuff, not the world of matter perceived by the senses), with its three component elements (*sattva—i.e.,* tension or harmony; *rajas—i.e.,* activity; and *tamas—i.e.,* inertia), and of the highest self (*puruṣottama*), whose higher and lower aspects are Matter and finite individuals, respectively. This knowledge of the highest self or the supreme lord, however, would only require a devotional attitude of complete self-surrender and performance of one's duties in the spirit of offering to him. Thus, *karma-yoga* (yoga of works) is made to depend on *jñāna-yoga* (yoga of knowledge), and the latter is shown to lead to *bhakti-yoga* (yoga of devotion). Instead of looking upon Krishna's teaching as laying down alternative ways for different persons in accordance with their aptitudes, it would seem more logical to suppose that he taught the essential unity and interdependence of these ways. How one should begin is left to one's aptitude and spiritual makeup.

DOCTRINES AND IDEAS OF THE BUDDHIST "TIPIṬAKA"

In the *Tipiṭaka* (Pali: "The Three Baskets"; Sanskrit *Tripiṭaka*), collected and compiled 300 years after the Buddha's *mahāparinirvāṇa* (attainment of Buddhahood), at the council at Pāṭaliputra (3rd century BC), both the canonical and philosophical doctrines of early Buddhism were codified. *Abhidamma piṭaka,* the last of the *piṭakas,* has seven parts: *Dhammasaṅgaṇi,* which gives an enumeration of *dhammas,* or elements of existence; *Vibhaṅga,* which gives further analysis of the *dhammas*; *Dhatukathā,* which is a detailed classification, following many different principles, of the elements; *Puggalapaññatti,* which gives descriptions of individual persons according to stages of their development; *Kathāvatthu,* which contains discussions and refutation of other schools (of Buddhism); *Yamaka,* which derives its name from the fact that it deals with pairs of questions; and *Paṭṭhāna,* which gives an analysis of relations among the elements.

The notion of the *dhammas*

The key notion in all this is that of the *dhammas.* Because Buddhist philosophers denied any permanence, whether in outer nature or in inner life, they felt compelled to undertake a detailed, systematic, and complete listing and classification of the different elements that constitute both the external world and the mental, inner life. Each of these elements, except for the three elements that are not composed of parts (*i.e.,* space, or *ākāśa,* and the two cessations, Nirvāṇa and a temporary stoppage, in states of meditation, of the flow of passions, or *apratisaṃkhyānivodha*), is momentary. The primary object of this exhaustive analysis was an understanding not so much of outer nature as of the human person (*pudgala*). The human person, however, consists in material (*rūpa*) and mental (*nāma*) factors, which led to an account of the various elements of matter. The primary interest, nevertheless, is in man, who is regarded as an aggregate of various elements. The analysis of these components, together with the underlying denial of an eternal self, was supposed to provide the theoretical basis for the possibility of a good life conducive to the attainment of Nirvāṇa.

The individual person was analyzed into five aggregates (*skandhas*): material form (*rūpa*); feeling (*vedanā*); conception (*saṃjñā*); disposition (*saṃskāra*); and consciousness (*vijñāna*). Of these, the last four constitute the mental; the first alone is the material factor. The material is further analyzed into 28 states, the *saṃskāra* into 50 (falling into three groups: intellectual, affectional, and volitional), and the *vijñāna* into 89 kinds of states of consciousness. Another principle of classification leads to a list of 18 elements (*dhātus*): five sense organs, five objects of those

senses, mind, the specific object of mind, and six kinds of consciousness (visual, auditory, olfactory, gustatory, tactual, and purely mental). A third classification is into 12 bases (*āyatanas*), which is a list of six cognitive faculties and their objects. The Buddhist analysis of matter was in terms of sensations and sense data, to which the sense organs were also added. The analysis of mind was also in terms of corresponding modes of consciousness and their objects.

EARLY SYSTEM BUILDING

The history of the sūtra style. A unique feature of the development of Indian thought was the systematization of each school of thought in the form of *sūtras*, or extremely concise expressions, intended to reduce the doctrines of a science or of a philosophy into a number of memorizable aphorisms, formulas, or rules. The word *sūtra,* originally meaning "thread," came to mean such concise expressions. A larger work containing such *sūtras* also came to be called a *sūtra.* The aid of commentaries becomes indispensable for the understanding of the *sūtras*, and it is not surprising that philosophical composition took the form of commentaries and subcommentaries. The earliest *sūtras*, the *Kalpa-sūtras*, however, are not philosophical but ritualistic. These *Kalpa-sūtras* fell into three major parts: the *Śrauta-sūtras*, dealing with Vedic sacrifices; the *Gṛhya-sūtras*, dealing with the ideal life of a householder; and the *Dharma-sūtras*, dealing with moral injunctions and prohibitions.

Rise of the *sūtra* form of philosophical and ritualistic works

In the works of Pāṇini, a Hindu grammarian, the *sūtra* style reached a perfection never attained before and only imperfectly approximated by the later practitioners. The *sūtra* literature began before the rise of Buddhism, though the philosophical *sūtras* all seem to have been composed afterward. The Buddhist *sutta* (Palī form of the Sanskrit word *sūtra*) differs markedly in style and content from the Hindu *sūtra.* The *suttas* are rather didactic texts, discourses, or sermons, possibly deriving their name from the sense in which they carry the thread of the tradition of the Buddha's teachings.

The "Pūrva-mīmāṃsā-sūtras" and Śabara's commentary. The Pūrva-mīmāṃsā ("First Reflection"), or Karma-mīmāṃsā ("Study of [Ritual] Action"), is the system that investigates the nature of Vedic injunctions. Though this is the primary purpose of the system, this task also led to the development of principles of scriptural interpretation and, therefore, to theories of meaning and hermeneutics (critical interpretations). Jaimini, who composed *sūtras* about the 4th century BC, was critical of earlier Mīmāṃsā authors, particularly of one Bādari, to whom is attributed the view that the Vedic injunctions are meant to be obeyed without the expectation of benefits for oneself. According to Jaimini, Vedic injunctions do not merely prescribe actions but also recommend these actions as means to the attainment of desirable goals. For both Jaimini and Śabara (3rd century), his chief commentator, performance of the Vedic sacrifices is conducive to the attainment of heaven; both emphasize that nothing is a duty unless it is instrumental to happiness in the long run.

Hermeneutical principles of Jaimini

Jaimini's central concern is *dharma,* which is defined as the desired object (*artha*), whose desirability is testified only by the injunctive statements of the scriptures (*codanā-lakṣaṇo*). In order to substantiate the implied thesis that what ought to be done—*i.e., dharma*—cannot be decided by either perception or reasoning, Jaimini proceeds to a discussion of the nature of ways of knowing. Because perceptual knowledge arises from contact of the sense organs with reality that is present, *dharma* that is not an existent reality but a future course of action cannot possibly be known by sense-experience. Reasoning based on such sense-experience is for the same reason useless. Only injunctive statements can state what ought to be done. Commands made by finite individuals are not reliable, because the validity of what they say depends upon the presumption that the persons concerned are free from those defects that render one's words dependable. Therefore, only the injunctions contained in the scriptures—which, according to Mīmāṃsā and the Hindu tradition, are not composed by any finite individual (*apauruṣeya*)—are the sources of all valid knowledge of *dharma.* The Mīmāṃsā rejects the belief that the scriptures are utterances of God. The words themselves are authoritative. In accordance with this thesis, Jaimini developed the theory that the relation between words and their meanings is natural (*autpattikastu śabdasyārthena sambandhah,* or "the relation of word to its meaning is eternal") and not conventional, that the primary meaning of a word is a universal (which is also eternal), that in a sentence the principal element is the verb, and that the principal force of the verb is that which specifically belongs to the verb with an optative ending and which instigates a person to take a certain course of action in order to effect the desired end.

Though this theory provided the Mīmāṃsā with a psychological and semantic technique for interpreting the sentences of the scriptures that are clearly in the injunctive form, there are also other kinds of sentences: prayers, glorifications, those referring to a thing by a name, and prohibitions. Attempts were therefore made to show how each one of these types of sentences bears, directly or indirectly, on the central, injunctive texts. Furthermore, a systematic classification of the various forms of injunctions is undertaken: those that indicate the general nature of an action, those that show the connection of a subsidiary rite to the main course of action, those that suggest promptness in performance of the action, and those that indicate the right to enjoy the results to be produced by the course of action enjoined.

The commentary of Śabara elaborated on the epistemological themes of the *sūtras*; in particular, Śabara sought to establish the intrinsic validity of experiences and traced the possibility of error to the presence of defects in the ways of knowing. He also critically examined Buddhist subjective idealism and the theory of utter emptiness of things and proved the existence of soul as a separate entity that enjoys the results of one's actions in this or the next life.

The "Vedānta-sūtras." *Relation to the "Mīmāṃsā-sūtras."* Along with Bādari and Jaimini, Bādarāyaṇa, a contemporary of Jaimini, was the other major interpreter of Vedic thought. Just as the *Mīmāṃsā-sūtra* traditions of Bādari's tradition were revived by Prabhākara, a 7th–8th-century scholar, and Jaimini's defended by Śabara and Kumārila, a 7th–8th-century scholar, Bādarāyaṇa's *sūtras* laid the basis for the development of Vedānta philosophy. The relation of the *Vedānta-sūtras* to the *Mīmāṃsā-sūtras*, however, is difficult to ascertain. Bādarāyaṇa approves of the Mīmāṃsā view that the relation between words and their significations is eternal. There are, however, clear statements of difference: according to Jaimini, for example, the dispenser of the "fruits" of one's actions is *dharma,* the law of righteousness itself, but for Bādarāyaṇa it is the supreme lord, Īśvara. Often, Jaimini's interpretation is contrasted with that of Bādari; in such cases, Bādarāyaṇa sometimes supports Bādari's view and sometimes regards both as defensible.

Hermeneutical principles of Bādarāyaṇa

The overall difference that emerges is that whereas Jaimini lays stress on the ritualistic parts of the Vedas, Bādarāyaṇa lays stress on the philosophical portions—*i.e.,* the *Upaniṣads*. The former recommends the path of Vedic injunctions, hence the ideal of *karma;* the latter recommends the path of knowledge. The central concept of Jaimini's investigation is *dharma—i.e.,* what ought to be done; the central theme of Bādarāyaṇa's investigations is Brahman—*i.e.,* the absolute reality. The relationship between these two treatises remains a matter of controversy between later commentators—Rāmānuja, a great South Indian philosopher of the 11th–12th centuries, defending the thesis that they jointly constitute a single work with Jaimini's coming first and Bādarāyaṇa's coming after it in logical order, and Śaṅkara, an earlier great South Indian philosopher of the 8th–9th centuries, in favour of the view that the two are independent of each other and possibly also inconsistent in their central theses.

Contents and organization of the four books. Bādarāyaṇa's *sūtras* have four books (*adhyāyas*), each book having four chapters (*pādas*). The first book is concerned with the theme of *samanvaya* ("reconciliation"). The many conflicting statements of the scriptures are all said to

agree in converging on one central theme: the concept of Brahman, the one absolute being from whom all beings arise, in whom they are maintained, and into whom they return. The second book establishes *avirodha* ("consistency") by showing the following: (1) that dualism and Vaiśeṣika atomism are neither sustainable interpretations of the scriptures nor defensible rationally; (2) that though consciousness cannot conceivably arise out of a nonconscious nature, the material world could arise out of spirit; (3) that the effect in its essence is not different from the cause; and (4) that though Brahman is all-perfect and has no want, creation is an entirely unmotivated free act of delight (*līlā*). The Buddhist (Vijñānavāda) view that there are no external objects but only minds and their conceptions is refuted, as also the Buddhist doctrine of the momentariness of all that is. The Jaina pluralism and the theism of the Pāśupatas and the Bhāgavatas are also rejected. Because, according to Vedānta, only Brahman is external, the third and the fourth chapters of the second book undertake to show that nothing else is eternal. The third book concerns the spiritual discipline and the various stages by which the finite individual (*jīva*) may realize his essential identity with Brahman. The fourth and last book deals with the final result of the modes of discipline outlined in the preceding book and distinguishes between the results achieved by worshipping a personal Godhead and those achieved by knowing the one Brahman. Included is some discussion of the possible "worlds" through which the spirits travel after death, but all this discussion is subordinate to the one dominant goal of liberation and consequent escape from the chain of rebirth.

Variations in views. Bādarāyaṇa's *sūtra*s refer to interpreters of Vedānta before him who were concerned with such central issues as the relation between the finite individual (*jīva*) and the absolute spirit (Brahman) and the possible bodily existence of a liberated individual. To Āśmarthya, an early Vedānta interpreter, is ascribed the view that the finite individual and the absolute are both identical and different (as causes and their effects are different—a view that seems to have been the ancestor of the later theory of Bhedābheda). Auḍulomi, another pre-Bādarāyaṇa Vedānta philosopher, is said to have held the view that the finite individual becomes identical with Brahman after going through a process of purification. Another interpreter, Kāśakṛtsna, holds that the two are identical—a view that anticipates the later "unqualified monism" of Śaṅkara. Bādarāyaṇa's own views on this issue are difficult to ascertain: the *sūtras* are so concise that they are capable of various interpretations, though there are reasons to believe that Rāmānuja's is closer to their intentions than Śaṅkara's.

The relation between the finite individual and the absolute spirit

The "Sāṃkhya-kārikās." *Relation to orthodoxy.* Īśvarakṛṣṇa's *Sāṃkhya-kārikā* (or "Verses on Sāṃkhya," *c.* 2nd century AD) is the oldest available Sāṃkhya work. Īśvarakṛṣṇa describes himself as laying down the essential teachings of Kapila as taught to Āsuri and by Āsuri to Pañcaśikha. He refers also to *Ṣaṣṭitantra* ("Doctrine of 60 Conceptions"), the main doctrines of which he claims to have expounded in the *kārikā*s. The Sāṃkhya of Caraka, which is substantially the same as is attributed to Pañcaśikha in the *Mahābhārata,* is theistic and regards the unmanifested (*avyakta*) as being the same as the *puruṣa* (the self). The *Mahābhārata* refers to three kinds of Sāṃkhya doctrines: those that accept 24, 25, or 26 principles, the last of which are theistic. The later *Sāṃkhya-sūtra* is more sympathetic toward theism, but the *kārikā*s are atheistic, and the traditional expositions of the Sāṃkhya are based on this work.

The nature of the self (*puruṣa*). According to the *kārikā*s, there are many selves, each being of the nature of pure consciousness. The self is neither the original matter (*prakṛti*) nor an evolute of it. Though matter is composed of the three *guṇa*s (qualities), the self is not; though matter, being nonintelligent, cannot discriminate, the self is discriminating; though matter is object (*viṣaya*), the self is not; though matter is common, the self is an individual (*asāmānya*); unlike matter, the self is not creative (*aprasavadharmin*). The existence of selves is proved on the ground that nature exhibits an ordered arrangement the like of which is known to be meant for another (*parārthatva*). This other must be a conscious spirit. That there are many such selves is proved on the grounds that different persons are born and die at different times, that they do not always act simultaneously, and that they show different qualities, aptitudes, and propensities. All selves are, however, passive witnesses (*sākṣin*), essentially alone (*kevala*), neutral (*madhyastha*), and not agents (*akartā*).

The nature, origin, and structure of the world (prakṛti). Phenomenal nature, with its distinctions of things and persons (taken as psychophysical organisms), is regarded as an evolution out of a primitive state of matter. This conception is based on a theory of causality known as the *satkāryavāda,* according to which an effect is implicitly pre-existent in its cause prior to its production. This latter doctrine is established on the ground that if the effect were not already existent in its cause, then something would have to come out of nothing. The original *prakṛti* (primeval stuff) is the primary matrix out of which all differentiations arose and within which they all were contained in an undistinguished manner. Original Matter is uncaused, eternal, all-pervading, one, independent, self-complete, and has no distinguishable parts; the things that emerge out of this primitive matrix are, on the other hand, caused, noneternal, limited, many, dependent, wholes composed of parts, and manifested. But Matter, whether in its original unmanifested state or in its manifested forms, is composed of three *guṇa*s, nondiscriminating (*avivekin*), object (*viṣaya*), general, nonconscious, and yet creative.

The evolution of original matter

The order in which Matter evolves is laid down as follows: *prakṛti* → *mahat* or *buddhi* (Intelligence) → *ahaṃkāra* (ego-sense) → *manas* (mind) → five *tanmātras* (the sense data: colour, sound, smell, touch, and taste) → five sense organs → five organs of action (tongue, hands, feet, organs of evacuation and of reproduction) → five gross elements (ether, air, light, water, and earth). This emanation schema may be understood either as an account of cosmic evolution or as a logical–transcendental analysis of the various factors involved in experience or as an analysis of the concrete human personality.

The concept of the three qualities (guṇas). A striking feature of this account is the conception of *guṇa:* nature is said to consist of three *guṇa*s—originally in a state of equilibrium and subsequently in varying states of mutual preponderance. The *kārikā*s do not say much about whether the *guṇa*s are to be regarded as qualities or as component elements. Of the three, harmony or tension (*sattva*) is light (*laghu*), is pleasing, and is capable of manifesting others. Activity (*rajas*) is dynamic, exciting, and capable of hurting. Inertia (*tamas*) is characterized by heaviness, conceals, is static, and causes sadness. Man's varying psychological responses are thus hypostatized and made into component properties or elements of nature—an argument whose fallacy was exposed, among others, by Śaṅkara.

Epistemology. The *Sāṃkhya-kārikā* delineates three ways of knowing (*pramāṇa*): perception, inference, and verbal testimony. Perception is defined as the application of the sense organs to their respective objects (*prativiṣayādhyavasāya*). Inference, which is not defined, is divided first into three kinds, and then into two. According to the former classification, an inference is called *pūrvavat* if it is based on past experience (such as when one, on seeing a dark cloud, infers that it will rain); it is called *śeṣavat* when from the presence of a certain property in one part of a thing the presence of the same property is inferred in the rest (such as when, on finding a drop of sea water to be saline, one infers the rest to be so); it is called *sāmānyato-dṛṣṭa* when it is used to infer what is not perceivable (such as when one infers the movement of a star on seeing it occupy two different positions in the firmament at different times). According to the other classification, an inference may be either from the mark to that of which it is the mark or in the reverse direction. Verbal testimony, in order to be valid, must be the word of one who has authoritative knowledge.

Perception and inference

There is, in addition to the three ways of knowing, consideration of the modes of functioning of the sense organs. The outer senses apprehend only the present objects, the

inner senses (*manas, antaḥkaraṇa,* and *buddhi*) have the ability to apprehend all objects—past, present, and future. The sense organs, on apprehending their objects, are said to offer them to *buddhi,* or intelligence, which both makes judgments and enjoys the objects of the senses. *Buddhi* is also credited with the ability to perceive the distinction between the self and the natural components of the person.

Ethics. In its ethics, the *kārikās* manifest an intellectualism that is characteristic of the Sāṃkhya system. Suffering is due to ignorance of the true nature of the self, and freedom, the highest good, can be reached through knowledge of the distinction between the self and nature. In this state of freedom, the self becomes indifferent to nature; it ceases to be an agent and an enjoyer. It becomes what it in fact is, a pure witness consciousness.

The "Yoga-sūtras." *Relation to Sāṃkhya.* The *Yoga-sūtras* of Patañjali (2nd century BC) are the earliest extant textbook on Yoga. Scholars now generally agree that the author of the *Yoga-sūtras* is not the grammarian Patañjali. In any case, the *Yoga-sūtras* stand in close relation to the Sāṃkhya system, so much so that tradition regards the two systems as one. Yoga adds a 26th principle to the Sāṃkhya list of 25—*i.e.,* the supreme lord, or Īśvara—and has thus earned the name of Seśvara-Sāṃkhya, or theistic Sāṃkhya. Furthermore, there is a difference in their attitudes: Sāṃkhya is intellectualistic and emphasizes metaphysical knowledge as the means to liberation; Yoga is voluntaristic and emphasizes the need of going through severe self-control as the means of realizing intuitively the same principles.

God, self, and body. In the *Yoga-sūtras,* God is defined as a distinct self (*puruṣa*), untouched by sufferings, actions, and their effects; his existence is proved on the ground that the degrees of knowledge found in finite beings, in an ascending order, has an upper limit—*i.e.,* omniscience, which is what characterizes God. He is said to be the source of all secular and scriptural traditions; he both revealed the Vedas and taught the first fathers of mankind. Surrender of the effects of action to God is regarded as a recommended observance.

Distinction between the self and the mind

As in Sāṃkhya, the self is distinguished from the mind (*citta*): the mind is viewed as an object, an aggregate. This argument is used to prove the existence of a self other than the mind. The mental state is not self-intimating; it is known in introspection. It cannot know both itself and its object. It rather is known by the self, whose essence is pure, undefiled consciousness. That the self is not changeable is proved by the fact that were it changeable the mental states would be sometimes known and sometimes unknown—which, however, is not the case, because a mental state is always known. To say that the self knows means that the self is reflected in the mental state and makes the latter manifested. The aim of Yoga is to arrest mental modifications (*citta-vṛtti*) so that the self remains in its true, undefiled essence and is, thus, not subject to suffering.

The attitude of the *Yoga-sūtras* to the human body is ambivalent. The body is said to be filthy and unclean. Thus, the ascetic cultivates a disgust for it. Yet, much of the discipline laid down in the *Yoga-sūtras* concerns perfection of the body, with the intent to make it a fit instrument for spiritual perfection. Steadiness in bodily posture and control of the breathing process are accorded a high place. The perfection of body is said to consist in "beauty, grace, strength and adamantine hardness."

Theories and techniques of self-control and meditation. Patañjali lays down an eightfold path consisting of aids to Yoga: restraint (*yama*), observance (*niyama*), posture (*āsana*), regulation of breathing (*prāṇāyama*), abstraction of the senses (*pratyāhāra*), concentration (*dhāranā*), meditation (*dhyānā*), and trance (*samādhi*). The first two constitute the ethical core of the discipline: the restraints are abstinence from injury, veracity, abstinence from stealing, continence, and abstinence from greed. The observances are cleanliness, contentment, purificatory actions, study, and surrender of the fruits of one's actions to God. *Ahiṃsā* (nonviolence) also is glorified, as an ethics of detachment.

Various stages of *samādhi* are distinguished: the conscious and the superconscious, which are subdivided into achievements with different shades of perfection. In the final stage, all mental modifications cease to be and the self is left in its pure, undefiled state of utter isolation. This is freedom (*kaivalya*), or absolute independence.

The "Vaiśeṣika-sūtras." The *Vaiśeṣika-sūtras* were written by Kaṇāda, a philosopher who flourished *c.* 2nd–4th centuries. The system owes its name to the fact that it admits ultimate particularities (*viśeṣa*). The metaphysics is, therefore, pluralistic.

Organization and contents. The *Vaiśeṣika-sūtras* are divided into ten chapters, each with two sections. Chapter 1 states the purpose of the work: to explain *dharma,* defined as that which confers prosperity and ultimate good on man. This is followed by an enumeration of the categories of being recognized in the system: substance, quality (*guṇa*), action, universality, particularity, and inherence (*samavāya*). Later authors add a seventh category: negation (*abhāva*). This enumeration is followed by an account of the common features as well as dissimilarities among these categories: the categories of "universal" and "particularity" and the concepts of being and existence. Chapter 2 classifies substances into nine kinds: earth, water, fire, air, ether, space, time, self, and mind. There next follows a discussion of the question of whether sound is eternal or noneternal. Chapter 3 is an attempt to prove the existence of self by an inference. Chapter 4 explains the words "eternal" and "noneternal," the noneternal being identified with *avidyā,* and distinguishes between three different forms of the substances earth, water, fire, and air—each of these is either a body, a sense organ, or an object. Chapter 5 deals with the notion of action and the connected concept of effort, and the next traces various special phenomena of nature to the supersensible force, called *adṛṣṭa.* Chapter 6 argues that performance of Vedic injunctions generates this supersensible force and that the merits and demerits accumulated lead to *mokṣa.* Chapter 7 argues that qualities of eternal things are eternal and those of noneternal things are noneternal. Chapter 8 argues that the self and mind are not perceptible. Chapter 9 argues that neither action nor qualities may be ascribed to what is nonexistent and, further, that negation may be directly perceived. Chapter 9 also deals with the nature of *hetu,* or the "middle term" in syllogism, and argues that the knowledge derived from hearing words is not inferential. Chapter 10 argues that pleasure and pain are not cognitions because they do not leave room for either doubt or certainty.

The categories of being in the Vaiśeṣika system

Structure of the world. This account of the contents of the *sūtras* shows that the Vaiśeṣika advocates an atomistic cosmology (theory of order) and a pluralistic ontology (theory of being). The material universe arises out of the conjunction of four kinds of atoms: the earth atom, water atom, fire atom, and air atom. There also are the eternal substances: ether, in which sound inheres as a quality; space, which accounts for man's sense of direction and distinctions between far and near; and time, which accounts for the notions of simultaneity and nonsimultaneity and which, like space, is eternal and is the general cause of all that has origin.

Atomistic cosmology and pluralistic ontology

Naturalism. The overall naturalism of the Vaiśeṣika, its great interest in physics, and its atomism are all counterbalanced by the appeal to *adṛṣṭa* (a supersensible force), to account for whatever the other recognized entities cannot explain. Among things ascribed to this supersensible force are movements of needles toward a magnet, circulation of water in plant bodies, upward motion of fire, movement of mind, and movements of soul after death. These limit the naturalism of the system.

Epistemology. Knowledge belongs to the self; it appears or disappears with the contact of the self with the senses and of the senses with the objects. Perception of the self results from the conjunction of the self with the mind. Perception of objects results from proximity of the self, the senses, and the objects. Error exists because of defects of the senses. Inference is of three kinds: inference of the nonexistence of something from the existence of some other things, inference of the existence of something from nonexistence of some other, and inference of existence of something from the existence of some other thing.

Ethics. *Mokṣa* is a state in which there is no body and no rebirth. It is achieved by knowledge. Works in accordance with the Vedic injunction may help in its attainment.

The "Nyāya-sūtras." The *Nyāya-sūtras* probably were composed by Gautama or Akṣapāda about the 2nd century BC, though there is ample evidence that many *sūtras* were subsequently interpolated.

Content and organization. The *sūtras* are divided into five chapters, each with two sections. The work begins with a statement of the subject matter, purpose, and relation of the subject matter to the attainment of that purpose. The ultimate purpose is salvation—*i.e.,* complete freedom from pain—and salvation is attained by knowledge of the 16 categories: hence the concern with these categories, which are means of valid knowledge (*pramāṇa*); objects of valid knowledge (*prameya*); doubt (*saṃśaya*); purpose (*prayojana*); example (*dṛṣṭānta*); conclusion (*siddhānta*); the constituents of a syllogism (*avayava*); argumentation (*tarka*); ascertainment (*nirṇaya*); debate (*vāda*); disputations (*jalpa*); destructive criticism (*vitaṇḍā*); fallacy (*hetvābhāsa*); quibble (*chala*); refutations (*jāti*); and points of the opponent's defeat (*nigrahasthāna*).

The categories of the *Nyāya* system

Epistemology. The words "knowledge," *buddhi,* and "consciousness" are used synonymously. Four means of valid knowledge are admitted: perception, inference, comparison, and verbal testimony. Perception is defined as the knowledge that arises from the contact of the senses with the object, which is nonjudgmental, or unerring or judgmental. Inference is defined as the knowledge that is preceded by perception (of the mark) and classified into three kinds: that from the perception of a cause to its effect; that from perception of the effect to its cause; and that in which knowledge of one thing is derived from the perception of another with which it is commonly seen together. Comparison is defined as the knowledge of a thing through its similarity to another thing previously well-known.

The validity of the means of knowing is established as against Buddhist skepticism, the main argument being that if no means of knowledge is valid then the demonstration of their invalidity cannot itself claim validity. Perception is shown to be irreducible to inference, inference is shown to yield certain knowledge, and errors in inference are viewed as being faults in the person, not in the method itself. Knowledge derived from verbal testimony is viewed as noninferential.

Theory of causation and metaphysics. Although the *sūtras* do not explicitly develop a detailed theory of causation, the later Nyāya theory is sufficiently delineated in Chapter 4. No event is uncaused. No positive entity could arise out of mere absence—a thesis that is pressed against what seems to be a Buddhist view that in a series of momentary events every member is caused by the destruction of the preceding member. Cause and effect should be homogeneous in nature, and yet the effect is a new beginning and was not already contained in the cause. The Buddhist thesis that all things are negative in nature (inasmuch as a thing's nature is constituted by its differences from others) is rejected, as is the view that all things are eternal or that all things are noneternal. Both these latter views are untrue to experience. Thus, the resulting metaphysics admits two kinds of entities: eternal and noneternal. The whole is a new entity over and above the parts that constitute it. Also, the idea that God is the material cause of the universe is rejected. God is viewed as the efficient cause, and human deeds produce their results under the control and cooperation of God.

The syllogism and its predecessors. Of the four main topics of the *Nyāya-sūtras* (art of debate, means of valid knowledge, syllogism, and examination of opposed views) there is a long history. There is no direct evidence for the theory that though inference (*anumāna*) is of Indian origin, the syllogism (*avayava*) is of Greek origin. Vātsāyana, the commentator on the *sūtras,* referred to some logicians who held a theory of a ten-membered syllogism (the Greeks had three). The *Vaiśeṣika-sūtras* give five propositions as constituting a syllogism but give them different names. Gautama also supports a five-membered syllogism with the following structure:

The propositions of the *Nyāya* syllogism

1. This hill is fiery (*pratijñā:* a statement of that which is to be proved).
2. Because it is smoky (*hetu:* statement of reason).
3. Whatever is smoky is fiery, as is a kitchen (*udāharaṇa:* statement of a general rule supported by an example).
4. So is this hill (*upanaya:* application of the rule of this case).
5. Therefore this hill is fiery (*nigamana:* drawing the conclusion).

The most characteristic feature of the Nyāya syllogism is its insistence on the example—which seems to suggest that the Nyāya logician wanted to be assured not only of formal validity but also of material truth. Five kinds of fallacious "middle" (*hetu*) are distinguished: the inconclusive (*savyabhicāra*), which leads to more conclusions than one; the contradictory (*viruddha*), which opposes that which is to be established; the controversial (*prakaraṇasama*), which provokes the very question that it is meant to settle; the counterquestioned (*sādhyasama*), which itself is unproved; and the mistimed (*kālātīta*), which is adduced "when the time in which it might hold good does not apply."

Other characteristic philosophic matters. Other philosophical theses stated in the *sūtras* are as follows: the relation of words to their meanings is not natural but conventional; a word means neither the bare individual nor the universal by itself but all three—the individual, the universal, and structure (*ākṛti*); desire, aversion, volition, pleasure, pain, and cognition are the marks of the self; body is defined as the locus of gestures, senses, and sentiments; and the existence and atomicity of mind are inferred from the fact that there do not arise in the self more acts of knowledge than one at a time.

The beginnings of Mahāyāna Buddhist philosophy. *Contributions of the Mahāsaṅgikas.* When the Mahāsaṅgikas ("School of the Great Assembly") seceded from the Elders (Theravādins) about 400 BC, the germs were laid for the rise of the Mahāyāna Buddhism. The Mahāsaṅgikas admitted non-*arhat* monks and worshippers (*i.e.,* those who had not attained perfection), defied the Buddha, taught the doctrine of the emptiness of the elements of being, distinguished between the mundane and the supramundane reality, and considered consciousness (*vijñāna*) to be intrinsically free from all impurities. These ideas found varied expression among the various groups into which the Mahāsaṅgikas later divided (see also BUDDHISM).

Contributions of the Sarvāstivādins. The Sarvāstivādins ("realists" who believe that all things, mental and material, exist or also that all *dharmas*—past, present, and future—exist) seceded from the Elders about the middle of the 3rd century BC. They rejected, in common with all other sects, *pudgalātmā,* or a self of the individual, but admitted *dharmātman*—*i.e.,* self-existence of the *dharmas* (categories), or the elements of being. Each *dharma* is a self-being; the law of causality applies to the formation of aggregates, not to the elements themselves. *Dharmas,* whether they are past or are in future, exist all the same. Of these, three are said to be unconditioned: space (*ākāśa*) and the two cessations (*nirodha*)—the cessation that arises from knowledge and the cessation that arises prior to the attainment of knowledge, the former being Nirvāṇa, the latter being an arrest of the flow of passions through meditation prior to the achievement of Nirvāṇa. By *śūnyatā* the Sarvāstivādins mean only the truth that there is no eternal substance called "I." Because all elements—past, present, or future—exist, the Sarvāstivādins are obliged to account for these temporal predicates, and several different theories are advanced. Of these, the theory advanced by Vasumitra, a 1st–2nd-century-AD Sarvāstivādin, viz., that temporal predicates are determined by the function of a *dharma,* is accepted by the Vaibhāṣikas—*i.e.,* those among the Sarvāstivādins who follow the authority of the texts known as the *Vibhāṣā.*

The concept of the existence of all things and all *dharmas*

Contributions of the Sautrāntikas. The Vaibhāṣika doctrine of eternal elements is believed to be inconsistent with the fundamental teachings of the Buddha. The Sautrāntikas (so-called because they rest their case on the *sūtras*) insist on the noneternality of the *dharma* as well. The past

and the future *dharmas* do not exist, only the present ones do. The so-called unconditioned *dharmas* are mere absences, not positive entities. Thus, the Sautrāntikas seem to be the only major school of Buddhist philosophy that comes near to regarding Nirvāṇa as entirely negative. In epistemology, whereas the Vaibhāṣikas are direct realists, the Sautrāntikas hold a sort of representationism, according to which the external world is only inferred from the mental conceptions that alone are directly apprehended.

The worldview of the "Arthaśāstra." Kauṭilya's *Arthaśāstra* (*c.* 321–296 BC) is the science of *artha,* or material prosperity, which is one of the four goals of human life. By *artha,* Kauṭilya meant "the means of subsistence of man," which is, primarily, wealth and, secondarily, earth. The work is concerned with the means of fruitfully maintaining and using the latter—*i.e.,* land. It is a work on politics and diplomacy.

Political philosophical ideas

Theories of kingship and statecraft. Though Kauṭilya recognized that sovereignty may belong to a clan (*kula*), he was himself concerned with monarchies. He advocated the idea of the king's divine nature, or divine sanction of the king's office, but he also attempted to reconcile it with a theory of the elective origin of the king. He referred to a state of nature, without king, as an anarchy in which the stronger devours the weaker. The four functions of the king are to acquire what is not gained, to protect what is gained, to increase what is protected, and to bestow the surplus upon the deserving. The political organization is held to have seven elements: the king, the minister, the territory, the fort, the treasury, the army, and the ally. These are viewed as being organically related. The three "powers" of the king are power of good counsel, the majesty of the king himself, and the power to inspire. The priest is not made an element of the state organization. The king, however, is not exempt from the laws of *dharma.* Being the "promulgator of *dharma,*" the king should himself be free from the six passions of sex, anger, greed, vanity, haughtiness, and overjoy. What Kauṭilya advocated was an enlightened monarchical paternalism.

Concepts of the public good. In the happiness of the subjects lies the king's happiness. The main task of the king is to offer protection. Monarchy is viewed as the only guarantee against anarchy. Thus, the king's duty is to avert providential visitations such as famine, flood, and pestilence; he ought also to protect agriculture, industry, and mining, the orphan, the aged, the sick, and the poor, to control crime with the help of spies, and to settle legal disputes.

Relations between states. Regarding relations with other states, Kauṭilya's thoughts were based not so much on high moral idealism as on the needs of self-interest. He wrote of six types of foreign policy: treaty (*sandhi*), war (*vigraha*), marching against the enemy (*yāna*), neutrality (*āsana*), seeking protection from a powerful king (*saṃśraya*), and dual policy (*dvaidhībhāva*). The rules concerning these are: he who is losing strength in comparison to the other shall make peace; he who is gaining strength shall make war; he who thinks neither he nor the enemy can win shall be neutral; he who has an excess of advantage shall march; he who is wanting in strength shall seek protection; he who undertakes work requiring assistance shall adopt a dual policy.

The formation and implementation of policy. Kauṭilya's views about the formation and implementation of policy were as follows: a treaty based on truth and oath is binding for temporal and spiritual consequences; a treaty based on security is binding only as long as the party is strong. He who inflicts severe punishments becomes oppressive; he who inflicts mild punishments is overpowered; and he who inflicts just punishments is respected. Kauṭilya advocated an elaborate system of espionage for domestic as well as foreign affairs.

Fragments from the Ājīvikas and the Cārvākas. *The Ājīvikas.* About the time of the rise of Buddhism, there was a sect of religious mendicants, the Ājīvikas, who held unorthodox views. In the strict sense, this name is applied to the followers of one Makkhali Gosāla, but in a wide sense it is also applied to those who taught many different shades of heretical teachings. Primary sources of knowledge about these are the *Dīgha Nikāya, Aṅguttara Nikāya, Saṃyutta Nikāya,* the *Sūtrakṛtanga-sūtra,* Śīlāṅka's commentary on the *Sutrakṛtanga-sūtra,* the *Bhagavatī-sūtra,* the *Nandī-sūtra,* and Abhayadeva's commentary on *Samavayaṅga-sūtra.*

Makkhali's views may be thus summarized. There is no cause of the depravity of things; they become depraved without any reason or cause. There is also no cause of the purity of beings; they become pure without any reason or cause. Nothing depends either on one's own efforts or on the efforts of others. All things are destitute of power, force, or energy. Their changing states are due to destiny, environment, and their own nature. Thus, Makkhali denies sin, or *dharma,* and denies freedom of man in shaping his own future. He is thus a determinist, although scholars have held the view that he might leave room for chance, if not for freedom of will. He is supposed to have held an atomistic cosmology and that all beings, in the course of time, are destined to culminate in a state of final salvation. He believes not only in rebirth but also in a special doctrine of reanimation according to which it is possible for one person's soul to be reanimated in the dead bodies of others. Thus, the Ājīvikas are far from being materialists.

Materialistic views of the Cārvākas

The Cārvākas. Another pre-Buddhistic system of philosophy, the Cārvāka, or the Lokāyata, is one of the earliest materialistic schools of philosophy. The name Cārvāka is traced back to one Cārvāka, supposed to have been one of the great teachers of the school. The other name, Lokāyata, means "the view held by the common people," "the system which has its base in the common, profane world," "the art of sophistry," and also "the philosophy that denies that there is any world other than this one." Bṛhaspati probably was the founder of this school. Much knowledge of the Cārvākas, however, is derived from the expositions of the later Hindu writings, particularly from Mādhava's *Sarva-darśana-saṃgraha* ("Compendium of All Philosophies," 14th century). Haribhadra in his *Ṣaḍdarśanasamuccaya* ("Compendium of the Six Philosophies," 5th century AD) attributes to the Cārvākas the view that this world extends only to the limits of possible sense experience.

The Cārvākas apparently sought to establish their materialism on an epistemological basis. In their epistemology, they viewed sense perception alone as a means of valid knowledge. The validity of inferential knowledge was challenged on the ground that all inference requires a universal major premise ("All that possesses smoke possesses fire") whereas there is no means of arriving at a certainty about such a proposition. No amount of finite observations could possibly yield the required universal premise. The supposed "invariable connection" may be vitiated by some unknown "condition," and there is no means of knowing that such a vitiating factor does not exist. Since inference is not a means of valid knowledge, all such supersensible objects as "afterlife," "destiny," or "soul" do not exist. To say that such entities exist though there is no means of knowing them is regarded as absurd, for no unverifiable assertion of existence is meaningful.

The authority of the scriptures also is denied. First, knowledge based on verbal testimony is inferential and therefore vitiated by all the defects of inference. The Cārvākas regard the scriptures as characterized by the three faults: falsity, self-contradiction, and tautology. On the basis of such a theory of knowledge, the Cārvākas defended a complete reductive materialism according to which the four elements of earth, water, fire, and air are the only original components of being and all other forms are products of their composition. Consciousness thus is viewed as a product of the material structure of the body and characterizes the body itself—rather than a soul—and perishes with the body. In their ethics, the Cārvākas upheld a hedonistic theory according to which enjoyment of the maximum amount of sensual pleasure here in this life and avoidance of pain that is likely to accompany such enjoyment are the only two goals that men ought to pursue.

FURTHER DEVELOPMENTS OF THE SYSTEM

Developments in Mahāyāna. *Nāgārjuna and Śūnyavāda.* Though the beginnings of Mahāyāna are to be found

in the Mahāsaṅgikas and many of their early sects, Nāgārjuna gave it a philosophical basis. Not only is the individual person empty and lacking an eternal self, according to Nāgārjuna, but the *dharma*s also are empty. He extended the concept of *śūnyatā* to cover all concepts and all entities. "Emptiness" thus means subjection to the law of causality or "dependent origination" and lack of an immutable essence and an invariant mark (*niḥsvabhāvatā*). It also entails a repudiation of dualities between the conditioned and the unconditioned, between subject and object, relative and absolute, and between *saṃsāra* and Nirvāṇa. Thus, Nāgārjuna arrived at an ontological monism; but he carried through an epistemological dualism (*i.e.,* a theory of knowledge based on two sets of criteria) between two orders of truth: the conventional (*samvṛtti*) and the transcendental (*paramārtha*). The one reality is ineffable. Nāgārjuna undertook a critical examination of all the major categories with which philosophers had sought to understand reality and showed them all to involve self-contradictions. The world is viewed as a network of relations, but relations are unintelligible. If two terms, A and B, are related by the relation R, then either A and B are different or they are identical. If they are identical, they cannot be related; if they are altogether different then they cannot also be related, for they would have no common ground. The notion of "partial identity and partial difference" is also rejected as unintelligible. The notion of causality is rejected on the basis of similar reasonings. The concepts of change, substance, self, knowledge, and universals do not fare any better. Nāgārjuna also directed criticism against the concept of *pramāṇa,* or the means of valid knowledge.

Ontological monism and epistemological dualism

Nāgārjuna's philosophy is also called Mādhyamika, because it claims to tread the middle path, which consists not in synthesizing opposed views such as "The real is permanent" and "The real is changing" but in showing the hollowness of both the claims. To say that reality is both permanent and changing is to make another metaphysical assertion, another viewpoint, whose opposite is "Reality is neither permanent nor changing." In relation to the former, the latter is a higher truth, but the latter is still a point of view, a *dṛṣṭi,* expressed in a metaphysical statement, though Nāgārjuna condemned all metaphysical statements as false.

Nāgārjuna used reason to condemn reason. Those of his disciples who continued to limit the use of logic to this negative and indirect method, known as *prasaṅga,* are called the *prāsaṅgikas:* of these, Āryadeva, Buddhapālita, and Candrakīrti are the most important. Bhāvaviveka, however, followed the method of direct reasoning and thus founded what is called the *svatantra* (independent) school of Mādhyamika philosophy. With him Buddhist logic comes to its own, and during his time the Yogācāras split away from the Śūnyavādins.

Contributions of Vasubandhu and Asaṅga. Converted by his brother Asaṅga to the Yogācāra, Vasubandhu wrote the *Vijñapti-mātratā-siddhi* ("Establishment of the Thesis of Cognitions—Only"), in which he defended the thesis that the supposedly external objects are merely mental conceptions. Yogācāra idealism is a logical development of Sautrāntika representationism: the conception of a merely inferred external world is not satisfying. If consciousness is self-intimating (*svaprakāśa*) and if consciousness can assume forms (*sākāravijñāna*), it seems more logical to hold that the forms ascribed to alleged external objects are really forms of consciousness. One only needs another conception: a beginningless power that would account for this tendency of consciousness to take up forms and to externalize them. This is the power of *kalpanā,* or imagination. Yogācāra added two other modes of consciousness to the traditional six: ego consciousness (*manovijñāna*) and storehouse consciousness (*ālaya-vijñāna*). The *ālaya-vijñāna* contains stored traces of past experiences, both pure and defiled seeds. Early anticipations of the notions of the subconscious or the unconscious, they are theoretical constructs to account for the order of individual experience. It still remained, however, to account for a common world—which in fact remains the main difficulty of Yogācāra. The state of Nirvāṇa becomes a state in which the *ālaya* with its stored "seeds" would wither away (*ālayaparāvṛtti*). Though the individual ideas are in the last resort mere imaginations, in its essential nature consciousness is without distinctions of subject and object. This ineffable consciousness is the "suchness" (*tathatā*) underlying all things. Neither the *ālaya* nor the *tathatā,* however, is to be construed as being substantial.

Ego consciousness and stored consciousness

Vasubandhu and Asaṅga are also responsible for the growth of Buddhist logic. Vasubandhu defined "perception" as the knowledge that is caused by the object, but this was rejected by Dignāga, a 5th-century logician, as a definition belonging to his earlier realistic phase. Vasubandhu defined "inference" as a knowledge of an object through its mark, but Dharmottara, an 8th-century commentator pointed out that this is not a definition of the essence of inference but only of its origin.

Contributions of Dignāga and Dharmakīrti. Dignāga's *Pramāṇasamuccaya* ("Compendium of the Means of True Knowledge") is one of the greatest works on Buddhist logic. Dignāga gave a new definition of "perception": a knowledge that is free from all conceptual constructions, including name and class concepts. In effect, he regarded only the pure sensation as perception. In his theory of inference, he distinguished between inference for oneself and inference for the other and laid down three criteria of a valid middle term (*hetu*), viz., that it should "cover" the minor premise (*pakṣa*), be present in the similar instances (*sapakṣa*), and be absent in dissimilar instances (*vipakṣa*). In his *Hetucakra* ("The Wheel of 'Reason'"), Dignāga set up a matrix of nine types of middle terms, of which two yield valid conclusions, two contradictory, and the rest uncertain conclusions. Dignāga's tradition is further developed in the 7th century by Dharmakīrti, who modified his definition of perception to include the condition "unerring" and distinguished, in his *Nyāyabindu,* between four kinds of perception: that by the five senses, that by the mind, self-consciousness, and perception of the *yogin*s. He also introduced a threefold distinction of valid middle terms: the middle must be related to the major either by identity ("This is a tree, because this is an oak") or as cause and effect ("This is fiery, because it is smoky"), or the *hetu* is a nonperception from which the absence of the major could be inferred. Dharmakīrti consolidated the central epistemological thesis of the Buddhists that perception and inference have their own exclusive objects. The object of the former is the pure particular (*svalakṣaṇa*), and the object of the latter (he regarded judgments as containing elements of inference) is the universal (*sāmānyalakṣaṇa*). In their metaphysical positions, Dignāga and Dharmakīrti represent a moderate form of idealism.

Pure sensation as perception

Pūrva-mīmāṃsā: the Bhāṭṭa and Prābhākara schools. *Principal texts and relation to Śabara.* Kumārila commented on Jaimini's *sūtra* as well as on Śabara's *bhāṣya.* The *Vārttika* (critical gloss) that he wrote was commented upon by Sucarita Miśra in his *Kāśikā* ("The Shining"), by Someśvara Bhaṭṭa in his *Nyāyasudhā* ("The Nectar of Logic"), and Pārthasārathi Miśra in *Nyāyaratnākara* ("The Abode of Jewels of Logic"). Pārthasārathi's *Śāstradīpikā* ("Light on the Scripture") is a famous independent Mīmāṃsā treatise belonging to Kumārila's school.

Prabhākara, who most likely lived after Kumārila, was the author of the commentary *Bṛhatī* ("The Large Commentary"), on Śabara's *bhāṣya.* On many essential matters, Prabhākara differs radically from the views of Kumārila. Prabhākara's *Bṛhatī* has been commented upon by Śālikanātha in his *Ṛjuvimalā* ("The Straight and Free from Blemishes"), whereas the same author's *Prakaraṇapañcikā* ("Commentary of Five Topics") is a very useful exposition of the Prābhākara system. Other works belonging to this school are Mādhava's *Jaiminīyā-nyāyamālā-vistara* ("Expansion of the String of Reasonings by Jaimini"). Appaya Dīkṣita's *Vidhirasāyana* ("The Elixir of Duty"), Āpadeva's *Mīmāṃsā-nyāyaprakāśa* (Illumination of the Reasonings of *Mīmāṃsā*) and Laugākṣi Bhāskara's *Artha-saṃgraha* ("Collection of Treasures").

Where Kumārila and Prabhākara differed, Kumārila remained closer to both Jaimini and Śabara. Kumārila,

like Jaimini and Śabara, restricted Mīmāṃsā to an investigation into *dharma,* whereas Prabhākara assigned to it the wider task of enquiring into the meaning of the Vedic texts. Kumārila understood the Vedic injunction to include a statement of the results to be attained; Prabhākara—following Bādari—excluded all consideration of the result from the injunction itself and suggested that the sense of duty alone should instigate a person to act.

Metaphysics and epistemology. Both the Bhāṭṭa (the name for Kumārila's school) and the Prabhākara schools, in their metaphysics, were realists; both undertook to refute Buddhist idealism and nihilism. The Bhāṭṭa ontology recognized five types of entities: substance (*dravya*), quality (*guṇa*), action (*karma*), universals (*sāmānya*), and negation (*abhāva*). Of these, substance was held to be of ten kinds: the nine substances recognized by the Vaiśeṣikas and the additional substance "darkness." The Prābhākara ontology recognized eight types of entities; from the Bhāṭṭa list, negation was rejected, and four more were added: power (*śakti*), resemblance (*sadṛsa*), inherence-relation (*samavāya*), and number (*sāṃkhyā*). Under the type "substance," the claim of "darkness" was rejected on the ground that it is nothing but absence of perception of colour; the resulting list of nine substances is the same as that of the Vaiśeṣikas. Though both the schools admitted the reality of the universals, their views on this point differed considerably. The Prābhākaras admitted only such universals as inhere in perceptible instances and insisted that true universals themselves must be perceivable. Thus, they rejected abstract universals, such as "existence," and merely postulated universals, such as "Brahminhood" (which cannot be perceptually recognized in a person).

Refutation of Buddhist idealism and nihilism

The epistemologies of the two schools differ as much as their ontologies. As ways of valid knowing, the Bhāṭṭas recognized perception, inference, verbal testimony (*śabda*), comparison (*upamāna*), postulation (*arthāpatti*), and nonperception (*anupalabdhi*). The last is regarded as the way men validly, and directly, apprehend an absence: this was in conformity with Śabara's statement that *abhāva* (nonexistence) itself is a *pramāṇa* (way of true knowledge). Postulation is viewed as the sort of process by which one may come to know for certain the truth of a certain proposition, and yet the Bhāṭṭas refused to include such cases under inference on the grounds that in such cases one does not say to himself "I am inferring" but rather says "I am postulating." "Comparison" is the name given to the perception of resemblance with a perceived thing of another thing that is not present at that moment. It is supposed that because the latter thing is not itself being perceived, the resemblance belonging to it could not have been perceived; thus, it is not a case of perception when one says "My cow at home is similar to this animal."

The Prābhākaras rejected nonperception as a way of knowing and were left with a list of five concerning definitions of perception. The Bhāṭṭas, following the *sūtra,* define perception in terms of sensory contact with the object, whereas the Prābhākaras define it in terms of immediacy of the apprehension.

Ethics. As pointed out earlier, Kumārila supported the thesis that all moral injunctions are meant to bring about a desired benefit and that knowledge of such benefit and of the efficacy of the recommended course of action to bring it about is necessary for instigating a person to act. Prābhākara defended the ethical theory of duty for its own sake, the sense of duty alone being the proper incentive. The Bhāṭṭas recognize *apūrva* (supersensible efficacy of actions to produce remote effects) as a supersensible link connecting the moral action performed in this life and the supersensible effect (such as going to heaven) to be realized afterward. Prābhākara understood by *apūrva* only the action that ought to be done.

Hermeneutics and semantics. In their principles of interpretation of the scriptures, and consequently in their theories of meaning (of words and of sentences), the two schools differ radically. Prābhākara defended the thesis that words primarily mean either some course of action (*kārya*) or things connected with action. Connected with this is the further Prābhākara thesis that the sentence forms the unit of meaningful discourse, that a word is never used by itself to express a single unrelated idea, and that a sentence signifies a relational complex that is not a mere juxtaposition of word meanings. Prābhākara's theory of language learning follows these contentions: the child learns the meanings of sentences by observing the elders issuing orders like "Bring the cow" and the juniors obeying them, and he learns the meaning of words subsequently by a close observation of the insertion (*āvāpa*) and extraction (*uddhāra*) of words in sentences and the resulting variations in the meaning of those sentences. From this semantic approach follows Prābhākara's principle of Vedic interpretation: all Vedic texts are to be interpreted as bearing on courses of action prescribed, and there are no merely descriptive statements in the scriptures. Furthermore, only the Vedic injunctions yield the authoritative verbal testimony that may be regarded as a unique way of knowing, whereas all other verbal knowledge is really inferential in character. In matters concerning what ought to be done, Prābhākara therefore regarded only the Vedas as authoritative.

Language and learning

Kumārila's theory is very different. In his view, words convey their own meanings, not relatedness to something else. He therefore was more willing to accommodate purely descriptive sentences as significant. Furthermore, he regarded sentence meaning as composed of separate word meanings held together in a relational structure; the word meaning formed, for him, the simplest unit of sense. Persons thus learn the meaning of words by seeing others talking as well as from advice of the elders.

Religious consequences. The Mīmāṃsā views the universe as being eternal and does not admit the need of tracing it back to a creator. It also does not admit the need of admitting a being who is to distribute moral rewards and inflict punishments—this function being taken over by the notion of *apūrva,* or supersensible power generated by each action. Theoretically not requiring a God, the system, however, posits a number of deities as entailed by various ritualistic procedures, with no ontological status assigned to the gods.

The linguistic philosophies: Bhartṛhari and Maṇḍana-Miśra. The linguistic philosophers considered here are the grammarians led by Bhartṛhari (7th century AD) and Maṇḍana-Miśra (8th century AD); the latter, reputed to be a disciple of Kumārila, held views widely different from the Mīmāṃsākas. The grammarians share with the Mīmāṃsākas their interest in the problems of language and meaning. But their own theories are so different that they cut at the roots of the Mīmāṃsā realism. The chief text of this school is Bhartṛhari's *Vākyapadīya.* Maṇḍana's chief works are *Brahma-siddhi* ("Establishment of Brahman"), *Sphoṭa-siddhi* ("Establishment of Word Essence"), and *Vidhiviveka* ("Inquiry into the Nature of Injunctions").

As his first principle, Bhartṛhari rejects one of the doctrines on which the realism of Mīmāṃsā and Nyāya had been built up—the view that there is a kind of perception that is nonconceptualized and that places persons in direct contract with things as they are. For Bhartṛhari this is not possible, for all knowledge is "penetrated" by words and "illuminated" by words. Thus, all knowledge is linguistic, and the distinctions of objects are traceable to distinctions among words. The metaphysical monism of word (*śabdādvaita*) is not far from this—*i.e.,* the view that the one word essence appears as this world of "names and forms" because of man's imaginative construction (*kalpanā*). Metaphysically, Bhartṛhari comes close both to Śaṅkara's Advaita and the Buddhist philosophers, such as Dharmakīrti. This metaphysical theory, however, makes use of another doctrine—the doctrine of *sphoṭa* ("that from which the meaning bursts forth"). Most Indian philosophical schools were concerned with the problem of what precisely is the bearer of the meaning of a word or a sentence. If the letters are evanescent and if, as one hears the sounds produced by the letters of a word, each sound is replaced by another, one never comes to perceive the word as a whole, and the question is how one grasps the meaning of the word. The same problem could be

All knowledge as linguistic

stated with regard to a sentence. The Mīmāṃsākas postulated an eternity of sounds and distinguished between the eternal sounds and sound complexes (words, sentences) from their manifestations. The grammarians, instead, distinguished between the word and sound and made the word itself the bearer of meaning. As bearer of meaning, the word is the *sphoṭa.*

Sounds have spatial and temporal relations; they are produced differently by different speakers. But the word as meaning bearer has to be regarded as having no size or temporal dimension. It is indivisible and eternal. Distinguished from the *sphoṭa* are the abstract sound pattern (*prākṛtadhvani*) and the utterances (*vikṛtadhvani*). Furthermore, Bhartṛhari held that the sentence is not a collection of words or an ordered series of them. A word is rather an abstraction from a sentence; thus, the sentence-*sphoṭa* is the primary unit of meaning. A word is also grasped as a unity by an instantaneous flash of insight (*pratibhā*). This theory of *sphoṭa,* which is itself a linguistic theory required by the problems arising from the theory of meaning, was employed by the grammarians to support their theory of word monism.

Maṇḍana-Miśra, in his *Vidhiviveka,* referred to three varieties of this monism: *śabdapratyāsavāda* (the doctrine of superimposition on the word; which is the same as *śabdādhyāsavāda*), *śabda-parināmavāda* (the doctrine of transformation of the word), and *śabdavivartavāda* (the doctrine of unreal appearance of the word). According to the first two, the phenomenal world is still real, though either falsely superimposed on words or a genuine transformation of the word essence. The last doctrine, perhaps more consistent than the other two, is the view that the phenomenal distinctions are unreal appearances of an immutable word essence.

Maṇḍana attempted to integrate this linguistic philosophy into his own form of *advaitavāda,* though later followers of Saṅkara did not accept the doctrine of *sphoṭa.* Even Vācaspati, who accepted many of Maṇḍana's theories, rejected the theory of *sphoṭa* and in general conformed to the Śaṅkarite's acceptance of the Bhāṭṭa epistemology.

Nyāya-Vaiśeṣika. *The old school.* Although as early as the commentators Praśastapāda (5th century AD) and Uddyotakara (7th century AD) the authors of the Nyāya-Vaisesika schools used each other's doctrines and the fusion of the two schools was well on its way, the two schools continued to have different authors and lines of commentators. About the 10th century AD, however, there arose a number of texts that sought to combine the two philosophies more successfully. Well known among these syncretist texts are the following: Bhāsarvajña's *Nyāyasāra* ("The Essence of Nyāya"; written *c.* 950), Varadarāja's *Tārkikarakṣā* ("In Defense of the Logician"; *c.* 1150), Vallabha's *Nyāyalīlāvatī* ("The Charm of Nyāya"; 12th century), Keśava Miśra's *Tarkabhāṣā* ("The Language of Reasoning"; *c.* 1275), Annam Bhāṭṭa's *Tarkasaṃgraha* ("Compendium of Logic"; *c.* 1623), and Viśvanātha's *Bhāṣāpariccheda* ("Determination of the Meaning of the Verses"; 1634).

Both the Nyāya-Vaiśeṣika schools are realistic with regard to things, properties, relations, and universals. Both schools are pluralistic (also with regard to individual selves) and theistic. Both schools admit external relations (the relation of inherence being only partly internal), atomistic cosmology, new production, and the concept of existence (*sattā*) as the most comprehensive universal. Both schools regard knowledge as a quality of the self, and they subscribe to a correspondence theory regarding the nature of truth and a theory of pragmatism-cum-coherence regarding the test of truth. The points that divide the schools are rather unimportant: they concern, for example, their theories of number, and some doctrines in their physical and chemical theories.

The Nyāya and Vaiśeṣika commentary tradition

Gautama's *sūtra*s were commented upon about AD 400 by Vātsāyana, who replied to the Buddhist doctrines, especially to some varieties of Śūnyavāda skepticism. Uddyotakara's *Vārttika* (*c.* 635) was written after a period during which major Buddhist works, but no major Hindu work, on logic were written. Uddyotakara undertook to refute Nāgārjuna and Dignāga. He criticized and refuted Dignāga's theory of perception, the Buddhist denial of soul, and the *anyāpoha* (exclusion of the other) theory of meaning. Positively, he introduced, for the first time, the doctrine of six modes of contact (*saṃnikarsa*) of the senses with their objects, which has remained a part of Nyāya-Vaiśeṣika epistemology. He divided inferences into those whose major premise (*sādhya*) is universally present, those in which one has to depend only upon the rule "Wherever there is absence of the major, there is absence of the middle (*hetu*)," and those in which both the positive and the negative rules are at one's disposal. He rejected the *sphoṭa* theory and argued that the meaning of a word is apprehended by hearing the last letter of the word together with recollection of the preceding ones. Vācaspati Miśra in the 9th century wrote his *Tātparyaṭīkā* (*c.* 840) on Uddyotakara's *Vārttika* and further strengthened the Nyāya viewpoint against the Buddhists. He divided perception into two kinds: the indeterminate, nonlinguistic, and nonjudgmental and the determinate and judgmental. In defining the invariable connection (*vyāpti*) between the middle and the major premises, he introduced the concept of a vitiating condition (*upādhi*) and stressed that the required sort of connection, if an inference is to be valid, should be unconditional. He also proposed a modified version of the theory of the extrinsic validity of knowledge by holding that inferences as well as knowledges that are the last verifiers (*phalajñāna*) are self-validating.

Praśastapāda's Vaiśeṣika commentary (*c.* 5th century) does not closely follow the *sūtra*s but is rather an independent explanation. Praśastapāda added seven more qualities to Kaṇāda's list: heaviness (*gurutva*), fluidity (*dravatva*), viscidity (*sneha*), traces (*saṃskara*), virtue (*dharma*), vice (*adharma*), and sound. The last quality was regarded by Kaṇāda merely as a mark of ether, whereas Praśastapāda elevated it to a defining quality of the latter. He also made the Vaiśeṣika fully theistic by introducing doctrines of creation and dissolution.

Metaphysics and epistemology

The Nyāya-Vaiśeṣika general metaphysical standpoint allows for both particulars and universals, both change and permanence. There are ultimate differences as well as a hierarchy of universals, the highest universal being existence. Substance is defined as the substrate of qualities and in terms of what alone can be an inherent cause. A quality may be defined as what is neither substance nor action and yet is the substratum of universals (for universals are supposed to inhere only in substances, qualities, and actions). Universal is defined as that which is eternal and inheres in many. Ultimate particularities belong to eternal substances, such as atoms and souls, and these account for all differences among particulars that cannot be accounted for otherwise. Inherence (*samavāya*) is the relation that is maintained between a universal and its instances, a substance and its qualities or actions, a whole and its parts, and an eternal substance and its particularity. This relation is such that one of the relations cannot exist without the other (*e.g.,* a whole cannot exist without the parts). Negation (*abhāva*), the seventh category, is initially classified into difference ("A is not B") and absence ("A is not in B"), absence being further divided into absence of a thing before its origin, its absence after its destruction, and its absence in places other than where it is present. For these schools, all that is is knowable and also nameable.

Knowledge is regarded as a distinguishing but not essential property of a self. It arises when the appropriate conditions are present. Consciousness is defined as a manifestation of object but is not itself self-manifesting; it is known by an act of inner perception (*anuvyavasāya*). Knowledge either is memory or is not; knowledge other than memory is either true or false; and knowledge that is not true is either doubt or error. In its theory of error, these philosophers maintained an uncompromising realism by holding that the object of error is still real but is only not here and now. True knowledge (*pramā*) apprehends its object as it is; false knowledge apprehends the object as what it is not. True knowledge is either perception, inference, or knowledge derived from verbal testimony or comparison. Perception is defined as knowledge that arises from the contact of the senses with their objects, and is viewed as either indeterminate and nonlinguistic

or as determinate and judgmental. Both aspects of the definition of perception are viewed as valid—a point that is made against both the Buddhists and the grammarians. Furthermore, perception is either ordinary (*laukika*) or extraordinary (*alaukika*). The former takes place through any of the six modes of sense-object contact recognized in the system. The latter takes place when one perceives the proper object of one sense through another sense ("The cushion looks soft") or when, on recognizing universal in a particular, one perceives all instances of the universal as its instances. Also extraordinary are the perceptions of the *yogin*s, who are supposed to be free from the ordinary spatiotemporal limitations.

Four conditions must be satisfied in order that a combination of words may form a meaningful sentence: a word should generate an intention or expectancy for the words to follow ("Bring"—"What?"—"A jar"); there should be mutual fitness ("Sprinkle"—"With what?"—"Water, not fire"); there should be proximity in space and time; and the proper intention of the speaker must be ascertained, otherwise there would be equivocation.

Among theistic proofs offered in the system, the most important are the causal argument ("The world is produced by an agent, since it is an effect, as is a jar"); the argument from a world order to a lawgiver; and the moral argument from the law of *karma* to a moral governor. Besides adducing these and other arguments, Udayana in his *Nyāyakusumāñjali* stressed the point that the nonexistence of God could not be proved by means of valid knowledge.

The new school. The founder of the school of Navya-(New) Nyāya, with an exclusive emphasis on the *pramaṇa*s, was Gaṅgeśa Upādhyāya (13th century), whose *Tattvacintāmaṇi* ("The Jewel of Thought on the Nature of Things") is the basic text for all later developments. The logicians of this school were primarily interested in defining their terms and concepts and for this purpose developed an elaborate technical vocabulary and logical apparatus that came to be used by, other than philosophers, writers on law, poetics, aesthetics, and ritualistic liturgy. The school may broadly be divided into two subschools: the Mithila school represented by Vardhamāna (Gaṅgeśa's son), Pakṣadhara or Jayadeva (author of *Āloka* gloss), and Śaṅkara Miśra (author of *Upaskāra*); and the Navadvīpa school, whose chief representatives were Vāsudeva Sārvabhauma (1450–1525), Raghunātha Śiromaṇi (*c.* 1475–*c.* 1550), Mathurānātha Tarkavāgīśa (fl. *c.* 1570), Jagadīśa Tarkālaṅkāra (fl. *c.* 1625), and Gadādhara Bhaṭṭacārya (fl. *c.* 1650).

Navya-Nyāya metaphysics and epistemology

By means of a new technique of analyzing knowledge, judgmental knowledge can be analyzed into three kinds of epistemological entities in their interrelations: "qualifiers" (*prakāra*); "qualificandum," or that which must be qualified (*viśeṣya*); and "relatedness" (*saṃsarga*). There also are corresponding abstract entities: qualifierness, qualificandumness, and relatedness. The knowledge expressed by the judgment "This is a blue pot" may then be analyzed into the following form: "The knowledge that has a qualificandumness in what is denoted by 'this' is conditioned by a qualifierness in blue and also conditioned by another qualifierness in potness."

A central concept in the Navya-Nyāya logical apparatus is that of "limiterness" (*avacchedakatā*), which has many different uses. If a mountain possesses fire in one region and not in another, it can be said, in the Navya-Nyāya language, "The mountain, as limited by the region *r*, possesses fire, but as limited by the region *r′* possesses the absence of fire." The same mode of speech may be extended to limitations of time, property, and relation, particularly when one is in need of constructing a description that is intended to suit exactly some specific situation and none other.

Inference is defined by Vātsāyana as the "posterior" knowledge of an object (*e.g.,* fire) with the help of knowledge of its mark (*e.g.,* smoke). For Navya-Nyāya, inference is definable as the knowledge caused by the knowledge that the minor term (*pakṣa,* "the hill") "possesses" the middle term (*hetu,* "smoke"), which is recognized as "pervaded by" the major (*sādhya,* "fire"). The relation of invariable connection, or "pervasion," between the middle (smoke) and the major (fire)—"Wherever there is smoke, there is fire"—is called *vyāpti.*

The logicians developed the notion of negation to a great degree of sophistication. Apart from the efforts to specify a negation with references to its limiting counterpositive (*pratiyogi*), limiting relation, and limiting locus, they were constrained to discuss and debate such typical issues as the following: Is one to recognize, as a significant negation, the absence of a thing *x* so that the limiter of the counterpositive *x* is not *x*-ness but *y*-ness? In other words, can one say that a jar is absent as a cloth even in a locus in which it is present as a jar? Also, is the absence of an absence itself a new absence or something positive? Furthermore, is the absence of colour in general nothing but the sum total of the absences of the particular colours, or is it a new kind of absence, a generic absence? Gaṅgeśa argued for the latter alternative, though he answers the first of the above three questions in the negative.

Introduction of new abstract entities

Though the philosophers of this school did not directly write on metaphysics, they nevertheless did tend to introduce many new kinds of abstract entities into their discourse. These entities are generally epistemological, though sometimes they are relational. Chief of these are entities called "qualifierness," "qualificandumness," and "limiterness." Various relations were introduced, such as direct and indirect temporal relations, *paryāpti* relation (in which a number reside, in sets rather than in individual members of those sets), *svarūpa* relation (which holds, for example, between an absence and its locus), and relation between a knowledge and its object.

Among the Navya-Nyāya philosophers, Raghunātha Śiromaṇi in *Padārthatattvanirūpaṇa* undertook a bold revision of the traditional categorial scheme by (1) identifying "time," "space," and "ether" with God; (2) eliminating the category of mind by reducing it to matter; (3) denying atoms (*paramāṇu*) and dyadic (paired) combinations of them (*dvyaṇuka*), (4) eliminating "number," "separateness," "remoteness," and "proximity" from the list of qualities; and (5) rejecting ultimate particularities (*viśeṣa*) on the grounds that it is more rational to suppose that the eternal substances are by nature distinct. He added some new categories, however, such as causal power (*śakti*) and the moment (*kṣaṇa*), and recognized that there are as many instances of the relation of inherence as there are cases of it (as contrasted with the older view that there is only one inherence that is itself present in all cases of inherence).

Sāṃkhya and Yoga. *Texts and commentaries until Vācaspati and the "Sāṃkhya-sūtras."* There are three commentaries on the *Sāṃkhya-kārikā:* that by Raja, much referred to but not extant; that by Gauḍapāda (7th century), on which there is a subcommentary *Candrikā* by Nārāyaṇatīrtha; and the *Tattva-kaumudī* by Vācaspati (9th century). The *Sāṃkhya-sūtra*s are a much later work (*c.* 14th century) on which Aniruddha (15th century) wrote a *vṛtti* and Vijñānabhikṣu (16th century) wrote the *Sāṃkhya-pravacana-bhāṣya* ("Commentary on the Sāṃkhya Doctrine"). Among independent works, mention may be made of *Tattvasamāsa* ("Collection of Truths"; *c.* 11th century).

The *Yoga-sūtra*s were commented upon by Vyāsa in his *Vyāsa-bhāṣya* (5th century), which again has two excellent subcommentaries: Vācaspati's *Tattvavaiśāradī* and Vijñānabhikṣu's *Yogavārttika,* besides the *vṛtti* by Bhoja (*c.* 1000).

Metaphysics and epistemology. For Vācaspati, creation was viewed in terms of the mere presence of the selves and the mere presentation to them of Matter (the undifferentiated primeval stuff). Such a view has obvious difficulties, for it would make creation eternal, because the selves and Matter are eternally copresent. Vijñānabhikṣu considered the relation between the selves and Matter to be a real relation that affects Matter but leaves the selves unaffected. Creation, in accordance with Bhikṣu's theism, is due to the influence of the chief self—*i.e.,* God. Furthermore, whereas the earlier Sāṃkhya authors, including Vācaspati, did not consider the question about the ontological status of the *guṇa*s, Bhikṣu regards them as real, as extremely subtle substances—so that each *guṇa* is held to be infinite in number. In general, the *Sāṃkhya-sūtra*s

The tendency to de-emphasize the difference between Sāṃkhya and Vedānta

show a greater Brahmanical influence, and there is a clear tendency to explain away the points of difference between the Sāṃkhya and the Vedānta. The author of the *sūtra*s tried to show that the Sāṃkhya doctrines are consistent with theism or even with the Upaniṣadic conception of Brahman. Vijñānabhikṣu made use of such contexts to emphasize that the atheism of Sāṃkhya is taught only to discourage men to try to be God, that originally the Sāṃkhya was theistic, and that the original Vedānta also was theistic. The Upaniṣadic doctrine of the unity of selves is interpreted by him to mean an absence of difference of kind among selves, which is consistent with the Sāṃkhya. *Māyā* (illusion) for Bhikṣu means nothing but the *prakṛti* (Matter) of the Sāṃkhya. The *sūtra*s also give cosmic significance to *mahat,* the first aspect to evolve from Matter, which then means cosmic Intelligence; a sense not found in the *kārikā*s.

In epistemology the idea of reflection of the spirit in the organs of knowing, particularly in the *buddhi,* or intelligence, comes to the forefront. Every cognition (*jñāna*) is a modification of the *buddhi,* with consciousness reflected in it. Though this is Vācaspati's account, it does not suffice according to Bhikṣu. If there is the mere reflection of the self in the state of the *buddhi,* this can only account for the fact that the state of cognition seems to be a conscious state; it cannot account for the fact that the self considers itself to be the owner and experiencer of that state. Accounting for this latter fact, Bhikṣu postulated a real contact between the self and *buddhi* as a reflection of the *buddhi* state back in the self.

Vācaspati, taking over a notion emphasized in Indian epistemology for the first time by Kumārila, introduced into the Sāṃkhya theory of knowledge a distinction between two stages of perceptual knowledge. In the first, a stage of nonconceptualized (*nirvikalpaka*) perception, the object of perception is apprehended vaguely and in a most general manner. In the second stage, this vague knowledge (*ālocanamātram*) is then interpreted and conceptualized by the mind. The interpretation is not so much synthesis as analysis of the vaguely presented totality into its parts. Bhikṣu, however, ascribed to the senses the ability to apprehend determinate properties, even independently of the aid of *manas.* For Sāṃkhya, in general, error is partial truth; there is no negation of error, only supplementation, though later Sāṃkhya authors tended to ascribe error to wrong interpretation.

An important contribution to epistemology was made by the writers on the Yoga: this concerns the key notion of *vikalpa,* which stands for mental states referring to pseudo-objects posited only by words. Such mental states are neither "valid" nor "invalid" and are said to be unavoidable accompaniments of one's use of language.

Ethics. Because the self is not truly an agent acting in the world, neither merit nor demerit, arising from one's actions, attaches to the self. Morality has empirical significance. In the long run, what really matters is knowledge. Nonattached performance of one's duties is an aid toward purifying intelligence so that it may be conducive to the attainment of knowledge: hence the importance of the restraints and observances laid down in the *Yoga-sūtra*s. The greatest good is freedom—*i.e.,* aloofness (*kaivalya*) from matter.

Rāja Yoga and Haṭha Yoga. Though Patañjali's yoga is known as Rāja Yoga (that in which one attains to self-rule), Haṭha Yoga (*haṭha* = "violence," "violent effort": *ha* = "sun," *ṭha* = "moon," *haṭha* = "sun and moon," breaths, or breaths travelling through the right and left nostrils) emphasizes bodily postures, regulation of breathing, and cleansing processes as means to spiritual perfection. A basic text on Haṭha Yoga is the *Haṭha-yoga-pradīpikā* ("Light on the Haṭha Yoga"; *c.* 15th century). As to the relation between the two yogas, a well-known maxim lays down that "No *rāja* without *haṭha,* and no *haṭha* without *rāja.*"

Religious consequences. The one religious consequence of the Sāṃkhya-Yoga is an emphasis on austere asceticism and a turning away from the ritualistic elements of Hinduism deriving from the Brahmanical sources. Though they continue to remain as an integral part of the Hindu faith, no major religious order thrived on the basis of these philosophies.

Vedānta. *Fragments from the Māṇḍukya-kārikā until Śaṅkara.* No commentary on the *Vedānta-sūtra*s survives from the period before Śaṅkara, though both Śaṅkara and Rāmānuja referred to the *vṛtti*s by Bodhāyana and Upavarṣa (the two may indeed be the same person). There are, however, pre-Śaṅkara monistic interpreters of the scriptures, three of whom are important: Bhartṛhari, Maṇḍana (both mentioned earlier), and Gauḍapāda. Śaṅkara referred to Gauḍapāda as the teacher of his own teacher Govinda, complimented him for having recovered the *advaita* (nondualism) doctrine from the Vedas, and also wrote a *bhāṣya* on Gauḍapāda's main work: the *kārikā*s on *Māṇḍukya Upaniṣad.*

The revival of non-dualism

Gauḍapāda's *kārikā*s are divided into four parts: the first part is an explanation of the *Upaniṣad* itself, the second part establishes the unreality of the world, the third part defends the oneness of reality, and the fourth part, called *Alātaśānti* ("Extinction of the Burning Coal"), deals with the state of release from suffering. It is not accidental that Gauḍapāda used as the title of the fourth part of his work a phrase in common usage among Buddhist authors. His philosophical views show a considerable influence of Mādhyamika Buddhism, particularly of the Yogācāra school, and one of his main purposes probably was to demonstrate that the teachings of the *Upaniṣad*s are compatible with the main doctrines of the Buddhist idealists. Among his principal philosophical theses were the following: All things are as unreal as those seen in a dream, for waking experience and dream are on a par in this regard. In reality, there is no production and no destruction. His criticisms of the categories of change and causality are reminiscent of Nāgārjuna's. Duality is imposed on this one reality by *māyā,* or the power of illusion-producing ignorance. Because there is no real coming into being, Gauḍapāda's philosophy is often called *ajātivāda* ("discourse on the unborn"). Though thus far agreeing with the Buddhist *Yogācārin*s, Gauḍapāda rejected their thesis that *citta,* or mind, is real and that there is a real flow of mental conception.

Śaṅkara greatly moderated Gauḍapāda's extreme illusionistic theory. Though he regarded the phenomenal world as a false appearance, he never made use of the analogy of dream. Rather, he contrasted the objectivity of the world with the subjectivity of dreams and hallucinations. The distinction between the empirical and the illusory—both being opposed to the transcendental—is central to his way of thinking.

Varieties of Vedānta schools. Though Vedānta is frequently referred to as one *darśana* (viewpoint), there are, in fact, radically different schools of Vedānta; what binds them together is common adherence to a common set of texts. These texts are the *Upaniṣad*s, the *Vedānta-sūtra*s, and the *Bhagavadgītā*—known as the three *prasthāna*s (the basic scriptures, or texts) of the Vedānta. The founders of the various schools of Vedānta have all substantiated their positions by commenting on these three source books. The problems and issues around which their differences centre are the nature of Brahman; the status of the phenomenal world; the relation of finite individuals to the Brahman; and the nature and the means to *mokṣa,* or liberation. The main schools are: Śaṅkara's unqualified nondualism (*śuddhādvaita*); Rāmānuja's qualified nondualism (*viśiṣṭādvaita*), Madhva's dualism (*dvaita*); Bhāskara's doctrine of identity and difference (*bhedābheda*); and the schools of Nimbārka and Vallabha, which assert both identity and difference though with different emphasis on either of the two aspects. From the religious point of view, Śaṅkara extolled metaphysical knowledge as the sole means to liberation and regarded even the concept of God as false; Rāmānuja recommended the path of *bhakti* combined with knowledge and showed a more tolerant attitude toward the tradition of Vedic ritualism; and Madhva, Nimbārka, and Vallabha all propounded a personalistic theism in which love and devotion to a personal God are rated highest. Although Śaṅkara's influence on Indian philosophy could not be matched by these other schools of Vedānta, in actual religious life the theistic Vedānta

schools have exercised a much greater influence than the abstract metaphysics of Śaṅkara.

The concepts of Advaita (nondualism)

Śaṅkara's philosophy is one among a number of other nondualistic philosophies: Bhartṛhari's *śabdādvaita,* the Buddhist's *vijñānadvaita,* and Gauḍapāda's *ajātivada.* Śaṅkara's system may then be called *ātmādvaita*—the thesis that the one, universal, eternal, and self-illuminating self whose essence is pure consciousness without a subject (*āśraya*) and without an object (*viṣaya*) from a transcendental point of view alone is real. The phenomenal world and finite individuals, though empirically real, are—from the higher point of view—merely false appearances. In substantiating this thesis Śaṅkara relied as much on the interpretation of scriptural texts as on reasoning. He set down a methodological principle that reason should be used only to justify truths revealed in the scriptures. His own use of reasoning was primarily negative; he showed great logical skill in refuting his opponents' theories. Śaṅkara's followers, however, supplied what is missed in his works—*i.e.,* a positive rational support for his thesis.

Śaṅkara's metaphysics is based on a criterion of reality, which may be briefly formulated as follows: the real is that whose negation is not possible. It is then argued that the only thing that satisfies this criterion is consciousness, because denial of consciousness presupposes the consciousness that denies. It is conceivable that any object is not existent, but the absence of consciousness is not conceivable. Negation may be either mutual negation (of difference) or absence. The latter is either absence of a thing prior to its origination or after its destruction or absence of a thing in a place other than where it is present. If the negation of consciousness is not conceivable, then none of these various kinds of negations can be predicated of consciousness. If difference cannot be predicated of it, then consciousness is the only reality and anything different from it would be unreal. If the other three kinds of absence are not predicable of it, then consciousness should be beginningless, without end, and ubiquitous. Consequently, it would be without change. Furthermore, consciousness is self-intimating; all objects depend upon consciousness for their manifestation. Difference may be either among members of the same class or of one individual from another of a different class or among parts of one entity. None of these is true of consciousness. In other words, there are not many consciousnesses; the plurality of many centres of consciousness should be viewed as an appearance. There is no reality other than consciousness—*i.e.,* no real *prakṛti;* such a thing would only be an unreal other. Also, consciousness does not have internal parts; there are not many conscious states. The distinction between consciousness of blue and consciousness of yellow is not a distinction within consciousness but one superimposed on it by a distinction among its objects, blue and yellow. With this, the Sāṃkhya, Vijñānavādin Buddhist, and Nyāya-Vaiśeṣika pluralism are refuted. Reality is one, infinite, eternal, and self-shining spirit; it is without any determination, for all determination is negation.

Śaṅkara's theory of error

The basic problem of Śaṅkara's philosophy is how such pure consciousness appears, in ordinary experience, to be individualized ("my consciousness") and to be of an object ("consciousness *of* blue"). As he stated it, subject and object are as opposed to each other as light and darkness, yet the properties of one are superimposed on the other. If something is a fact of experience and yet ought not to be so—*i.e.,* is rationally unintelligible—then this must be false. According to Śaṅkara's theory of error, the false appearance is a positive, presented entity that is characterized neither as existent (because it is sublated when the illusion is corrected) nor as nonexistent (because it is presented, given as much as the real is). The false, therefore, is indescribable either as being or as nonbeing, it is not a fiction, such as a round square. Śaṅkara thus introduced a new category of the "false" apart from the usual categories of the existent and the nonexistent. The world and finite individuals are false in this sense: they are rationally unintelligible, their reality is not logically deducible from Brahman, and their experience is cancelled with the knowledge of Brahman. The world and finite selves are not creations of Brahman; they are not real emanations or transformations of it. Brahman is not capable of such transformation or emanation. They are appearances that are superimposed on Brahman because of man's ignorance. This superimposition was sometimes called *adhyāsa* by Śaṅkara and was often identified with *avidyā.* Later writers referred to *avidyā* as the cause of the error. Thus, ignorance came to be regarded as a beginningless, positive something that conceals the nature of reality and projects the false appearances on it. Śaṅkara, however, did distinguish between three senses of being: the merely illusory (*prātibhāsika*), the empirical (*vyāvahārika;* which has unperceived existence and pragmatic efficacy), and transcendental being of one, indeterminate Brahman.

In his epistemology, Śaṅkara's followers in general accepted the point of view of the Mīmāṃsā of Kumārila's school. Like Kumārila, they accepted six ways of knowing: perception, inference, verbal testimony, comparison, nonperception, and postulation. In general, cognitions are regarded as modifications of the inner sense in which the pure spirit is reflected or as the pure spirit limited by respective mental modifications. The truth of cognitions is regarded as intrinsic to them, and a knowable fact is accepted as true so long as it is not rejected as false. In perception a sort of identity is achieved between the form of the object and the form of the inner sense; in fact, the inner sense is said to assume the form of the object. In their theory of inference, the Nyāya five-membered syllogism is rejected in favour of a three-membered one. Furthermore, the sort of inference admitted by the Nyāya, in which the major term is universally present, is rejected, because nothing save Brahman has this property according to the system.

Śaṅkara's ethical and religious concerns

Śaṅkara regarded moral life as a necessary preliminary to metaphysical knowledge and thus laid down strict ethical conditions to be fulfilled by one who wants to study Vedānta. For him, however, the highest goal of life is to know the essential identity of his own self with Brahman, and though moral life may indirectly help in purifying the mind and intellect, over an extended period of time knowledge comes from following the long and arduous process whose three major stages are study of the scriptures under appropriate conditions, reflection aimed at removing all possible intellectual doubts about the nondualistic thesis, and meditation on the identity of *ātman* and Brahman. *Mokṣa* is not, according to Śaṅkara, a perfection to be achieved; it is rather the essential reality of one's own self to be realized through destruction of the ignorance that conceals it. God is how Brahman appears to an ignorant mind that regards the world as real and looks for its creator and ruler. Religious life is sustained by dualistic concepts: the dualism between man and God, between virtue and vice, and between this life and the next. In the state of *mokṣa,* these dualisms are transcended. An important part of Śaṅkara's faith was that *mokṣa* was possible in bodily existence. Because what brings this supreme state is the destruction of ignorance, nothing need happen to the body; it is merely seen for what it really is—an illusory limitation on the spirit.

Śaṅkara's chief direct pupils were Sureśvara, the author of *Vārttika* ("Gloss") on his *bhāṣya* and of *Naiṣkarmyasiddhi* ("Establishment of the State of Non-Action"), and Padmapāda, author of *Pañcapādika,* a commentary on the first five *pāda*s, or sections, of the *bhāṣya.* These early pupils raised and settled issues that were not systematically discussed by Śaṅkara himself—issues that later divided his followers into two large groups: those who followed the *Vivaraṇa* (a work written on Padmapāda's *Pañcapādika* by one Prakāśātman in the 12th century) and those who followed Vācaspati's commentary (known as *Bhāmatī*) on Śaṅkara's *bhāṣya.* Among the chief issues that divided Śaṅkara's followers was the question about the locus and object of ignorance. The *Bhāmatī* school regarded the individual self as the locus of ignorance and sought to avoid the consequent circularity (arising from the fact that the individual self is itself a product of ignorance) by postulating a beginningless series of such selves and their ignorances. The *Vivaraṇa* school regarded both the locus and the object of ignorance to be Brahman and sought to avoid the contradiction (arising from the fact that

Brahman is said to be of the nature of knowledge) by distinguishing between pure consciousness and valid knowledge (*pramājñāna*). The latter, a mental modification, destroys ignorance, and the former, far from being opposed to ignorance, manifests ignorance itself, as evidenced by the judgment "I am ignorant." The two schools also differed in their explanations of the finite individual. The *Bhāmatī* school regarded the individual as a limitation of Brahman just as the space within the four walls of a room is a limitation of the big space. The *Vivaraṇa* school preferred to regard the finite individual as a reflection of Brahman in the inner sense. As the moon is one, but its reflections are many, so also Brahman is one, but its reflections are many. Later followers of Śaṅkara, such as Śrīharṣa in his *Khaṇḍanakhaṇḍakhādya* and his commentator Citsukha, used a destructive, negative dialectic in the manner of Nāgārjuna to criticize man's basic concepts about the world.

Concepts of *bhedābheda*

The philosophies of transcendence and immanence (*bhedābheda*) assert both identity and difference between the world and finite individuals, on the one hand, and Brahman, on the other. The world and finite individuals are real and yet both different and not different from the Brahman.

Among pre-Śaṅkara commentators on the *Vedāntasūtras*, Bhartṛprapañca defended the thesis of *bhedābheda,* and Bhāskara (*c.* 9th century) closely followed him. Bhartṛprapañca's commentary is not extant; the only known source of knowledge is Śaṅkara's reference to him in his commentary on the *Bṛhādaraṇyaka Upaniṣad,* in which Bhartṛprapañca is said to have held that though Brahman as cause is different from Brahman as effect, the two are identical inasmuch as the effect dissolves into the cause, as the waves return into the sea. Bhāskara viewed Brahman as both the material and the efficient cause of the world. The doctrine of *māyā* was totally rejected. Brahman undergoes the modifications by his own power. As waves are both different from and identical with the sea, so are the world and the finite individuals in relation to Brahman. The finite selves are parts of Brahman, as sparks of fire are parts of fire. But the finite soul exists, since beginningless time, under the influence of ignorance. It is atomic in extension and yet animates the whole body. Corresponding to the material world and the finite selves, Bhāskara ascribed to God two powers of self-modification. Bhāskara, in his theory of knowledge, distinguished between self-consciousness that is ever-present and objective knowledge that passively arises out of appropriate causal conditions but is not an activity. Mind, thus, is a sense organ. Bhāskara subscribed to the general Vedānta thesis that knowledge is intrinsically true, though falsity is extrinsic to it. In his ethical views, Bhāskara regarded religious duties as binding at all stages of life. He upheld a theory known as *jñāna-karmasamuccaya-vāda:* performance of duties together with knowledge of Brahman leads to liberation. In religious life, Bhāskara was an advocate of *bhakti,* but *bhakti* is not a mere feeling of love or affection for God, but rather is *dhyāna,* or meditation, directed toward the transcendent Brahman who is not exhausted in his manifestations. Bhāskara denied the possibility of liberation in bodily existence.

The *bhedābheda* point of view had various other adherents: Vijñānabhikṣu, Nimbārka, Vallabha, and Caitanya. Rāmānuja (11th century) sought to synthesize a long tradition of theistic religion with the absolutistic monism of the *Upaniṣads*, a task in which he had been preceded by no less an authority than the *Bhagavadgītā.* In his general philosophical position, he followed the *vṛttikāra* Bodhāyana, the Vākyakāra (to whom he referred but whose identity is not established except that he advocated a theory of real modification of Brahman), Nāthamuni (*c.* 1000), and his own teachers' teacher Yāmunācārya (*c.* 1050).

The main religious inspirations are from the theistic tradition of the Āḻvār poet-saints and their commentators known as the Ācāryas, who sought to combine knowledge with action (*karma*) as the right means to liberation. There is also, besides the Vedic tradition, the religious tradition of *Āgamas*, particularly of the Pāñcarātra literature. It is within this old tradition that Rāmānuja's philosophical and religious thought developed.

Rāmānuja rejected Śaṅkara's conception of Brahman as an indeterminate, qualityless, and differenceless reality on the ground that such a reality cannot be perceived, known, thought of, or even spoken about, in which case it is nothing short of a fiction. In substantiating this contention, Rāmānuja undertook, in his *Śrī-bhāṣya* on the *Vedāntasūtras*, a detailed examination of the different ways of knowing. Perception, either nonconceptualized or conceptualized, always apprehends its object as being something, the only difference between the two modes of perception being that the former takes place when one perceives an individual of a certain class for the first time and thus does not subsume it under the same class as some other individuals. Nor can inference provide one with knowledge of an indeterminate reality, because in inference one always knows something as coming under a general rule. The same holds true of verbal testimony. This kind of knowledge arises from understanding sentences. For Rāmānuja there is nothing like a pure consciousness without subject and without object. All consciousness is of something and belongs to someone. He also held that it is not true that consciousness cannot be the object of another consciousness. In fact, one's own past consciousness becomes the object of present consciousness. Consciousness is self-shining only when it reveals an object to its own owner—*i.e.,* the self.

Rāmānuja's views of Brahman

Rejecting Śaṅkara's conception of reality, Rāmānuja defended the thesis that Brahman is a being with infinitely perfect excellent virtues, a being whose perfection cannot be exceeded. The world and finite individuals are real, and together they constitute the body of Brahman. The category of body and soul is central to his way of thinking. Body is that which can be controlled and moved for the purpose of the spirit. The material world and the conscious spirits, though substantive realities, are yet inseparable from Brahman and thus qualify him in the same sense in which body qualifies the soul. Brahman is spiritual–material–qualified. Rāmānuja and his followers undertook criticisms of Śaṅkara's illusionism, particularly of his doctrine of *avidyā* (ignorance) and the falsity of the world. For Rāmānuja, such a beginningless, positive *avidyā* could not have any locus or any object, and if it does conceal the self-shining Brahman, then there would be no way of escaping from its clutches.

A most striking feature of Rāmānuja's epistemology is his uncompromising realism. Whatever is known is real, and only the real can be known. This led him to advocate the thesis that even the object of error is real—error is really incomplete knowledge—and correction of error is really completion of incomplete knowledge.

The state of *mokṣa* is not a state in which the individuality is negated. In fact, the sense of "I" persists even after liberation, for the self is truly the object of the notion of "I." What is destroyed is egoism, the false sense of independence. The means thereto is *bhakti,* leading to God's grace. But by *bhakti* Rāmānuja means *dhyāna,* or intense meditation with love. Obligation to perform one's scriptural duties is never transcended. Liberation is a state of blessedness in the company of God. A path emphasized by Rāmānuja for all persons is complete self-surrender (*prapatti*) to God's will and making oneself worthy of his grace. In his social outlook, Rāmānuja believed that *bhakti* does not recognize barriers of caste and classes.

The doctrinal differences among the followers of Rāmānuja is not so great as among Śaṅkara's. Writers such as Sudarśana Sūri and Veṅkatanātha continued to elaborate and defend the theses of the master, and much of their writing is polemical. Some differences are to be found regarding the nature of emancipation, the nature of devotion, and other ritual matters. The followers are divided into two schools: the Uttara-kalārya, led by Veṅkatanātha, and the Dakṣiṇa-kalārya, led by Lokācārya. One of the points at issue is whether or not emancipation is destructible; another, whether there is a difference between liberation attained by mere self-knowledge and that attained by knowledge of God. There also were differences in interpreting the exact nature of self-surrender to God and the degree of passivity or activity required of the worshipper.

Concepts of *Dvaita* (dualism)

Madhva (born 1199?) belonged to the tradition of Vaiṣṇava religious faith and showed a great polemical spirit in refuting Śaṅkara's philosophy and in converting people to his own fold. An uncompromising dualist, he traced back dualistic thought even to some of the *Upaniṣad*s. His main works are his commentaries on the *Upaniṣad*s, the *Gītā,* and the *Vedānta-sūtra*s. He also wrote a commentary on the *Mahābhārata* and several logical and polemical treatises.

He glorified difference. Five types of differences are central to Madhva's system: difference between soul and God, between soul and soul, between soul and matter, between God and matter, and that between matter and matter. Brahman is the fullness of qualities, and by his own intrinsic nature, Brahman produces the world. The individual, otherwise free, is dependent only upon God. The Advaita concepts of falsity and indescribability of the world were severely criticized and rejected. In his epistemology, Madhva admitted three ways of knowing: perception, inference, and verbal testimony. In Madhva's system the existence of God cannot be proved; it can be learned only from the scriptures.

Bondage and release both are real and devotion is the only way to release, but ultimately it is God's grace that saves. Scriptural duties, when performed without any ulterior motive, purify the mind and help one to receive God's grace.

Among the other theistic schools of Vedānta, brief mention may be made of the schools of Nimbārka (*c.* 12th century), Vallabha (15th century), and Caitanya (16th century).

Nimbārka's philosophy is known as Bhedābheda because he emphasized both identity and difference of the world and finite souls with Brahman. His religious sect is known as the Sanaka-sampradāya of Vaiṣṇavism. Nimbārka's commentary of the *Vedānta-sūtra*s is known as *Vedānta-pārijāta-saurabha* and is commented on by Śrīnivāsa in his *Vedānta-kaustubha.* Of the three realities admitted—God, souls, and matter—God is the independent reality, self-conscious, controller of the other two, free from all defects, abode of all good qualities, and both the material and efficient cause of the world. The souls are dependent, self-conscious, capable of enjoyment, controlled, atomic in size, many in number, and eternal but seemingly subject to birth and death because of ignorance and *karma.* Matter is of three kinds: nonnatural matter, which constitutes divine body; natural matter constituted by the three *guṇa*s; and time. Both souls and matter are pervaded by God. Their relation is one of difference-with-nondifference. Liberation is because of a knowledge that makes God's grace possible. There is no need for Vedic duties after knowledge is attained, nor is performance of such duties necessary for acquiring knowledge.

Vallabha's commentary on the *Vedānta-sūtra*s is known as *Aṇubhāṣya* ("The Brief Commentary"), which is commented upon by Puruṣottama in his *Bhāṣya-prakāśa* ("Lights on the Commentary"). His philosophy is called pure nondualism—"pure" meaning "undefiled by *māyā.*" His religious sect is known as the Rudra-sampradāya of Vaiṣṇavism and also Puṣṭimārga, or the path of grace. Brahman, or Śrī Krishna, is viewed as the only independent reality; in his essence he is existence, consciousness, and bliss, and souls and matter are his real manifestations. *Māyā* is but his power of self-manifestation. Vallabha admitted neither *pariṇāma* (of Sāṃkhya) nor *vivarta* (of Śaṅkara). According to him, the modifications are such that they leave Brahman unaffected. From his aspect of "existence" spring life, senses, and body. From "consciousness" spring the finite, atomic souls. From "bliss" spring the presiding deities, or *antaryāmin*s, for whom Vallabha finds place on his ontology. This threefold nature of God pervades all beings. World is real; but *saṃsāra,* the cycle of birth and death, is unreal, and time is regarded as God's power of action. Like all other Vedāntins, Vallabha rejected the Vaiśeṣika relation of *samavaya* and replaced it by *tādātmya,* or identity. The means to liberation is *bhakti,* which is defined as firm affection for God and also loving service (*sevā*). *Bhakti* does not lead to knowledge, but knowledge is regarded as a part of *bhakti.* The notion of "grace" plays an important role in Vallabha's religious thought. He is also opposed to renunciation.

The influence of Caitanya

Caitanya (1485–1533) was one of the most influential and remarkable of the medieval saints of India. His life is characterized by almost unique emotional fervour, hovering on the pathological, which was directed toward Śrī Krishna (the incarnation of Vishnu). He has not written anything, but the discourses recorded by contemporaries give an idea of his philosophical thought that was later developed by his followers, particularly by Rūpa Gosvāmin and Jīva Gosvāmin. Rūpa is the author of two great works: *Bhakti-rasāmṛta-sindhu* ("The Ocean of the Nectar of the Essence of Bhakti") and *Ujjvalanīlamani* ("The Shining Blue Jewel"). Jīva's main work is the great and voluminous *Ṣaṭsaṃdarbha.* These are the main sources of the philosophy of Bengal Vaiṣṇavism. Caitanya rejected the conception of an intermediate Brahman. Brahman, according to him, has three powers: the transcendent power that is threefold (the power of bliss, the power of being, and the power of consciousness) and the two immanent powers, namely, the powers of creating souls and the material world. Jīva Gosvāmin regarded bliss to be the very substance of Brahman who, with the totality of all his powers, is called God. Jīva distinguished between God's essential power, his peripheral power that creates the souls, and the external power (called *māyā*) that creates cosmic forms. The relation between God and his powers is neither identity nor difference, nor identity-with-difference. This relation, unthinkable and suprarational, is central to Caitanya's philosophy. For Jīva, the relation between any whole and its parts is unthinkable. *Bhakti* is the means to emancipation. *Bhakti* is conceived as a reciprocal relation between man and God, a manifestation of God's power in man. The works of Jīva and Rūpa delineated a detailed and fairly exhaustive classification of the types and gradations of *bhakti.*

Vaiṣṇava schools. The main philosophers of the medieval Vaiṣṇavism have been noted above. Vaiṣṇavism, however, has a long history, traceable to the Vishnu worship of the Rigveda, the Bhakti conception of the epics, and the Vāsudeva cult of the pre-Christian era. Of the two main Vaiṣṇava scriptures, or *āgama*s, the Pāñcarātra ("Relating to the Period of Five Nights") and the Vaikhānasa ("Relating to a Hermit or Ascetic") are the most important. Though Vaiṣṇava philosophers trace the Pāñcarātra works to Vedic origin, absolutists such as Śaṅkara refused to acknowledge this claim. The main topics of the Pāñcarātra literature concern rituals and forms of image worship and religious practices of the Vaiṣṇavas. Of philosophical importance are the *Ahirbudhnya-saṃhitā* ("Collection of Verses for Śiva") and *Jayākhya-saṃhitā* ("Collection of Verses Called Jayā"). The most well-known Pāñcarātra doctrine concerns the four spiritual forms of God: the absolute, transcendent state, known as Vāsudeva; the form in which knowledge and strength predominate (known as Saṃkarṣana); the form in which wealth and courage predominate (known as Pradyumna); and the form in which power and energy predominate (known as Aniruddha). Śaṅkara identified Saṃkarṣana with the individual soul, Pradyumna with mind, and Aniruddha with the ego sense. Furthermore, five powers of God are distinguished: creation, maintenance, destruction, favour, and disfavour. *Bhakti* is regarded as affection for God and associated with a sense of his majesty. The doctrine of *prapatti,* or complete self-surrender, is emphasized.

Śaiva schools. The Śaiva schools are the philosophical systems within the fold of Śaivism, a religious sect that worships Śiva as the highest deity. There is a long tradition of Śiva worship going back to the Rudra hymns of the Rigveda, the Śiva-Rudra of the Vājasaneyi-Saṃhitā, the Atharvaveda, and the *Brāhmaṇa*s. Mādhava in his *Sarva-darśana-saṃgraha* referred to three Śaiva systems: the Nakuliśa-Pāśupata, the Śaiva, and the Pratyabhijñā systems. The Śaiva system of Mādhava's classification probably corresponds to Śaiva-siddhānta of Tamil country, and the Pratyabhijñā is known as Kashmir Śaivism. The Śaiva-siddhānta is realistic and dualistic; the Kashmir system is idealistic and monistic.

Śaiva-siddhānta. The source literature of the Śaiva-

siddhānta school consists of the *Āgamas*, Tamil devotional hymns written by Śaiva saints but collected by Nambi (*c.* AD 1000) in a volume known as *Tirumurai, Civañāṇa-pōtam* ("Understanding of the Knowledge of Śiva") by Meykaṇṭatēvar (13th century), Śivācārya's *Śiva-jñānasiddhiyār* ("Attainment of the Knowledge of Śiva"), Umāpati's *Śivaprakāśam* ("Lights on Śiva") in the 14th century, Śrīkaṇṭha's commentary on the *Vedānta-sūtras* (14th century), and Appaya Dīkṣita's commentary thereon. This school admits three categories (*padārthas*): God (Śiva or Pati, Lord), soul (*paśu*), and the bonds (*pāśa*), and the 36 principles (*tattvas*). These 36 are divided into three groups: at the top, in order of manifestation from Śiva, are the five pure principles—*śivatattva* (the essence of Śiva), *śakti* (power), *sadā-śiva* (the eternal good), *īśivara* (lord), and *śuddha-vidyā* (true knowledge); seven mixed principles—pure *māyā,* five envelopes (destiny, time, interest, knowledge, and power), and *puruṣa,* or self; and 24 impure principles beginning with *prakṛti* (this list is broadly the same as that of Sāṃkhya). Śiva is the first cause: his *śakti,* or power, is the instrumental cause, *māyā* the material cause. This *māyā-śakti* is not God's essential power but is assumed by him; it is *parigraha-śakti* ("Assumed Power"). The relation of Śiva to his essential power is one of identity. Bonds are of three kinds: *karma, māyā,* and *avidyā.* The world and souls are real, and emancipation requires the grace of Śiva. The Śaiva-siddhānta always insisted on the preservation of the individuality of the finite soul, even in the state of emancipation, and rejected Śaṅkara's nondualism. Appaya Dīkṣita's commentary shows the tendency to attempt a reconciliation between the *Āgama* tradition of realism and pluralism with the Advaita tradition. The soul is eternal and all-pervasive, but, owing to original ignorance, it is reduced to the condition of *āṇava,* which consists in regarding oneself as finite and atomic. Knowledge of its own nature as well as God's is possible only by God's grace.

Śaivite categories and principles

Kashmir Śaivism. The source literature of this school consists in the *Śiva-sūtra,* Vasugupta's *Spanda-kārikā* ("Verses on Creation"; 8th–9th centuries), Utpala's *Pratyabhijñā-sūtra* ("Aphorisms on Recognition"; *c.* 900), Abhinavagupta's *Paramārthasāra* ("The Essence of the Highest Truth"), *Pratyabhijñā-vimarśini* ("Reflections on Recognition"), and *Tantrāloka* ("Lights on the Doctrine") in the 10th century, and Kṣemarāja's *Śiva-sūtra-vimarśini* ("Reflections on the Aphorisms on Śiva"). As contrasted with the Śaiva-siddhānta, this school is idealist and monist, and, although it accepts all the 36 *tattvas* and the three *padārthas,* it is Śiva, the Lord, who is the sole reality. God is viewed as both the material and efficient cause of the universe. Five aspects of God's power are distinguished: consciousness (*cit*), bliss (*ānanda*), desire (*icchā*), knowledge (*jñāna*), and action (*kriyā*). Śiva is one—without a second, infinite spirit. He has a transcendent aspect and an immanent aspect, and his power with its fivefold functions constitutes his immanent aspect. The individual soul of a person is identical with Śiva; recognition of this identity is essential to liberation.

JAINA PHILOSOPHY

Jainism, founded in about the 6th century BC by Vardhamāna Mahāvīra, the 24th in a succession of religious leaders known as Jinas (Conquerors), rejects the idea of God as the creator of the world but teaches the perfectibility of man, to be accomplished through the strictly moral and ascetic life. Central to the moral code of Jainism is the doctrine of *ahiṃsā*, or noninjury to all living beings, an idea that may have arisen in reaction to Vedic sacrifice ritual. There is also a great emphasis on vows (*vratas*) of various orders.

Although earlier scriptures, such as the *Bhagavatī-sūtra,* contained assorted ideas on logic and epistemology, Kundakunda of the 2nd century AD was the first to develop Jaina logic. The *Tattvarthādhigama-sūtra* of Umāsvatis, however, is the first systematic work, and Siddhasena (7th century AD) the first great logician. Other important figures are Akalanka (8th century), Mānikyanandi, Vādideva, Hemchandra (12th century), Prabhāchandra (11th century), and Yasovijaya (17th century).

The principal ingredients of Jaina metaphysics are: an ultimate distinction between "living substance" or "soul" (*jīva*) and "nonliving substance" (*ajīva*); the doctrine of *anekāntavāda,* or nonabsolutism (the thesis that things have infinite aspects that no determination can exhaust); the doctrine of *naya* (the thesis that there are many partial perspectives from which reality can be determined, none of which is, taken by itself, wholly true, but each of which is partially so); and the doctrine of *karma,* in Jainism a substance, rather than a process, that links all phenomena in a chain of cause and effect.

As a consequence of their metaphysical liberalism, the Jaina logicians developed a unique theory of seven-valued logic, according to which the three primary truth values are "true," "false," and "indefinite," and the other four values are "true and false," "true and indefinite," "false and indefinite," and "true, false, and indefinite." Every statement is regarded as having these seven values, considered from different standpoints.

The Jaina view of knowledge

Knowledge is defined as that which reveals both itself and another (*svaparabhāsi*). It is eternal, as an essential quality of the self; it is noneternal, as the perishable empirical knowledge. Whereas most Hindu epistemologists regarded *pramaṇā* as the cause of knowledge, the Jainas identified *pramaṇā* with valid knowledge. Knowledge is either perceptual or nonperceptual. Perception is either empirical or nonempirical. Empirical perception is either sensuous or nonsensuous. The latter arises directly in the self, not through the sense organs, but only when the covering ignorance is removed. With the complete extinction of all *karmas*, a person attains omniscience (*kevala-jñāna*). (See also JAINISM.)

MUGHAL PHILOSOPHY

Reference has been made earlier to the Ṣūfī (Islāmic mystics), who found a resemblance between the ontological monism of Ibn al-ʿArabi and that of Vedānta. The Shaṭṭārī order among the Indian Ṣūfīs practiced Yogic austerities and even physical postures. Various minor syncretistic religious sects attempted to harmonize Hindu and Muslim religious traditions at different levels and with varying degrees of success. Of these, the most famous are Rāmānanda, Kabīr, and Gurū Nānak. Kabīr harmonized the two religions in such a manner that, to an enquiry about whether he was a Hindu or a Muslim, the answer given by a contemporary was "It is a secret difficult to comprehend. One should try to understand." Gurū Nānak rejected the authority of both Hindu and Muslim scriptures alike and founded his religion (Sikhism) on a rigorously moralistic, monotheistic basis.

Among the great Mughals, Akbar attempted, in 1581, to promulgate a new religion, Dīn-e Ilāhī, which was to be based on reason and ethical teachings common to all religions and which was to be free from priestcraft. This effort, however, was short-lived, and a reaction of Muslim orthodoxy was led by Shaykh Aḥmed Sirhindī, who rejected ontological monism in favour of orthodox unitarianism and sought to channel mystical enthusiasm along Qurʾānic (Islāmic scriptural) lines. By the middle of the 17th century, the tragic figure of Dārā Shikōh, the Mughal emperor Shāh Jahān's son and disciple of the Qādirī sufis, translated Hindu scriptures, such as the *Bhagavadgītā* and the *Upaniṣads*, into Persian and in his translation of the latter closely followed Śaṅkara's commentaries. In his *Majmaʿ al-baḥrayn* he worked out correlations between Ṣūfī and Upaniṣadic cosmologies, beliefs, and practices. During this time, the Muslim elite of India virtually identified Vedānta with Ṣūfīsm. Later, Shāh Walī Allāh's son, Shāh ʿAbd-ul-ʿAzīz, regarded Krishna among the *awliyāʾ* (saints).

19TH- AND 20TH-CENTURY PHILOSOPHY IN INDIA AND PAKISTAN

Influence of Western philosophies

In the 19th century, India was not marked by any noteworthy philosophical achievements, but the period was one of great social and religious reform movements. The newly founded universities introduced Indian intellectuals to Western thought, particularly to the empiricistic, utilitarian, and agnostic philosophies in England, and John

Stuart Mill, Jeremy Bentham, and Herbert Spencer had become the most influential thinkers in the Indian universities by the end of the century. These Western-oriented ideas served to generate a secular and rational point of view and stimulated social and religious movements, most noteworthy among them being the Brahmo (Brahma) Samaj movement founded by Rammohan Ray. Toward the later decades of the century, the great saint Ramakrishna Paramahamsa of Calcutta renewed interest in mysticism, and many young rationalists and skeptics were converted into the faith exemplified in his person. Ramakrishna taught, among other things, an essential diversity of religious paths leading to the same goal, and this teaching was given an intellectual form by Swami Vivekananda, his famed disciple.

The first Indian graduate school in philosophy was founded in the University of Calcutta during the first decades of the 20th century, and the first incumbent of the chair of philosophy was Sir Brajendranath Seal, a versatile scholar in many branches of learning, both scientific and humanistic. Seal's major published work is *The Positive Sciences of the Ancient Hindus,* which, besides being a work on the history of science, shows interrelations among the ancient Hindu philosophical concepts and their scientific theories. Soon, however, the German philosophers Kant and Hegel came to be the most studied philosophers in the Indian universities. The ancient systems of philosophy came to be interpreted in the light of German idealism. The Hegelian notion of Absolute Spirit found a resonance in the age-old Vedānta notion of Brahman. The most eminent Indian Hegelian scholar is Hiralal Haldar, who was concerned with the problem of the relation of the human personality with the Absolute, as is evidenced by his book *Neo-Hegelianism.* The most eminent Kantian scholar is K.C. Bhattacharyya.

Contributions of modern Indian philosophers

Among those who deserve mention for their original contributions to philosophical thinking are Sri Aurobindo (died 1950), Mahatma Gandhi (died 1948), Rabindranath Tagore (died 1941), Sir Muḥammed Iqbāl (died 1938), K.C. Bhattacharyya (died 1949), and Sarvepalli Radhakrishnan (died 1975). Of these, Sri Aurobindo was first a political activist and then a *yogin,* Tagore and Iqbāl poets, Gandhi a political and social leader, and only Radhakrishnan and Bhattacharyya university professors. This fact throws some light on the state of Indian philosophy in this century.

In his major work, *The Life Divine,* Sri Aurobindo starts from the fact of human aspiration for a kingdom of heaven on earth and proceeds to give a theoretical framework in which such an aspiration would be not a figment of imagination but a drive in nature, working through man toward a higher stage of perfection. Both the denial of the materialist and that of the ascetic are rejected as being one-sided. The gulf between unconscious matter and fully self-conscious spirit is sought to be bridged by exhibiting them as two poles of a series in which spirit continuously manifests itself. The Vedāntic concept of a transcendent and all-inclusive Brahman is sought to be harmonized with a theory of emergent evolution. Illusionism is totally rejected. The purpose of man is to go beyond his present form of consciousness. Yoga is interpreted as a technique not for personal liberation but for cooperating with the cosmic evolutionary urge that is destined to take mankind ahead from the present mental stage to a higher, supramental stage of consciousness. A theory of history, in accordance with this point of view, is worked out in his *The Human Cycle.*

Rabindranath Tagore's philosophical thinking is no less based on the *Upaniṣad*s, but his interpretation of the *Upaniṣad*s is closer to Vaiṣṇava theism and the Bhakti cults than to traditional monism. He characterized the absolute as supreme person and placed love higher than knowledge. In his *Religion of Man,* Tagore sought to give a philosophy of man in which human nature is characterized by a concept of surplus energy that finds expression in creative art. In his lectures on *Nationalism,* Tagore placed the concept of society above that of the modern nation-state.

Mahatma Gandhi preferred to say that the truth is God rather than God is the truth, because the former proposition expresses a belief that even the atheists share. The belief in the presence of an all-pervading spirit in the universe led Gandhi to a strict formulation of the ethics of nonviolence (*ahiṃsā*). But he gave this age-old ethical principle a wealth of meaning so that *ahiṃsā* for him became at once a potent means of collective struggle against social and economic injustice, the basis of a decentralized economy and decentralized power structure, and the guiding principle of one's individual life in relation both to nature and to other persons. The unity of existence, which he called the truth, can be realized through the practice of *ahiṃsā*, which requires reducing oneself to zero and reaching the furthest limit of humility.

Influenced by the British philosopher J.M.E. McTaggart's form of Hegelian idealism and the French philosopher Henri Bergson's philosophy of change, Muḥammed Iqbāl conceived reality as creative and essentially spiritual, consisting of egos. "The truth, however, is that matter is spirit," he wrote,

> in space-time reference. The unity called man is body when we look at it as acting in regard to what we call external world; it is mind or soul when we look at it as acting in regard to the ultimate aim and ideal of such acting.

Influenced by British Neo-Hegelianism in his interpretation of the Vedāntic tradition, Sarvepalli Radhakrishnan was primarily an interpreter of Indian thought to the Western world. He defended a realistic interpretation of the concept of *māyā*—thereby playing down its illusionistic connotation, a theory of intuition as the means of knowing reality, and a theory of emergent evolution of spirit (not unlike Sri Aurobindo, but without his doctrine of supermind) in nature and history. The most original among modern Indian thinkers, however, is K.C. Bhattacharyya, who rejected the conception of philosophy as a construction of a worldview and undertook a phenomenological description of the various grades of subjectivity: (1) the bodily, (2) the psychic, and (3) the spiritual. With regard to (1), he distinguished between the objective body and the felt body and regarded the latter as the most primitive level of the subjective sense of freedom from the objective world. The stage (2) includes the range of mental life from image to free thought. In introspection, the level (2) is transcended, but various levels of introspection are distinguished, all leading to greater freedom from objectivity. It would seem, however, that for Bhattacharyya absolute freedom from objectivity was a spiritual demand. According to his theory of value, value is not an adjective of the object but a feeling absolute, of which the object evaluated appears as an adjective, and his logic of alternation is a modern working out of the Jaina theories of *anekānta* (non-absolutism) and *syādvāda* (doctrine of "may be").

Later developments

Among later philosophers, N.V. Banerjee (1901–81) and Kalidas Bhattacharyya (1911–84), the son of K.C. Bhattacharyya, have made important contributions. In *Language, Meaning and Persons* (1963), Banerjee examines the development of personhood from a stage of individualized bondage to liberation in a collective identity, a life-with-others. This liberation, according to Banerjee, also entails an awareness of time and freedom from spatialized objects.

In his earlier writings such as *Object, Content and Relation* (1951) and *Alternative Standpoints in Philosophy* (1953), Bhattacharyya developed his father's idea of theoretically undecidable alternatives in philosophy. In the later works *Philosophy, Logic and Language* (1965) and *Presuppositions of Science and Philosophy* (1974), he developed the concept of metaphysics as a science of the nonempirical a priori essences that are initially discerned as the structure of the empirical but are subsequently recognized as autonomous entities. The method of metaphysics for him is reflection, phenomenological and transcendental. Kalidas Bhattacharyya was concerned with the nature and function of philosophical reflection and its relation of unreflective experience. What reflection brings to light, he held, is present in pre-reflective experience, but only as undistinguished and fused, in a state of objective implicitness. The essences as such are not real but de-

mand realization in pure reflective consciousness. At the same time, he emphasized the limitations of any doctrine positing the constitution of nature in consciousness. Such a doctrine, he insisted, cannot be carried out in details.

Among those who apply the phenomenological method and concepts to understanding the traditional Indian philosophies, D. Sinha, R.K. Sinari, and J.N. Mohanty are especially noteworthy. Others who interpret the Indian philosophies by means of the methods and concepts of analytical philosophy include B.K. Matilal and G. Misra. In the field of philosophy of logic, P.K. Sen has worked on the paradoxes of confirmation and the concept of quantification, and Sibajiban on the liar paradox and on epistemic logic. Sibajiban and Matilal have made important contributions toward rendering the concepts of Navya-Nyāya logic into the language of modern logic. In ethics and social philosophy, notable work has been done by Abu Sayyid Ayub, Daya Krishna, Rajendra Prasad, and D.P. Chattopadhyaya.

BIBLIOGRAPHY

General: S.N. DASGUPTA, *A History of Indian Philosophy*, 5 vol. (1922–55, reprinted 1977), a comprehensive account, though its scholarship tends to outweigh philosophical insight; M. HIRIYANNA, *Outlines of Indian Philosophy* (1932, reissued 1975), lucidly written, based on reliable acquaintance with original source material but leaving out many minor, though important, schools of thought; S. RADHAKRISHNAN, *Indian Philosophy*, 2 vol. (1923–27), a very readable account written from an idealistic point of view—may often mislead; S.C. VIDYABHUSAN, *A History of the Mediæval School of Indian Logic* (1909), still indispensable, though outdated and containing many inaccuracies; U.N. GHOSHAL, *A History of Indian Political Ideas: The Ancient Period and the Period of Transition to the Middle Ages* (1959); DONALD H. BISHOP (ed.), *Indian Thought: An Introduction* (1975), 15 historical essays by Indian scholars; BALBIR SINGH, *The Conceptual Framework of Indian Philosophy* (1976), a study of 12 fundamental concepts; KARL H. POTTER (ed.), *The Encyclopedia of Indian Philosophies* (1970–), a very valuable source. For current articles on Indian philosophy, see *Philosophy East and West* (quarterly) and *The Journal of Indian Philosophy* (quarterly).

Critical studies from the point of view of modern Western thought: KARL H. POTTER, *Presuppositions of India's Philosophies* (1963, reprinted 1976); NINIAN SMART, *Doctrine and Argument in Indian Philosophy* (1964); and B.K. MATILAL, *Epistemology, Logic and Grammar in Indian Philosophical Analysis* (1971), three books that, together, form a good introduction to the logical, dialectical, and analytical aspects of Indian philosophy; KEWAL KRISHAN MITTAL, *Materialism in Indian Thought* (1974).

English translations of Sanskrit sources: S. RADHAKRISHNAN and C.A. MOORE (eds.), *A Source Book in Indian Philosophy* (1957), an excellent one-volume collection of source materials (does not include many medieval masterpieces on logic and epistemology); *The Thirteen Principal Upanishads*, 2nd ed., trans. by R.E. HUME (1931); *The Bhagavadgita*, trans. by S. RADHAKRISHNAN (1948).

Selected readings on the systems and texts: (*Upaniṣads*): ARUN SHOURIE, *Hinduism, Essence and Consequence: A Study of the Upanishads, the Gita, and the Brahma-Sutras* (1980), an analysis and assessment of Brahmanical Hinduism; R.D. RANADE, *A Constructive Survey of Upanishadic Philosophy*, 2nd ed. (1968). (*Bhagavadgītā*): SRI AUROBINDO, *Essays on the Gita* (1916–20, reissued 1974). (*Mahābhārata*): EDWARD W. HOPKINS, *The Great Epic of India* (1901, reprinted 1973). (*Cārvākas and Ājīvikas*): DAKSHINARANJAN SHASTRI, *A Short History of Indian Materialism, Sensationalism and Hedonism* (1930); DALE RIEPE, *Early Indian Philosophical Materialism* (1954); A.L. BASHAM, *The History and Doctrines of the Ājīvikas* (1951, reprinted 1981). (*Buddhism*): BENIMADHAB BARUA, *Prolegomena to a History of Buddhist Philosophy*, 2nd ed. (1974); DAVID J. KALUPAHANA, *Buddhist Philosophy: A Historical Analysis* (1976); A.L. HERMAN, *An Introduction to Buddhist Thought: A Philosophic History of Indian Buddhism* (1983); TH. STCHERBATSKY, *Buddhist Logic*, 2 vol. (Eng. trans. 1930–32, reissued 1970), a work of great scholarship, marred by too-hasty comparisons with 19th-century European philosophers, and containing an English translation of Dharmakīrti's *Nyayavindu;* T.R.V. MURTY, *The Central Philosophy of Buddhism* (1955, reissued 1980); (*Mīmāmsā*): A. BERRIEDALE KEITH, *Karma-mīmāmsā* (1921, reprinted 1978); P. SHASTRI, *Introduction to the Pūrva Mīmāmsā*, 2nd ed. (1980). (*Vedānta*): ERIC J. LOTT, *Vedantic Approaches to God* (1980), a clear introduction to the religious philosophies of Vedānta; JACOB KATTACKAL, *Religion and Ethics in Advaita* (1980, reprinted 1982); T.M.P. MAHADEVAN, *Gaudapāda: A Study in Early Advaita* (1952; 4th ed., 1975); ELIOT DEUTSCH, *Advaita Vedānta: A Philosophical Reconstruction* (1969, reissued 1973); P.N. SRINIVASACHARI, *The Philosophy of Viśiṣṭādvaita* (1943, reprinted 1973). (*Vaiṣṇavism and Śaivism*): R.G. BHANDARKAR, *Vaiṣṇavism, Śaivism and Minor Religious Systems* (1913, reissued 1980). (*Nyāya-Vaiśeṣika*): H. UI, *The Vaiśeṣhika Philosophy*, 2nd ed. (1962); N.S. JUNANKAR, *Gautama: The Nyāya Philosophy* (1978); S.C. CHATTERJEE, *Nyāya Theory of Knowledge* (1939); D.H. INGALLS, *Materials for the Study of Navya-Nyāya Logic* (1951); J.N. MOHANTY (trans.), *Gangesa's Theory of Truth* (1966); B.K. MATILAL, *The Navya-Nyāya Doctrine of Negation* (1969). (*Sāṃkhya-Yoga*): S.N. DASGUPTA, *The Study of Patanjali* (1920); MIRCEA ELIADE, *Yoga: Immortality and Freedom*, 2nd ed. (1969; originally published in French, 1954); KAREL WERNER, *Yoga and Indian Philosophy* (1977), a wide-ranging introduction. (*Mughal philosophy*): AZĪZ AHMAD, *Studies in Islamic Culture in the Indian Environment* (1964).

Contemporary Indian philosophy: S. RADHAKRISHNAN, *The Reign of Religion in Contemporary Philosophy* (1920); RABINDRANATH TAGORE, *Religion of Man* (1931, reprinted 1981); S. RADHAKRISHNAN, *Eastern Religions and Western Thought* (1939, reissued 1974); SRI AUROBINDO, *The Life Divine* (1949, reissued 1982); S. RADHAKRISHNAN and J.H. MUIRHEAD (eds.), *Contemporary Indian Philosophy*, 2nd ed. (1952); P.T. RAJU, *Idealistic Thought of India* (1953, reprinted 1973); KALIDAS BHATTACHARYYA, *Studies in Philosophy*, 2 vol. (1956–58), and (ed.), *Recent Indian Philosophy* (1962), and *Philosophical Papers* (1969); G. MISRA, *Analytical Studies in Indian Philosophical Problems* (1970); K.S. MURTY and K.R. RAO (eds.), *Current Trends in Indian Philosophy* (1972); MARGARET CHATTERJEE (ed.), *Contemporary Indian Philosophy* (1974); J.N. MOHANTY, "Philosophy in India: 1967–1973," *Review of Metaphysics*, 28:54–84 (1974); N.K. DEVARAJA (ed.), *Indian Philosophy Today* (1975); DALE RIEPE, *Indian Philosophy Since Independence* (1979), an assessment from the perspective of historical and dialectical materialism; T.M.P. MAHADEVAN and G.V. SAROJA, *Contemporary Indian Philosophy* (1981), brief accounts of the lives and thought of eight philosophers of the first half of the 20th century.

(J.N.M.)

Indonesia

The Republic of Indonesia (Indonesian: Republik Indonesia), an archipelago nation in the Indian and Pacific oceans, lies across the equator for one-eighth of the Earth's circumference off the coast of the Southeast Asian mainland. Its islands can be grouped into the Greater Sunda Islands of Sumatra (Sumatera), Java (Jawa), the southern extent of Borneo known as Kalimantan, and Celebes (Sulawesi); the Lesser Sunda Islands (Nusa Tenggara) of Bali and a chain of islands that runs eastward through Timor; the Moluccas (Maluku) between Celebes and the island of New Guinea; and the western extent of New Guinea known as Irian Jaya (formerly Irian Barat).

The country is the largest in Southeast Asia, about 752,400 square miles (1,948,700 square kilometres) in area, with a maximum dimension from east to west of about 3,200 miles (5,100 kilometres) and a dimension from north to south of 1,100 miles. It is composed of some 13,670 islands, of which more than 7,000 are uninhabited. Almost 75 percent of Indonesia's area is included in the three largest islands of Borneo, of which about three-quarters, or 212,000 square miles, is part of Indonesia; Sumatra, with 186,000 square miles; and the Irian Jaya portion of New Guinea, with 162,000 square miles. Nearly all of the total land area is accounted for with the addition of Celebes (75,100 square miles) and Java and the Moluccas (83,400 square miles).

Indonesia was formerly known as the Dutch, or Netherlands, East Indies; the islands were first named Indonesia in modern times by a German geographer in 1884, although this name is thought to derive from Indos Nesos, "Indian Islands," in the ancient trading language of the region. The capital, Jakarta, is located near the northwestern coast of Java.

After a period of occupation by the Japanese (1942–45) during World War II, Indonesia declared its independence from The Netherlands in 1945. Its struggle for independence, however, continued until 1949; and it was not until the official recognition by the United Nations of Irian Barat as a part of Indonesia in 1969 and the incorporation of the former Portuguese territory of East Timor in 1975–76 that the nation took on its present form.

The Indonesian archipelago represents one of the most unusual areas in the world, encompassing a major juncture of the Earth's tectonic plates, the dividing line between two faunal realms, and the meeting point for the peoples and cultures of mainland Asia and Oceania. These factors have created a highly diverse environment and society in which the only common elements are the susceptibility to seismic and volcanic activity, close proximity to the sea, and a moist, tropical climate.

In its economic development the country relies heavily upon its agricultural capacity—particularly rice cultivation—and the export of such cash crops as coconuts, rubber, and tea; its petroleum products, of which it is the major producer in Asia; its rich deposits of tin and other minerals; and timber.

Indonesia is the most populous nation in Southeast Asia, and it is advantageously located between mainland Asia and Australia. As such, the country has a critical role to play in the development of its part of the world. In keeping with its size and importance, it is active in such regional and international groupings as the Association of Southeast Asian Nations (ASEAN), the Economic and Social Commission for Asia and the Pacific (ESCAP), the Organization of Petroleum Exporting Countries (OPEC), and the United Nations.

The article is divided into the following sections:

Physical and human geography 201B
- The land 201B
  - Geologic framework
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant life
  - Animal life
  - Traditional regions
  - Settlement patterns
- The people 201J
  - Ethnic groups
  - Languages
  - Religions
  - Demographic trends
- The economy 201K
  - Resources
  - Mining and forestry
  - Agriculture and fishing
  - Industry
  - Finance
  - Commerce and trade
  - Transportation
- Administration and social conditions 201M
  - Government
  - Justice
  - Armed forces
  - Education
  - Health and welfare
  - Housing
- Cultural life 201O
  - The ancient culture
  - Traditional arts
  - Modern art forms

History 201P
- The archipelago and its early historical records 201P
- Indonesian "Hinduism" 201Q
  - Hindu religious conceptions
  - Indonesian religious conceptions
- The Malay kingdom of Śrīvijaya-Palembang 201R
  - Buddhism in Palembang
  - The maritime influence
- Central Java from the 8th to the 10th century 201S
- Eastern Java and the archipelago from 1019 to 1292 201T
  - Government and politics
  - The empire of Kertanagara
- The Majapahit era 201U
- Islāmic influence in Indonesia 201V
  - Muslim kingdoms of northern Sumatra
  - Muslims in Java
- Growth of the Dutch empire 201W
  - Dutch East India Company
  - The French and British in Java, 1806–15
- Dutch rule from 1815 to *c.* 1920 201Y
  - The Cultivation System
  - Dutch territorial expansion
  - The Ethical Policy
- Toward independence 201AA
  - The rise of nationalism
  - Japanese occupation
  - The revolution
- Independent Indonesia to 1965 201BB
  - The years of constitutional democracy
  - Introduction of Guided Democracy
  - Sukarno's policies
- Indonesia since 1965 201DD
  - The coup and after
  - Suharto's New Order

Bibliography 201EE

Physical and human geography

THE LAND

General nature of the islands' topography

The major Indonesian islands are characterized by rugged volcanic mountains, covered by dense tropical forests, which slope down to coastal plains often covered by thick alluvial swamps and bordered by shallow seas and coral reefs. Cultivated land is mainly devoted to rice, which in many areas is grown on mountain terraces, or to such cash crops as rubber. In the highly populated areas palm-shaded villages are scattered among green rice terraces, which are overlooked by the forest-clad cone of an active volcano.

Geologic framework. The physical structure of Indonesia is unique and complicated because it encompasses the junction of three major sections of the Earth's crust and involves a complex series of shelves, volcanic mountain chains, and deep-sea trenches. The island of Borneo and the arc of islands including Sumatra, Java, Bali, and the Lesser Sunda chain sit on the Sunda Shelf, a southward extension of the continental mass of Asia. The shelf is bounded on the south and west by deep-sea trenches such as the Java Trench (24,442 feet deep at its lowest point) that form the true continental boundary. The island of New Guinea and adjacent islands, possibly including Halmahera Island, sit on the Sahul Shelf, which is a northwestern extension of the Australian continental mass; the shelf is bounded to the northeast by a series of deep-sea troughs and to the northwest by troughs, a chain of coral reefs, and a series of submarine ridges. The third major unit of the Earth's crust in Indonesia is an extension of the belt of mountains of Japan and the Philippines that runs south between Borneo and New Guinea. It includes a series of mountain volcanoes and deep-sea trenches on and around Celebes and the Moluccas.

The interrelation of these units is not clearly understood. The present land-sea relations are somewhat misleading, because the seas that lie on the Sunda Shelf and on the Sahul Shelf are shallow and of geologically recent origin; they rest on the continental mass rather than on a true ocean floor. The Sunda Shelf in the vicinity of the Java Sea has relatively low relief, contains several coral reefs, and is not volcanic. The mountain system that is welded along the South China and Celebes seas of this shelf and which comprises the outer edge of the continental mass of Asia, however, is an area of strong relief and is perhaps the most active volcanic zone in the world.

The outer, or southern, side of the chain of islands from Sumatra through Java and the Lesser Sundas forms the active leading edge of the Southeast Asian landmass. It is characterized by active volcanoes, bounded on the south and west by a series of deep-sea trenches and grading off on the north, or inner, edge to swamps, lowlands, and the shallow Java Sea. This sheltered sea was formed at the close of the Pleistocene Epoch (the Pleistocene lasted from about 1,600,000 to 10,000 years ago), and there is evidence of former land bridges that facilitated the migration of plants and animals from the Asian continent.

Relief. *Islands of the Sunda Shelf.* Borneo, the third largest island in the world and the main island on the Sunda Shelf, is hilly and mountainous. Its relief, however, seldom exceeds 5,600 feet above sea level, and most of the island is below 1,000 feet. Structural trends are not as well defined as on adjacent islands, although a broad mountain system runs roughly from northeast to southwest. It includes the island's highest peak of Mount Kinabalu, which rises to 13,455 feet in Sabah, Malaysia. Indonesian Borneo, or Kalimantan, constitutes about three-quarters of the island and is mainly mountainous and forested, with coastal alluvial swamps.

The Riau Islands lie east of Sumatra. They have a granite core and can be considered as a physical extension of the Malay Peninsula. Like Malaysia, they are rich in tin, which is recovered both on land and offshore, mainly off the islands of Bangka, Billiton, and Singkep.

Sumatra is flanked on its outer (western) edge by a string of nonvolcanic islands, including Simeulue, Nias, and the Mentawai group, none of which is densely populated. Sumatra runs from northwest to southeast for a length of more than 1,000 miles and a maximum width, including offshore islands, of about 325 miles and is bisected by the equator. The island divides into four main physical regions: the narrow coastal plain along the west; the Barisan Mountains, which extend the length of the island close to its western edge and include 10 active volcanoes; an inner nonvolcanic zone of low hills grading down toward the stable platform of the Asian mainland; and the broad alluvial lowland, as much as 150 miles wide and no more than 100 feet above sea level, that constitutes the eastern half of the island.

Sub-regions of Sumatra

Much of eastern Sumatra is a low-lying swampy forest that is difficult to penetrate, seriously impeding the development of the inland area. The mountain watershed is close to the west coast, and much of the soil cover in the hills and lowlands is built up by debris from the volcanoes. There are a number of beautiful lakes in Sumatra, the most famous of which is Lake Toba, which lies in the north at an elevation of 2,953 feet above sea level and covers some 440 square miles.

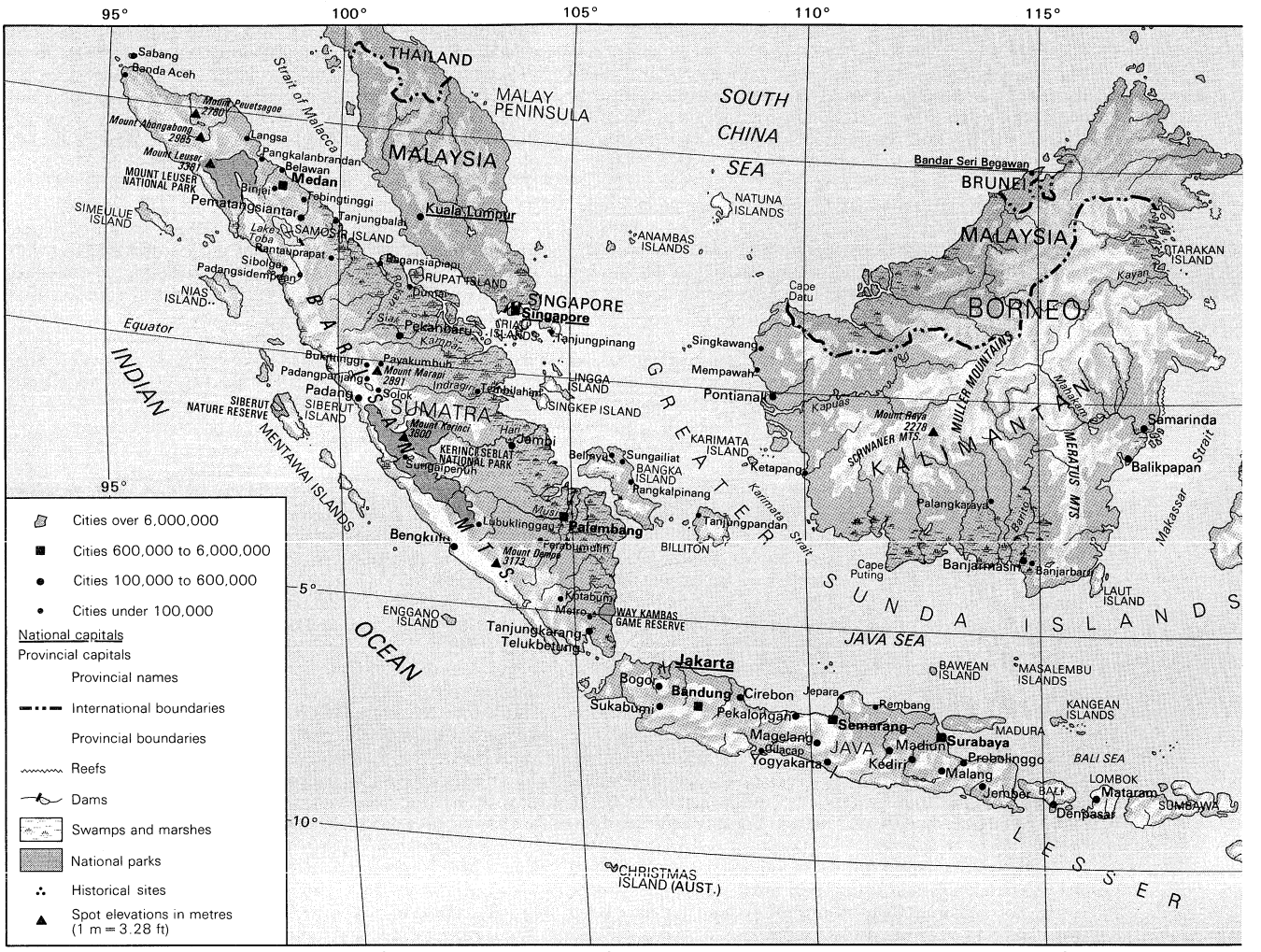
Java is some 660 miles long and has a maximum width of about 125 miles. Its physical divisions are not as distinct as those of Sumatra, because the continental shelf drops sharply to the Indian Ocean in the southern part of the island. Java can be divided into five latitudinal physiographic regions. A series of limestone platforms extends along the southern coast; in some areas they form an eroded karst region (*i.e.,* marked by sinks interspersed with abrupt ridges, irregular rocks, caverns, and underground streams) that makes communication and habitation difficult. A mountain belt just to the north is partially composed of sediments derived from eroded volcanoes and includes a number of alluvial basins that are heavily cultivated, as around Bandung and Garut. The belt of volcanoes through the centre of the island constitutes the third region and contains more than 55 active cones and 22 volcanoes with a geologically recent history of eruption. A northern alluvial belt, the fourth region, spreads across the Sunda Shelf toward the sea and is extended by delta formations, particularly during volcanic activity. There are deep inland extensions of the alluvial region, which in central Java cut through to the southern coast. Finally, there is a second limestone platform area along the northern coast of Madura and the adjacent section of eastern Java.

The many islands east of Java are much smaller, less densely populated, and less developed than Java. The physiography in Bali and Lombok is similar to that of eastern Java. The Lesser Sunda chain continues through Sumbawa and Flores islands, narrowing progressively until it appears on a map as a spine of volcanic islands that loops northeast into the Banda Islands. The same volcanic system may be considered to reappear in northern Celebes. Sumba and Timor form an outer (southerly) fringe of nonvolcanic islands, which resemble those off the western edge of the Sunda Shelf near Sumatra.

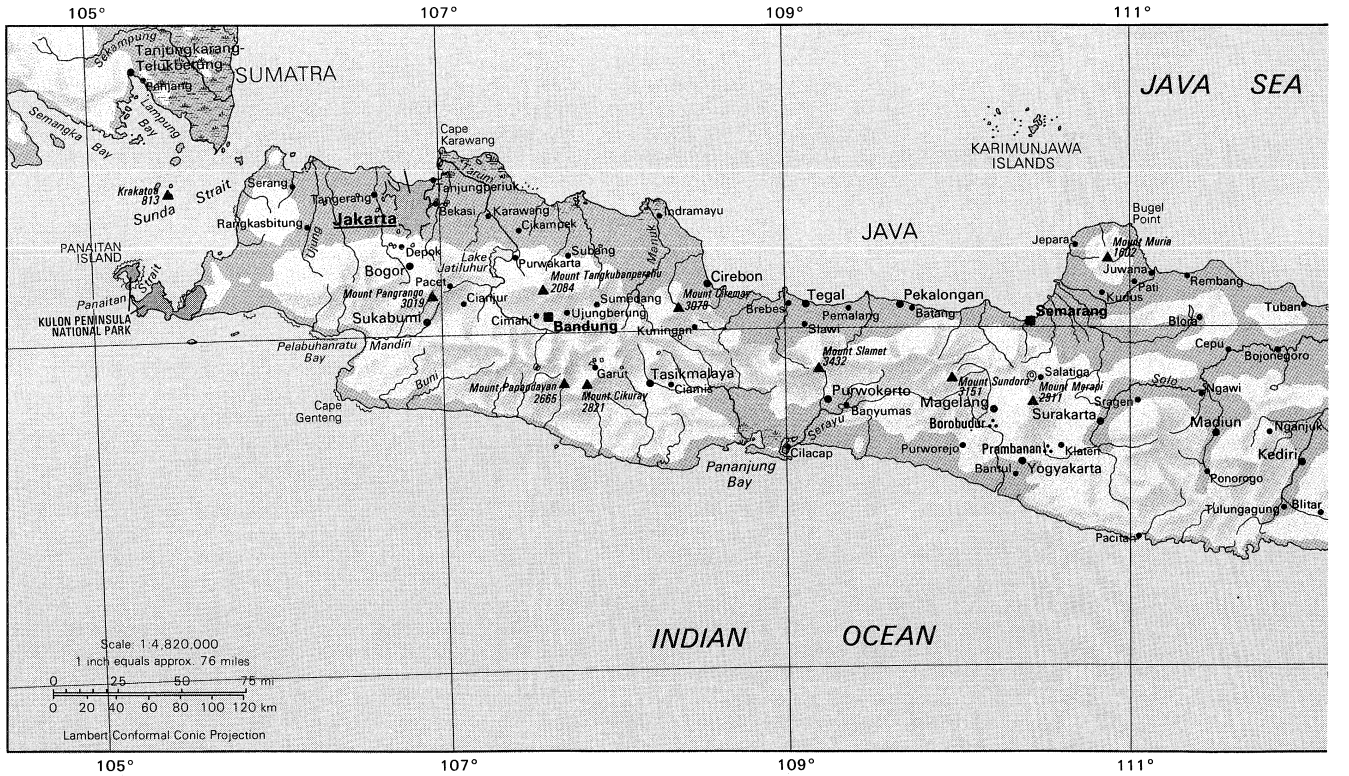
Islands of the Sahul Shelf. The islands of the Sahul Shelf appear to have a physiographic structure similar to those of the Sunda Shelf. They include the northern Moluccas (see below) and New Guinea. The western portion of New Guinea constitutes the Indonesian province of Irian Jaya, which accounts for more than 20 percent of the total area of Indonesia but less than 1 percent of the country's population. Irian Jaya is a remote region with a spectacular and varied landscape. Mangrove swamps seal much of the southern and western coastline, while the Maoke Mountains—including Jaya Peak, which at 16,502 feet (5,030 metres) is the highest point in Indonesia—form a natural barrier across the central area. There is a narrow coastal plain in the north. More than two-thirds of the province is heavily forested.

Irian Jaya

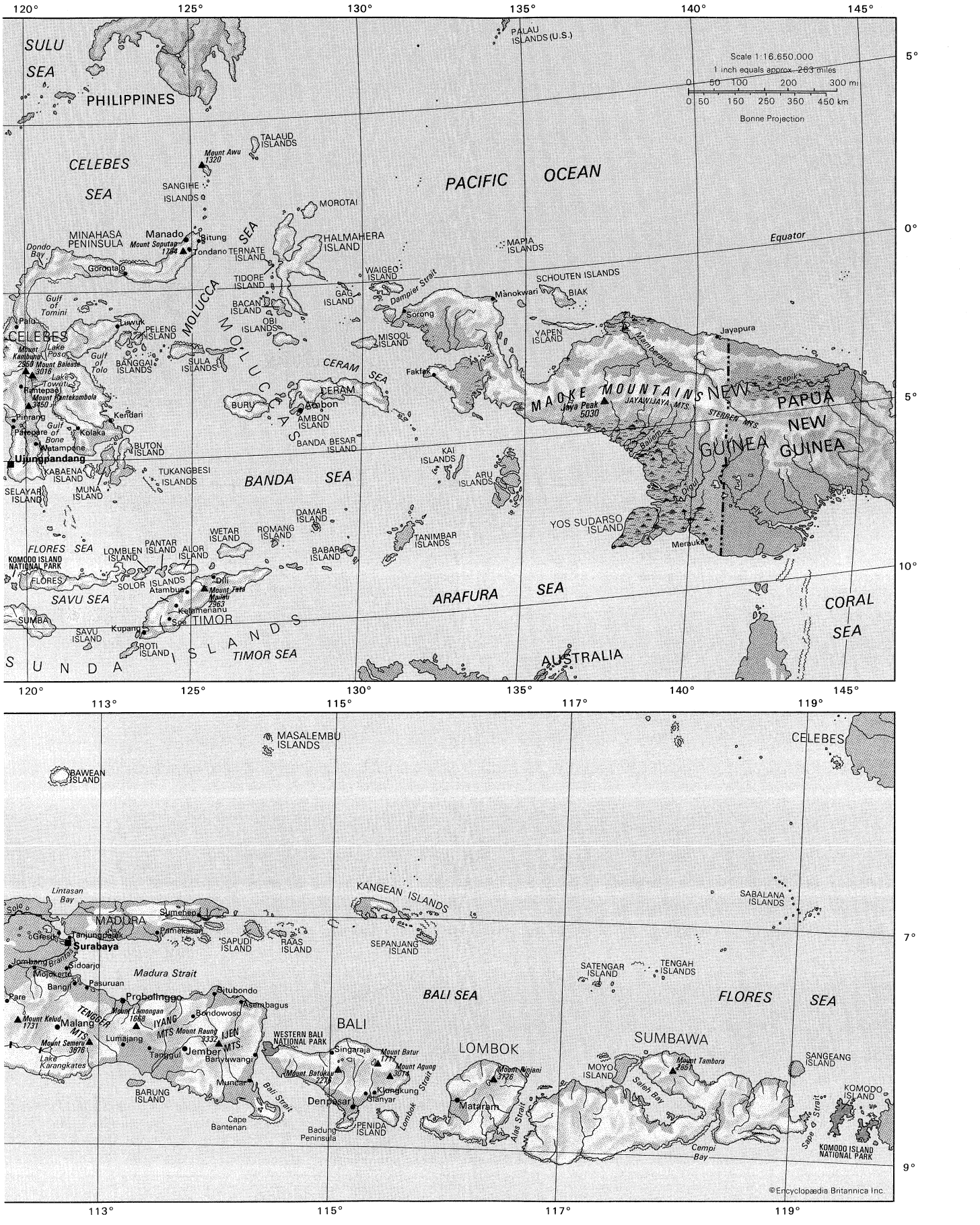
Celebes and the Moluccas. Celebes shows some evidence of being squeezed between the conflicting forces of the more stable surrounding masses of the Sunda and Sahul shelves. Its complex shape somewhat resembles a capital K, with an extremely long peninsula running northeast from its north-south backbone. There are, therefore, three large gulfs: Tomini (or Gorontalo) on the north, Tolo on the east, and Bone on the south. The coastline is long in relation to the size of the island. Celebes consists of ranges



95°
100°
105°
110°
115°
5°
0°
10°
THAILAND
MALAY PENINSULA
MALAYSIA
SOUTH CHINA SEA
Strait of Malacca
Sabang
Banda Aceh
Mount Peuetsagoe 2780
Mount Abongabong 2985
Langsa
Pangkalanbrandan
Belawan
Medan
Binjai
Tebingtinggi
Pematangsiantar
Tanjungbalai
Kuala Lumpur
MOUNT LEUSER NATIONAL PARK
SIMEULUE ISLAND
Lake Toba
SAMOSIR ISLAND
Rantauprapat
Sibolga
Padangsidempuan
NIAS ISLAND
Bagansiapiapi
RUPAT ISLAND
Duntai
SINGAPORE
Singapore
Pekanbaru
RIAU ISLANDS
Tanjungpinang
Equator
INDIAN OCEAN
Bukittinggi
Payakumbuh
Mount Marapi 2891
Padangpanjang
Padang
Solok
Indragiri
Tembilahan
INGGA ISLAND
SINGKEP ISLAND
SIBERUT NATURE RESERVE
SIBERUT ISLAND
SUMATRA
Mount Kerinci 3800
Hari
Jambi
KERINCI SEBLAT NATIONAL PARK
Sungaipenuh
MENTAWAI ISLANDS
Bengkulu
Lubuklinggau
Musi
Palembang
Ferabumulih
Mount Dempo 3173
Kotabumi
Metro
WAY KAMBAS GAME RESERVE
ENGGANO ISLAND
Tanjungkarang-Telukbetung
Jakarta
Bogor
Bandung
Cirebon
Sukabumi
Pekalongan
Jepara
Rembang
Semarang
Magelang
Cilacap
Yogyakarta
JAVA
Madiun
Kediri
Surabaya
Probolinggo
Malang
Jember
BALI
Denpasar
LOMBOK
Mataram
SUMBAWA
BALI SEA
MADURA
KANGEAN ISLANDS
BAWEAN ISLAND
MASALEMBU ISLANDS
JAVA SEA
SUNDA ISLANDS
LESSER
CHRISTMAS ISLAND (AUST.)
GREATER SUNDA ISLANDS
Karimata Strait
Tanjungpandan
BILLITON
KARIMATA ISLAND
Ketapang
Singkawang
Mempawah
Pontianak
Kapuas
Cape Datu
NATUNA ISLANDS
ANAMBAS ISLANDS
Bandar Seri Begawan
BRUNEI
MALAYSIA
BORNEO
TARAKAN ISLAND
Kayan
Mount Raya 2278
SCHWANER MTS.
MULLER MOUNTAINS
KALIMANTAN
Mahakam
MERATUS MTS.
Samarinda
Balikpapan
Makassar Strait
Palangkaraya
Barito
Cape Puting
Banjarmasin
Banjarbaru
LAUT ISLAND
Cities over 6,000,000
Cities 600,000 to 6,000,000
Cities 100,000 to 600,000
Cities under 100,000
National capitals
Provincial capitals
Provincial names
International boundaries
Provincial boundaries
Reefs
Dams
Swamps and marshes
National parks
Historical sites
Spot elevations in metres (1 m = 3.28 ft)



105°
107°
109°
111°
7°
9°
Tanjungkarang-Telukbetung
Bahjang
SUMATRA
Sekampung
Lampung Bay
Semangka Bay
Krakatoa 813
Sunda Strait
Serang
Tangerang
Rangkasbitung
Jakarta
Cape Karawang
Tanjungperiuk
Bekasi
Karawang
Cikampek
Indramayu
Manuk
Subang
Purwakarta
Mount Tangkubanperahu 2084
Cirebon
Mount Ciremai 3078
Depok
Bogor
Pacet
Mount Pangrango 3019
Cianjur
Sukabumi
Cimahi
Bandung
Sumedang
Ujungberung
Kuningan
Brebes
Tegal
Pemalang
Pekalongan
Batang
Semarang
Slawi
Mount Slamet 3432
Purwokerto
Banyumas
Magelang
Borobudur
Mount Sundoro 3151
Mount Merapi 2911
Salatiga
Surakarta
Sragen
Solo
Ngawi
Madiun
Nganjuk
Kediri
Ponorogo
Tulungagung
Blitar
Pacitan
Bugel Point
Jepara
Mount Muria 1602
Juwana
Pati
Kudus
Rembang
Tuban
Blora
Cepu
Bojonegoro
Cilacap
Serayu
Purworejo
Prambanan
Klaten
Bantul
Yogyakarta
Pananjung Bay
Cape Genteng
Garut
Tasikmalaya
Ciamis
Mount Papandayan 2665
Mount Cikuray 2821
Buni
Mandiri
Pelabuhanratu Bay
PANAITAN ISLAND
Panaitan Strait
KULON PENINSULA NATIONAL PARK
JAVA SEA
KARIMUNJAWA ISLANDS
JAVA
INDIAN OCEAN
Scale: 1:4,820,000
1 inch equals approx. 76 miles
0 25 50 75 mi
0 20 40 60 80 100 120 km
Lambert Conformal Conic Projection


120°
125°
130°
135°
140°
145°
5°
0°
10°
Scale 1:16,650,000
1 inch equals approx. 263 miles
0 50 100 200 300 mi
0 50 150 250 350 450 km
Bonne Projection
SULU SEA
PHILIPPINES
PALAU ISLANDS (U.S.)
CELEBES SEA
PACIFIC OCEAN
Equator
TALAUD ISLANDS
Mount Awu 1320
SANGIHE ISLANDS
MOROTAI
MINAHASA PENINSULA
Manado
Mount Soputan 1784
Bitung
Tondano
TERNATE ISLAND
HALMAHERA ISLAND
MAPIA ISLANDS
Dondo Bay
Gorontalo
TIDORE ISLAND
WAIGEO ISLAND
SCHOUTEN ISLANDS
Dampier Strait
Manokwari
BIAK
Gulf of Tomini
BACAN ISLAND
GAG ISLAND
Sorong
Palu
Luwuk
PELENG ISLAND
OBI ISLANDS
MISOOL ISLAND
YAPEN ISLAND
Jayapura
CELEBES
MOLUCCA SEA
Mount Kambuno 2959
Lake Poso
Mount Balease 3018
Gulf of Tolo
BANGGAI ISLANDS
SULA ISLANDS
CERAM SEA
Fakfak
MAOKE MOUNTAINS
JAYAWIJAYA MTS.
Jaya Peak 5030
STERREN MTS
NEW GUINEA
PAPUA NEW GUINEA
Lake Towuti
Rantepao
Mount Rantekombola 3450
BURU
CERAM
Ambon
AMBON ISLAND
BANDA BESAR ISLAND
Pinrang
Kendari
Gulf of Bone
Parepare
Kolaka
Watampone
BUTON ISLAND
Ujungpandang
KABAENA ISLAND
TUKANGBESI ISLANDS
KAI ISLANDS
ARU ISLANDS
BANDA SEA
SELAYAR ISLAND
MUNA ISLAND
DAMAR ISLAND
YOS SUDARSO ISLAND
Merauke
WETAR ISLAND
ROMANG ISLAND
TANIMBAR ISLANDS
FLORES SEA
PANTAR ISLAND
LOMBLEN ISLAND
ALOR ISLAND
BABAR ISLAND
KOMODO ISLAND NATIONAL PARK
FLORES
SOLOR ISLANDS
Atambua
Dili
Mount Tatamailau 2963
SAVU SEA
ARAFURA SEA
CORAL SEA
Kefamenanu
SUMBA
TIMOR
Soe
Kupang
SAVU ISLAND
ROTI ISLAND
SUNDA ISLANDS
TIMOR SEA
AUSTRALIA
113°
115°
117°
119°
7°
9°
MASALEMBU ISLANDS
CELEBES
BAWEAN ISLAND
Lintasan Bay
KANGEAN ISLANDS
SABALANA ISLANDS
MADURA
Sumenep
Pamekasan
Gresik
Tanjungperak
Surabaya
SAPUDI ISLAND
RAAS ISLAND
SEPANJANG ISLAND
Jombang
Sidoarjo
Mojokerto
Madura Strait
SATENGAR ISLAND
TENGAH ISLANDS
Bangil
Pasuruan
Bitubondo
BALI SEA
FLORES SEA
Rare
Probolinggo
Asembagus
Mount Lamongan 1668
Bondowoso
Mount Kelud 1731
Malang
IYANG MTS
Mount Raung 3332
IJEN MTS
TENGGER MTS
Mount Semeru 3676
Lumajang
Tanggul
Jember
Baryuwangi
WESTERN BALI NATIONAL PARK
BALI
LOMBOK
SUMBAWA
Singaraja
Mount Batur 1717
Mount Agung 3014
Mount Tambora 2851
SANGEANG ISLAND
MOYO ISLAND
Lake Karangkates
Mount Betukau 2274
Klungkung
Gianyar
Mount Rinjani 3726
Saleh Bay
KOMODO ISLAND
BARUNG ISLAND
Muncar
Bali Strait
Denpasar
Mataram
Alas Strait
Lombok Strait
Cape Bantenan
Badung Peninsula
PENIDA ISLAND
Cempi Bay
Sape Strait
KOMODO ISLAND NATIONAL PARK
©Encyclopædia Britannica Inc.

MAP INDEX

Political subdivisions

Aceh . . . . . 4 00 N 97 00 E
Bali . . . . . 8 30 S 115 00 E
Bengkulu . . . . . 3 48 S 102 16 E
Irian Jaya . . . . . 5 00 S 138 00 E
Jakarta Raya . . . . 6 10 S 106 45 E
Jambi . . . . . 1 30 S 103 00 E
Jawa Barat . . . . . 7 00 S 107 00 E
Jawa Tengah . . . 7 30 S 110 00 E
Jawa Timur . . . . . 8 00 S 113 00 E
Kalimantan Barat . 0 00 110 30 E
Kalimantan Selatan . . . . . 2 30 S 115 30 E
Kalimantan Tengah . . . . . 2 00 S 113 30 E
Kalimantan Timur . . . . . 1 30 S 116 30 E
Lampung . . . . . 5 00 S 105 00 E
Maluku . . . . . 4 00 S 128 00 E
Nusa Tenggara Barat . . . . . 8 50 S 117 30 E
Nusa Tenggara Timur . . . . . 9 30 S 122 00 E
Riau . . . . . 1 00 N 102 00 E
Sulawesi Selatan . 4 00 S 120 00 E
Sulawesi Tengah . 1 00 S 121 00 E
Sulawesi Tenggara . . . . . 4 00 S 122 30 E
Sulawesi Utara . . 1 00 N 123 00 E
Sumatera Barat . . 1 00 S 100 30 E
Sumatera Selatan . . . . . 3 30 S 104 00 E
Sumatera Utara . . 2 00 N 99 00 E
Timor Timur . . . . 8 35 S 126 00 E
Yogyakarta . . . . . 7 45 S 110 30 E

Cities and towns

Ambon . . . . . 3 43 S 128 12 E
Asembagus . . . . . 7 45 S 114 14 E
Atambua . . . . . 9 07 S 124 54 E
Bagansiapiapi . . . 2 09 N 100 49 E
Balikpapan . . . . . 1 17 S 116 50 E
Banda Aceh (Kuta Raja) . . . . 5 34 N 95 20 E
Bandung . . . . . 6 54 S 107 36 E
Bangil . . . . . 7 36 S 112 47 E
Banjarbaru . . . . . 3 25 S 114 50 E
Banjarmasin . . . . 3 20 S 114 35 E
Bantul . . . . . 7 54 S 110 20 E
Banyumas . . . . . 7 31 S 109 17 E
Banyuwangi . . . . 8 12 S 114 21 E
Batang . . . . . 6 55 S 109 45 E
Bekasi . . . . . 6 14 S 106 59 E
Belawan . . . . . 3 47 N 98 41 E
Belinyu . . . . . 1 38 S 105 46 E
Bengkulu . . . . . 3 48 S 102 16 E
Binjai . . . . . 3 36 N 98 30 E
Bitung . . . . . 1 27 N 125 11 E
Blitar . . . . . 8 06 S 112 09 E
Blora . . . . . 6 57 S 111 25 E
Bogor . . . . . 6 35 S 106 47 E
Bojonegoro . . . . . 7 09 S 111 52 E
Bondowoso . . . . . 7 55 S 113 49 E
Brebes . . . . . 6 53 S 109 03 E
Bukittinggi . . . . . 0 19 S 100 22 E
Cepu . . . . . 7 09 S 111 35 E
Ciamis . . . . . 7 20 S 108 21 E
Cianjur . . . . . 6 49 S 107 08 E
Cikampek . . . . . 6 24 S 107 27 E
Cilacap . . . . . 7 44 S 109 00 E
Cimahi . . . . . 6 53 S 107 32 E
Cirebon . . . . . 6 44 S 108 34 E
Denpasar . . . . . 8 39 S 115 13 E
Depok . . . . . 6 24 S 106 50 E
Dili . . . . . 8 33 S 125 34 E
Dumai . . . . . 1 41 N 101 27 E
Fakfak . . . . . 2 55 S 132 18 E
Garut . . . . . 7 13 S 107 54 E
Gianyar . . . . . 8 32 S 115 20 E
Gorontalo . . . . . 0 33 N 123 03 E
Gresik . . . . . 7 09 S 112 38 E
Indramayu . . . . . 6 20 S 108 19 E
Jakarta . . . . . 6 10 S 106 48 E
Jambi . . . . . 1 36 S 103 37 E
Jayapura . . . . . 2 32 S 140 42 E
Jember . . . . . 8 10 S 113 42 E
Jepara (Japara) . . 6 35 S 110 39 E
Jombang . . . . . 7 33 S 112 14 E
Juwana . . . . . 6 42 S 111 09 E
Karawang . . . . . 6 19 S 107 17 E
Kediri . . . . . 7 49 S 112 01 E
Kefamenanu . . . . 9 27 S 124 29 E
Kendari . . . . . 3 57 S 122 35 E
Ketapang . . . . . 1 52 S 109 59 E
Klaten . . . . . 7 42 S 110 35 E
Klungkung . . . . . 8 32 S 115 24 E
Kolaka . . . . . 4 03 S 121 36 E
Kotabumi . . . . . 4 50 S 104 54 E
Kudus . . . . . 6 48 S 110 50 E
Kuningan . . . . . 6 59 S 108 29 E
Kupang . . . . . 10 10 S 123 35 E
Kuta Raja, see Banda Aceh
Langsa . . . . . 4 28 N 97 58 E
Lubuklinggau . . . 3 10 S 102 52 E
Lumajang . . . . . 8 08 S 113 13 E
Luwuk . . . . . 0 56 S 122 47 E
Madiun . . . . . 7 37 S 111 31 E
Magelang . . . . . 7 28 S 110 13 E
Malang . . . . . 7 59 S 112 37 E
Mampawah, see Mempawah
Manado . . . . . 1 29 N 124 51 E
Manokwari . . . . . 0 52 S 134 05 E
Mataram . . . . . 8 35 S 116 07 E
Medan . . . . . 3 35 N 98 40 E
Mempawah (Mampawah) . . . 0 22 N 108 58 E
Merauke . . . . . 8 28 S 140 20 E
Metro . . . . . 5 05 S 105 20 E
Mojokerto . . . . . 7 28 S 112 26 E
Muncar . . . . . 8 26 S 114 20 E
Nganjuk . . . . . 7 36 S 111 55 E
Ngawi . . . . . 7 24 S 111 26 E
Pacet . . . . . 6 45 S 107 03 E
Pacitan . . . . . 8 12 S 111 07 E
Padang . . . . . 0 57 S 100 21 E
Padangpanjang . . 0 27 S 100 25 E
Padang-sidempuan . . . . 1 22 N 99 16 E
Pakanbaru, see Pekanbaru
Palangkaraya . . . 2 16 S 113 56 E
Palembang . . . . . 2 55 S 104 45 E
Palu . . . . . 0 53 S 119 53 E
Pamekasan . . . . . 7 10 S 113 28 E
Pangkalan-brandan . . . . . 4 01 N 98 17 E
Pangkalpinang . . 2 08 S 106 08 E
Panjang . . . . . 5 29 S 105 18 E
Pare . . . . . 7 46 S 112 11 E
Parepare . . . . . 4 01 S 119 38 E
Pasuruan . . . . . 7 38 S 112 54 E
Pati . . . . . 6 45 S 111 01 E
Payakumbuh . . . . 0 14 S 100 38 E
Pekalongan . . . . . 6 53 S 109 40 E
Pekanbaru (Pakanbaru) . . . . 0 32 N 101 27 E
Pemalang . . . . . 6 54 S 109 22 E
Pematangsiantar . 2 57 N 99 03 E
Perabumulih . . . . 3 27 S 104 15 E
Pinrang . . . . . 3 48 S 119 38 E
Ponorogo . . . . . 7 52 S 111 27 E
Pontianak . . . . . 0 02 S 109 20 E
Probolinggo . . . . 7 45 S 113 13 E
Purwakarta . . . . . 6 34 S 107 26 E
Purwokerto . . . . . 7 25 S 109 14 E
Purworejo . . . . . 7 43 S 110 01 E
Rangkasbitung . . 6 21 S 106 15 E
Rantauprapat (Rantaupa-rapat) . . . . . 2 06 N 99 50 E
Rantepao . . . . . 2 59 S 119 54 E
Rembang . . . . . 6 42 S 111 20 E
Sabang . . . . . 5 55 N 95 19 E
Salatiga . . . . . 7 19 S 110 30 E
Samarinda . . . . . 0 30 S 117 09 E
Semarang . . . . . 6 58 S 110 25 E
Serang . . . . . 6 07 S 106 09 E
Sibolga . . . . . 1 45 N 98 48 E
Sidoarjo . . . . . 7 27 S 112 43 E
Singaraja . . . . . 8 07 S 115 06 E
Singkawang . . . . 0 54 N 109 00 E
Situbondo . . . . . 7 42 S 114 00 E
Slawi . . . . . 6 59 S 109 08 E
Soe . . . . . 9 52 S 124 17 E
Solok . . . . . 0 48 S 100 39 E
Sorong . . . . . 0 53 S 131 15 E
Sragen . . . . . 7 26 S 111 02 E
Subang . . . . . 6 34 S 107 45 E
Sukabumi . . . . . 6 55 S 106 56 E
Sumedang . . . . . 6 52 S 107 55 E
Sumenep . . . . . 7 01 S 113 52 E
Sungailiat . . . . . 1 51 S 106 08 E
Sungaipenuh . . . . 2 05 S 101 23 E
Surabaya . . . . . 7 15 S 112 45 E
Surakarta . . . . . 7 35 S 110 50 E
Tangerang . . . . . 6 11 S 106 37 E
Tanggul . . . . . 8 10 S 113 26 E
Tanjungbalai . . . . 2 58 N 99 48 E
Tanjungkarang-Telukbetung . . . 5 27 S 105 16 E
Tanjungpandan . . 2 45 S 107 39 E
Tanjungperak . . . 7 12 S 112 43 E
Tanjungperiuk . . . 6 06 S 106 53 E
Tanjungpinang . . 0 55 N 104 27 E
Tasikmalaya . . . . 7 20 S 108 12 E
Tebingtinggi . . . . 3 20 N 99 09 E
Tegal . . . . . 6 52 S 109 08 E
Tembilahan . . . . . 0 19 S 103 09 E
Tondano . . . . . 1 19 N 124 54 E
Tuban . . . . . 6 54 S 112 03 E
Tulungagung . . . . 8 04 S 111 54 E
Ujungberung . . . . 6 55 S 107 42 E
Ujungpandang . . 5 07 S 119 24 E
Watampone . . . . 4 32 S 120 20 E
Yogyakarta . . . . . 7 48 S 110 22 E

Physical features and points of interest

Abongabong, Mount . . . . . 4 15 N 96 48 E
Agung, Mount, *volcano* . . . . . 8 21 S 115 30 E
Alas Strait . . . . . 8 40 S 116 40 E
Aior Island . . . . . 8 15 S 124 45 E
Ambon (Amboina) Island . . . . . 3 40 S 128 10 E
Anambas Islands . 3 00 N 106 00 E
Arafura Sea . . . . . 9 00 S 133 00 E
Aru Islands . . . . . 6 00 S 134 30 E
Awu, Mount, *volcano* . . . . . 3 43 N 125 27 E
Babar Island . . . . 7 55 S 129 45 E
Bacan Island . . . . 0 35 S 127 30 E
Badung Peninsula . . . . . 8 49 S 115 10 E
Balease, Mount . . 2 24 S 120 33 E
Bali, *island* . . . . . 8 20 S 115 00 E
Bali Sea . . . . . 7 45 S 115 30 E
Bali Strait . . . . . 8 18 S 114 25 E
Baliem, *river* . . . . 4 25 S 138 59 E
Banda Besar Island . . . . . 4 34 S 129 55 E
Banda Sea . . . . . 5 00 S 128 00 E
Banggai Islands . 1 30 S 123 15 E
Bangka Island . . . 2 15 S 106 00 E
Bantenan, Cape . 8 47 S 114 33 E
Barisan Mountains . . . . . 3 00 S 102 15 E
Barito, *river* . . . . . 3 32 S 114 29 E
Barung Island . . . 8 28 S 113 20 E
Batukau, Mount . . 8 20 S 115 05 E
Batur, Mount . . . . 8 14 S 115 23 E
Bawean Island . . . 5 46 S 112 40 E
Biak, *island* . . . . . 1 00 S 136 00 E
Billiton (Belitung), *island* . . . . . 2 50 S 107 56 E
Bone, Gulf of . . . . 4 00 S 120 40 E
Borneo, *island* . . . 1 00 N 114 00 E
Borobudur, *historical site* . . . 7 36 S 110 12 E
Brantas, *river* . . . 7 28 S 112 25 E
Bugel Point . . . . . 6 26 S 111 03 E
Buni, *river* . . . . . 7 26 S 106 47 E
Buru, *island* . . . . 3 24 S 126 40 E
Buton (Butung) Island . . . . . 5 00 S 122 55 E
Carstensz, see Jaya Peak
Celebes (Sulawesi), *island* . . . . . 2 00 S 121 00 E
Celebes Sea . . . . 3 00 N 122 00 E
Cempi Bay . . . . . 8 44 S 118 25 E
Ceram (Seram), *island* . . . . . 3 00 S 129 00 E
Ceram (Seram) Sea . . . . . 2 30 S 128 00 E
Cikuray, Mount . . 7 19 S 107 52 E
Ciremay, Mount, *volcano* . . . . . 6 54 S 108 24 E
Damar Island . . . . 1 00 S 128 24 E
Dampier Strait . . . 0 40 S 130 40 E
Datu, Cape . . . . . 2 05 N 109 39 E
Dempo, Mount, *volcano* . . . . . 4 02 S 103 09 E
Digul, *river* . . . . . 7 07 S 138 42 E
Dondo Bay . . . . . 0 55 N 120 30 E
Enggano Island . . 5 24 S 102 16 E
Ewab, see Kai Islands
Flores, *island* . . . 8 30 S 121 00 E
Flores Sea . . . . . 8 00 S 121 00 E
Gag Island . . . . . 0 27 S 129 52 E
Genteng, Cape . . 7 23 S 106 24 E
Gorontalo, see Tomini, Gulf of
Greater Sunda Islands . . . . . 0 00 110 00 E
Halmahera Island . . . . . 1 00 N 128 00 E
Hari, *river* . . . . . 1 16 S 104 05 E
Ijen Mountains . . . 8 05 S 114 15 E
Indian Ocean . . . 5 00 S 100 00 E
Indragiri (Inderagiri), *river* . . . . . 0 22 S 103 26 E
Iyang Mountains . 7 58 S 113 38 E
Jatiluhur, Lake, *reservoir* . . . . . 6 35 S 107 20 E
Java (Jawa), *island* . . . . . 7 30 S 110 00 E
Java (Jawa) Sea . 5 00 S 110 00 E
Jaya (Carstensz) Peak . . . . . 4 05 S 137 11 E
Jayawijaya Mountains . . . . . 4 30 S 139 30 E
Kabaena Island . . 5 15 S 121 55 E
Kai (Ewab) Islands . . . . . 5 35 S 132 45 E
Kalimantan, *region* . . . . . 1 00 S 114 00 E
Kambuno, Mount . 2 20 S 120 04 E
Kampar, *river* . . . 0 16 N 101 41 E
Kangean Islands . 6 55 S 115 30 E
Kapuas, *river* . . . . 0 25 S 109 40 E
Karangkates, Lake, *reservoir* . . 8 12 S 112 30 E
Karawang, Cape . 5 56 S 107 00 E
Karimata Island . . 1 36 S 108 55 E
Karimata Strait . . 2 05 S 108 40 E
Karimunjawa Islands . . . . . 5 50 S 110 25 E
Kayan, *river* . . . . . 2 55 N 117 35 E
Kelud, Mount, *volcano* . . . . . 7 56 S 112 18 E
Kerinci, Mount, *volcano* . . . . . 1 42 S 101 16 E
Kerinci Seblat National Park . . . 3 00 S 102 00 E
Komodo Island . . 8 36 S 119 30 E
Komodo Island National Park . . 8 35 S 119 28 E
Krakatoa, *volcano* . . . . . 6 07 S 105 24 E
Kulon Peninsula National Park . . 6 45 S 105 30 E
Lamongan, Mount, *volcano* . 7 58 S 113 20 E
Lampung Bay . . . 5 40 S 105 20 E
Laut Island . . . . . 3 40 S 116 10 E
Lesser Sunda (Nusa Tenggara) Islands . . . . . 9 00 S 120 00 E
Leuser, Mount, *volcano* . . . . . 3 45 N 97 11 E
Leuser, Mount, National Park . . 3 30 N 97 00 E
Lingga Island . . . 0 12 S 104 35 E
Lintasan Bay . . . . 6 54 S 112 50 E
Lomblen Island . . 8 25 S 123 30 E
Lombok, *island* . . 8 45 S 116 30 E
Lombok Strait . . . 8 30 S 115 50 E
Madura, *island* . . 7 00 S 113 20 E
Madura Strait . . . 7 25 S 113 25 E
Mahakam, *river* . . 0 35 S 117 17 E
Makassar Strait . . 2 00 S 117 30 E
Malacca, Strait of . . . . . 2 30 N 101 20 E
Maluku, see Moluccas
Maluku, see Molucca Sea
Mamberamo, *river* . . . . . 1 26 S 137 53 E
Mandiri, *river* . . . . 7 02 S 106 32 E
Manuk, *river* . . . . 6 14 S 108 13 E
Maoke Mountains . . . . 4 00 S 138 00 E
Mapia Islands . . . 0 50 N 134 20 E
Marapi, Mount, *volcano* . . . . . 0 23 S 100 28 E

Masalembu Islands ..... 0 50 S 114 26 E
Mentawai Islands 2 00 S 99 30 E
Merapi, Mount, *volcano* ..... 7 33 S 110 26 E
Meratus Mountains ..... 2 45 S 115 40 E
Minahasa Peninsula ..... 1 00 N 124 35 E
Misool Island .... 1 52 S 130 10 E
Molucca (Maluku) Sea ..... 0 00 125 00 E
Moluccas (Maluku), *islands* ..... 2 00 S 128 00 E
Morotai, *island* .. 2 20 N 128 25 E
Moyo Island .... 8 15 S 117 34 E
Muller Mountains . 0 40 N 113 50 E
Muna Island .... 5 00 S 122 30 E
Muria, Mount .... 6 36 S 110 53 E
Musi, *river* ..... 2 20 S 104 56 E
Natuna Islands .. 4 00 N 108 15 E
New Guinea .... 5 00 S 140 00 E
Nias Island ..... 1 05 N 97 35 E
Nusa Tenggara, see Lesser Sunda Islands
Obi Islands ..... 1 30 S 127 45 E
Pacific Ocean ... 3 00 N 136 00 E
Panaitan Island .. 6 36 S 105 12 E
Panaitan Strait .. 6 40 S 105 16 E
Pananjung Bay .. 7 45 S 108 37 E
Pangrango, Mount ..... 6 46 S 106 57 E
Pantar Island .... 8 25 S 124 07 E
Papandayan, Mount, *volcano* ..... 7 20 S 107 44 E
Pelabuhanratu Bay ..... 7 03 S 106 27 E
Peleng Island ... 1 20 S 123 10 E
Penida Island ... 8 44 S 115 32 E
Peuetsagoe, Mount, *volcano* . 4 55 N 96 20 E
Prambanan, *historical site* ... 7 45 S 110 28 E
Poso, Lake ..... 1 52 S 120 35 E
Puting, Cape .... 3 31 S 111 46 E
Raas Island ..... 7 09 S 114 32 E
Ramelau, see Tata Mailau, Mount
Rantekombola, Mount ..... 3 21 S 120 01 E
Raung, Mount, *volcano* ..... 8 08 S 114 03 E
Raya, Mount .... 0 40 S 112 41 E
Riau Islands .... 1 00 N 104 30 E
Rinjani, Mount, *volcano* ..... 8 24 S 116 28 E
Rokan, *river* ..... 2 00 N 100 52 E
Romang Island .. 7 35 S 127 26 E
Roti Island ..... 10 45 S 123 10 E
Rupat Island .... 1 50 N 101 35 E
Sabalana Islands . 6 45 S 118 50 E
Saleh Bay ..... 8 34 S 117 57 E
Samosir Island .. 2 35 N 98 50 E
Sangeang Island . 8 12 S 119 04 E
Sangihe Islands .. 3 00 N 125 30 E
Sape Strait ..... 8 39 S 119 18 E
Sapudi Island ... 7 06 S 114 20 E
Satengar Island .. 7 31 S 117 17 E
Savu (Sawu) Island ..... 10 30 S 121 54 E
Savu (Sawu) Sea . 9 40 S 122 00 E
Schouten Islands . 0 55 S 135 55 E
Schwaner Mountains ..... 0 40 S 112 40 E
Sekampung, *river* ..... 5 36 S 105 50 E
Selayar Island ... 6 05 S 120 30 E
Semangka Bay .. 5 36 S 104 42 E
Semeru, Mount, *volcano* ..... 8 06 S 112 55 E
Sepanjang Island . 7 10 S 115 50 E
Sepik, *river* ..... 3 51 S 144 32 E
Seram, see Ceram
Serayu, *river* .... 7 41 S 109 06 E
Siak, *river* ..... 1 13 N 102 09 E
Siberut Island ... 1 20 S 98 55 E
Siberut Nature Reserve ..... 1 30 S 98 30 E
Simeulue Island .. 2 35 N 96 05 E
Singkep Island .. 0 30 S 104 25 E
Slamet, Mount, *volcano* ..... 7 14 S 109 12 E
Solo, *river* ..... 6 47 S 112 33 E
Solor Islands .... 8 25 S 123 30 E
Soputan, Mount, *volcano* ..... 1 07 N 124 45 E
South China Sea . 5 00 N 110 00 E
Sterren (Star) Mountains ..... 5 00 S 141 00 E
Sula Islands .... 1 52 S 125 22 E
Sulawesi, see Celebes
Sumatra, *island* .. 0 00 102 00 E
Sumba, *island* ... 10 00 S 120 00 E
Sumbawa, *island* . 8 40 S 118 00 E
Sunda Strait .... 6 00 S 105 45 E
Sundoro, Mount, *volcano* ..... 7 17 S 109 59 E
Talaud Islands ... 4 20 N 126 50 E
Tambora, Mount, *volcano* ..... 8 14 S 117 55 E
Tangkubanperahu, Mount, *volcano* . 6 44 S 107 36 E
Tanimbar Islands . 7 30 S 131 30 E
Tarakan Island ... 3 21 N 117 36 E
Tarum, *river* .... 5 59 S 107 03 E
Tata Mailau (Ramelau), Mount ..... 8 55 S 125 30 E
Tengah Islands .. 7 30 S 117 30 E
Tengger Mountains ..... 7 55 S 112 55 E
Ternate Island ... 0 48 N 127 20 E
Tidore Island .... 0 41 N 127 24 E
Timor, *island* .... 8 50 S 126 00 E
Timor Sea ..... 10 30 S 126 00 E
Toba, Lake ..... 2 35 N 98 50 E
Tolo, Gulf of .... 2 00 S 122 30 E
Tomini (Gorontalo), Gulf of ..... 0 20 S 121 00 E
Towuti, Lake .... 2 45 S 121 32 E
Tukangbesi Islands ..... 5 40 S 123 50 E
Ujung, *river* ..... 5 56 S 106 15 E
Waigeo Island ... 0 14 S 130 45 E
Way Kambas Game Reserve . 5 00 S 105 45 E
Western Bali National Park .. 8 08 S 114 30 E
Wetar Island .... 7 48 S 126 18 E
Yapen Island .... 1 45 S 136 15 E
Yos Sudarso Island ..... 7 50 S 138 30 E

of mountains cut by deep rift valleys, many of which contain lakes. The island is fringed by coral reefs and is bordered by deep sea troughs in the south. Its northeastern arm, the Minahasa Peninsula, is volcanic and structurally different from the rest of the island, which is composed of a complex of igneous and metamorphic rocks.

The Moluccas consist of a group of approximately 1,000 islands with a combined area that is two-thirds the size of Java. Halmahera Island, at 6,865 square miles, is the largest of the group, followed by Ceram and Buru. The Moluccas lie in the same geologically unstable zone as Celebes, although the northern islands are associated more with the Sahul Shelf. Halmahera Island in the north and the islands of the Banda Sea are volcanic, the latter group also experiencing a high frequency of earthquakes. Most of the northern and central Moluccas have dense vegetation and rugged mountainous interiors where elevations often exceed 3,000 feet. The Moluccas were synonymous with the "Spice Islands," and especially during the 16th and 17th centuries the islands of Ternate, Tidore, Ambon, and Banda Besan were a source of cloves, nutmeg, and mace.

Volcanoes. There are some 220 active volcanoes in Indonesia and many hundreds that are considered extinct. They run in a crescent-shaped line along the outer margin of the country through Sumatra and Java as far as Flores and then loop north through the Banda Sea to a junction with the volcanoes of northern Celebes.

Volcanoes play a major role in soil development and enrichment, and there is a strong relationship between agricultural development, density of population, and location of volcanoes. Of the 80 volcanoes that have a recent history of eruption, the greatest concentration (22) is on Java. The greatest population densities occur in such areas as the region south and east of Merapi volcano in central Java, where the soil is enriched by volcanic ash and debris. The same pattern occurs on Bali and in northern Sumatra, where the rich soils are directly related to flows from volcanic eruptions. The chemical composition of the ejecta, however, is not uniform throughout the country; in central and southern Sumatra, for example, the ash and lava are largely acidic, resulting in relatively poor soils.

Volcanic eruptions

Volcanic eruptions are by no means uncommon. Merapi, which rises to 9,550 feet near Yogyakarta (Jogjakarta), erupts frequently—often accompanied by extensive destruction of roads, fields, and villages but always to the great benefit of the soil. Kelud (5,679 feet), near Kediri in eastern Java, can be particularly devastating, because its large crater lake is thrown out during eruption, causing great mudflows (lahars) that rush down into the plains and sweep all before them.

Perhaps the best-known volcano is Krakatoa (Krakatau), which erupted from the seabed of the Sunda Strait between Sumatra and Java in 1883. All life on the surrounding island group was destroyed. The eruptions caused tidal waves throughout Southeast Asia, killing tens of thousands of people, and ash clouds that circled the Earth decreased solar radiation and produced spectacular sunsets for more than a year. Another major incident occurred in 1963 when Mount Agung in Bali erupted after being dormant for more than 140 years, causing more than 1,500 deaths and heavy property destruction.

Drainage. Because of its insularity, Indonesia has no large rivers comparable to those on the Asian mainland. Indonesian rivers generally are relatively short and flow from interior mountains to the sea. The Kapuas (710 miles long), Barito (560 miles), and Mahakam (480 miles) rivers of Kalimantan are among the longest, but shifting sandbars across their mouths reduce their importance for transportation. Irian Jaya, most of which receives heavy rainfall, is drained by a number of large rivers, including the Baliem, the Mamberamo, and the Digul.

The seas surrounding Indonesia, however, must also be viewed as a dominant physical feature, having an important effect on climate, transportation, and the development of culture. They serve both as channels of communication and as barriers protecting distinctive features. The shallow seas between many of the islands are a significant resource not only of offshore oil, gas, and minerals but also of fisheries and food.

Soils. Indonesia illustrates the relation between climate and source rock in the formation of soils. The rocks on Java are primarily andesitic volcanics (dark gray rocks consisting essentially of the minerals oligoclase or feldspar), while rhyolites (the acidic lava form of granite) are dominant on Sumatra, granites on the Riau Islands, granites and sediments in Kalimantan, and sediments in Irian Jaya. The resulting soils in humid regions are mainly lateritic (containing iron oxides and aluminum hydroxide) and of varying fertility depending on the source rock; they include heavy black or gray-black margalite soils and limestone soils. Black soils occur in regions with a distinct dry season, and highly localized soils include the fertile ando

soils, which developed on the andesitic volcanic sediments of the northeastern coast of Sumatra.

In general, the perpetual high temperatures and heavy precipitation throughout much of Indonesia has led to rapid erosion and deep chemical weathering and leaching, which usually results in impoverished soil. In areas covered with tropical rain forests, such as Kalimantan, the soils are protected by the forest cycle; as plants die, rapid decomposition releases nutrients that are reabsorbed by new vegetation growth. Although such soils support a luxurious growth, they cannot support a large agricultural population because clearing of the forest breaks the cycle and can lead to accelerated soil deterioration.

Importance of volcanic ash

Minerals that are leached from the soil are replaced by alluvial deposition from rivers, as in some parts of Kalimantan, or by deposition in impounded water or rice terraces. Most valuable in Indonesia is the volcanic ash, which is transported by wind to be deposited as a layer of homogeneous, fresh inorganic material over wide areas; it is also carried as suspended material in streams and irrigation channels. The best soils are derived from or enriched by basic andesitic volcanic material, the ejecta from rhyolitic volcanoes being less rich. The andesitic volcanoes occur in Sumatra, Java, and western Celebes.

Climate. The climate of Indonesia is controlled by its island structure and position astride the equator, which assure high, even temperatures, and by its location between the two landmasses of Asia and Australia, which strongly influences the monsoonal rainfall patterns. Temperatures are uniformly high and are a function of elevation rather than latitude. They are highest along the coast, where mean annual temperatures range from 74° to 88° F (23° to 31° C) and are moderated considerably above 2,000 feet. The only area high enough to receive snow is the Maoke Mountains of Irian Jaya. The diurnal difference of temperature in Jakarta is at least five times as great as the difference between high and low temperatures of January and July; the highest temperature ever recorded in Jakarta was 99° F (37° C), and the lowest was 66° F (19° C).

Rainfall is more varied in extremes and distribution. Most of Indonesia receives heavy precipitation throughout the year, the greatest amounts occurring from December to March. From central Java eastward toward Australia, however, the dry season (June to October) is progressively more pronounced; in Timor and Sumba there is little rain during these months. The highest amount of rainfall occurs in the mountainous regions of Sumatra, Kalimantan, Celebes, and Irian Jaya, where more than 120 inches (3,000 millimetres) falls annually. The rest of Kalimantan, Sumatra, and Irian Jaya; western and central Java; and much of Celebes and the Moluccas average at least 80 inches of rainfall per year. Eastern Java, Bali, southern and central Celebes, and Timor generally receive between 60 and 80 inches; while the Lesser Sunda Islands that are closest to Australia have only 40 to 60 inches.

The absolute daily maximum of rainfall can be extremely high, with a number of stations recording between 20 and 28 inches. Local variations, caused in large part by geographic features, are great. Jakarta, for example, near sea level, has a mean annual rainfall of 70 inches, while Bogor, which is 30 miles south toward the mountains, at an elevation of about 790 feet, records nearly 170 inches of rainfall.

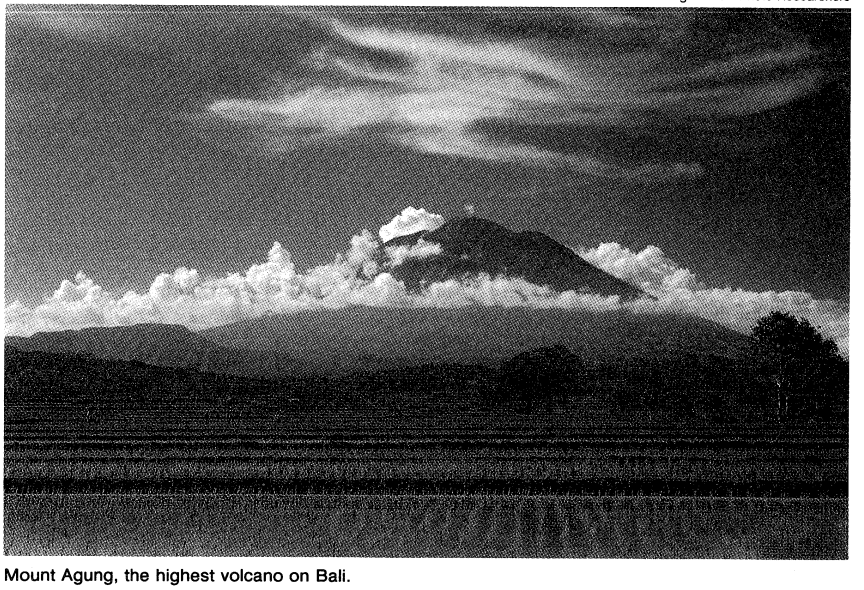
Seasonal variations are caused by monsoonal Asian air drifts and the convergence of tropical air masses from both north and south of the equator along an intertropical front of low pressure. The monsoon pattern in any given part of the archipelago depends on location either north or south of the equator, proximity to Australia or mainland Asia, and the position of the intertropical front. During December, January, and February, the west monsoon, reflecting Asian influence, brings heavy rain to southern Sumatra, Java, and the Lesser Sunda Islands. In June, July, and August, these areas are affected by the east monsoon, which brings dry air from Australia. Only the Lesser Sunda Islands and eastern Java have a well-developed dry season, which increases in length toward Australia. By the time the east monsoon has crossed the equator—becoming the southwest monsoon of the Northern Hemisphere—its winds have become humid and a source of rain. Sumatra and Kalimantan, which are located close to the equator and far from Australia, have no dry season, although precipitation tends to be slightly lower during July and August. Strong cyclones and typhoons, which normally occur in higher latitudes, are absent in Indonesia, but afternoon thunderstorms are common.

Tropical rain forests

Plant life. Much of Indonesia is still covered with the natural growth of tropical rain forest, of which only a fraction is primeval forest. Irian Jaya and eastern Kalimantan are mostly forest-covered, while on the densely populated islands of Java and Bali a much smaller part of the land is covered with forest.

The vegetation is similar to that of the Philippines, Malaysia, and Papua New Guinea. There are some 40,000 species of flowering plants, including 5,000 species of orchids, as well as the monster flower (*Rafflesia arnoldii*), which is the world's largest flower. There are more than 3,000 tree species, including durian, sandalwood, illupi nut, valuable timber varieties such as teak and ironwood, and rattans. More than 6,000 species are exploited for economic purposes, either directly or indirectly. Major timber operations are located in Kalimantan, where the trees are not differentiated but are referred to as "broad-leaved species"; they include, for example, meranti, which has a soft, lightweight, pinkish to darkish red wood. Teak, which is also broad-leaved, comes mainly from Java.

© George Love—Photo Researchers



Mount Agung, the highest volcano on Bali.

The most important vegetation type is the mixed lowland and hill tropical rain forest, which occurs below 5,000 feet. It is characterized by a large number of species, including high-canopied and buttressed trees and woody, thick-stemmed lianas (climbing plants). Epiphytes (plants that derive nourishment from the air and usually live on another plant) such as orchids and ferns, saprophytes (plants that live on dead or decaying matter), and parasites are well developed. Above 5,000 feet this forest gives way to temperate upland forest dominated by oak, laurel, tea, and magnolia species. Another typical feature of Indonesian vegetation is the mangrove forest, characterized by the formation of stilt- or prop-rooted trees, which grow only in salty or brackish water along muddy shores. Mangrove swamps are extensively developed along the shallow seas on eastern Sumatra, southern Kalimantan, and southeastern Irian Jaya.

Animal life. Indonesia is located in the transitional zone between two of the world's major faunal realms: the Megagaean of Asia in the west and the Notogaean of Australia in the east. The boundary of these realms, called Wallace's Line, runs between Borneo and Celebes in the north and Bali and Lombok in the south. To the west, the Asian animal community includes such mammals as the rhinoceros, orangutan, tapir, tiger, and elephant. Animals related to Australian fauna include birds such as cockatoos, bowerbirds, and birds of paradise, as well as marsupials such as bandicoots (small insectivorous, herbivorous, marsupial mammals) and cuscuses (brightly coloured, woolly-haired arboreal marsupials).

Endemic animal species

Many of the islands contain endemic species. Peacocks are found only in Java, and the related kuau only in Sumatra. The mountain goat lives only on the rugged slopes of the Barisan Mountains of Sumatra. A unique species of proboscis monkey lives only in Kalimantan, and the babirusa (a large hoglike quadruped) and anoa (a small wild ox with nearly straight horns) can be found only in Celebes. A giant lizard—the prehistoric Komodo dragon, which attains a length of 12 feet—occurs on two small islands, Rinca and Komodo, between Sumbawa and Flores.

Some of these endemic species have become exceedingly rare. Most of the remaining single-horned Sumatran rhinoceroses, for example, are now restricted to the Kulon Peninsula National Park on the western tip of Java. This nearly extinct species is one of the most highly protected forms of wildlife. Another such endangered species is the orangutan, which is native to Borneo and Sumatra. Orangutan rehabilitation centres have been established on the edge of the Mount Leuser National Park in northern Sumatra and in a game preserve on Cape Puting in southern Kalimantan in an effort to prevent the capture and slaughter of the animals and to train those that have been held captive to return to the wild.

Indonesia has an enormous and varied insect life that includes many unusual species. Examples include giant walkingsticks that can attain eight inches in length, walking leaves, huge atlas beetles, elegant luna moths, and beautiful bird-wing swallowtails.

Traditional regions. The island structure of Indonesia provides natural boundaries that strongly influence the traditional regions. On the smaller islands administrative and traditional regions generally overlap, while on the larger islands the administrative structure generally harmonizes with traditional and cultural divisions.

Central Java

The three provinces of Java, and the special autonomous district of Yogyakarta, represent the most populous and culturally sophisticated part of the country and illustrate the overlap of traditional and administrative regions. Jawa Tengah (Central Java) is the geographic, cultural, and historical focus of the island. Yogyakarta is a stronghold of Javanese culture and maintains a traditional sultan ruler; a variety of Javanese historic monuments (*candi*) are located in the vicinity, including the great Buddhist monument of Borobudur. Jawa Tengah and Jawa Timur (East Java) are densely settled, largely agricultural areas dominated by ethnic Javanese. The major port, trading, and industrial city of Surabaya is Jawa Timur's largest city and the second largest in Indonesia; Semarang assumes these functions in Jawa Tengah. Jawa Barat (West Java) is the land of the Sundanese, who are related to but quite distinct from the Javanese in language and tradition. In addition, Java contains the strongly contrasting metropolitan district (*daerah khusus ibukota*) of Jakarta, which does not coincide with cultural or traditional patterns.

The provinces on Sumatra (seven provinces plus the special autonomous district of Aceh) also have a degree of traditional integrity. Located in the north, Aceh is a region of strict Muslims who were long noted for their resistance to European influence. Sumatera Utara (North Sumatra), with its major city of Medan, includes a rich plantation area along the coast and, at a higher elevation, the region inhabited by the Batak, who were largely isolated until the 19th century. Riau and Jambi, oil-rich provinces in the east, are inhabited by Malay people and are the areas in which the Indonesian language developed. Sumatera Barat (West Sumatra) is the region of the Minangkabau people, who are devout Muslims and are noted for their matrilineal society in which property is passed on through the female line. Lampung and Bengkulu provinces, in the southern part of Sumatra, are the sites of major oil fields.

On Kalimantan, as elsewhere, there is a contrast between the coast and the inland region. Chinese and Malays dominate the coastal regions, while a variety of Dayak tribes live in the interior, where they engage in traditional shifting cultivation. A similar pattern applies in Irian Jaya and on many of the other islands where maritime-trading communities have been developed along the coast and agrarian, noncommercial societies, with strongly developed and highly localized customs, inhabit the interior.

East of Java each island or group of islands has maintained its own distinct character, in many cases strongly influenced by religion. Bali—with its long tradition of Hindu and Buddhist influences rooted in animism—is quite different in character and customs from any other part of Indonesia. Lombok is partly Hindu, but the influence of Islām is stronger. Sumbawa is Muslim, Flores is largely Roman Catholic, and Timor contains strong Protestant groups. These variations also prevail in Celebes and the Moluccas, where the Makasarese and Buginese of southern Celebes are Muslims noted as seafarers and shipbuilders, while the Menadonese in northern Celebes (on the Minahasa Peninsula) and the Ambonese are Christian.

The barriers of the mountains and the sea have protected the character and traditions of many groups. Away from the major cities and areas of dense population, there are significant variations from one valley to the next and almost from one village to the next. In many cases the tribal groups—the Toraja of Celebes, the Dayak of Kalimantan, and the Gayo, Lampung, and Batak peoples of Sumatra—were relatively untouched by outside influences until the arrival of Christian missionaries during the 19th century, and even today they display a wide range of cultures.

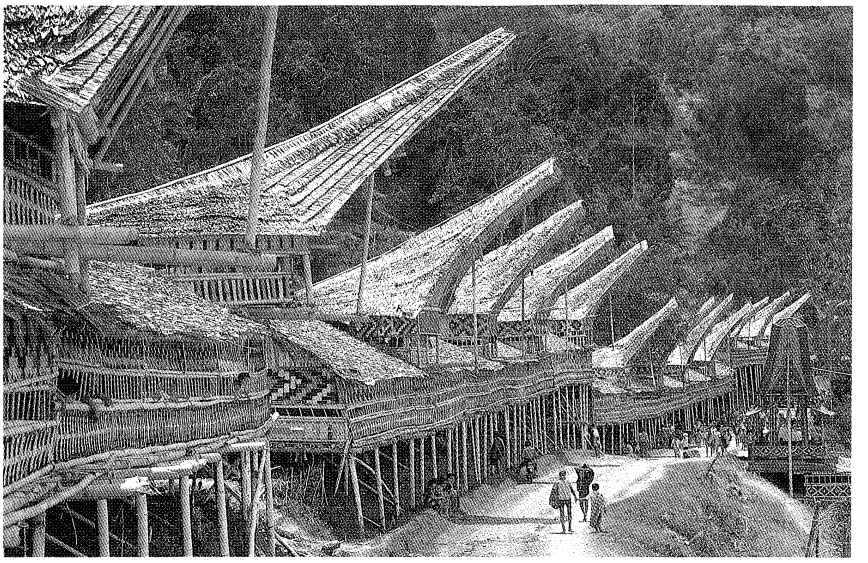
Settlement patterns. *Rural settlement.* Indonesia is primarily a rural country, with the majority of the population living in agricultural areas. About two-thirds of the total population inhabits the islands of Java, Madura, and Bali, which have a highly sophisticated rural structure that is based largely on wet-rice cultivation. Other areas of high rural population are found in parts of Sumatra and Celebes. Most of the rest of the country is sparsely settled by tribal groups who engage in subsistence agriculture.

The Javanese village

The Javanese rural village is the most common settlement. Paddy rice fields cover the flat land and in many places rise up the hillsides in terraces. Scattered throughout are clusters of coconut, palm, and fruit trees, which indicate the location of the villages. In the heavily populated areas of central and eastern Java, there are thousands of such villages, some of which have sizable populations.

The people of each village form a group that is homogeneous both in economic condition and in social interest and outlook. In many cases, and particularly in irrigated areas, there is much mutual exchange of labour. Overpopulation in the densely populated areas has led to the decrease in size of the average farm and to an increase in the numbers of the landless rural group, who work mainly as farm labourers or sharecroppers.

Each village has a stream or well as its source of water, a mosque and elementary school, and a network of



Temporary housing in a Toraja village, constructed for guests and relatives attending a funeral, Celebes.

© George Holton—Photo Researchers

swept-earth paths. There is little formal commercial activity; goods are obtained from peddlers and small shops (*warung*s) or from the market towns, which often are also local government centres. Houses are well separated and are normally of frame and bamboo with roofs of red tile or coconut fibres; houses constructed of locally made bricks are increasingly common, especially among the wealthier families. Goats, chickens, banana and papaya trees, and a host of small children are characteristic of village life.

Rural structure varies considerably from region to region. Balinese villages are clusters of walled family complexes with Hindu shrines, public buildings, and larger temples. The Batak villages around Lake Toba in northern Sumatra, Minangkabau villages in western Sumatra, Toraja villages in southern Celebes, and Dayak longhouses in Kalimantan each have their characteristic structure and building style. The social pattern also varies considerably. On Java the pattern is very simple, with few organized groupings above the level of the household, while on neighbouring Bali there are strong groups related to working, dancing, and other functions, many of which are associated with Hindu festivals.

The rural mode of life is controlled by the growing season and by the productivity of the land. It ranges from the seminomadic shifting cultivation of tribal groups, through cassava and sago gardening, smallholder plantations, and irrigated rice farming, to large mechanized plantations. In some cases these activities are combined with some form of cottage industry. Most Indonesians are small-scale, independent peasant farmers who operate at or near the subsistence level and sell some produce but usually do not accumulate substantial capital. In general, the villages are small, independent, and largely self-sufficient.

Urban settlement. The overall level of urbanization in Indonesia is low in relation to other countries that are at a comparable stage of economic growth. This can be explained in part by the phenomenon of nonpermanent, or "circular," migration on Java and elsewhere: individuals from rural families live and work in the cities, but they return to their homes at least once every six months. Although there is some regional variation in urban growth rates, generally cities of every population size are growing rapidly.

Few of the cities, except for Jakarta, Surabaya, and Medan, have the heterogeneity of a true urban centre. Instead, they are the economic, governmental, cultural, and social centres for highly populated and distinct regions. The growth of the cities has not been accompanied by a parallel growth of industry, and the outlook of much of the urban population is still rural. Large parts of the population, even in Jakarta, live in replicas of rural villages, or kampongs, characterized by rural customs. Urban dwellers generally are better off than their rural counterparts, and urban services have gradually improved; but the availability of adequate housing, potable water, and public transportation services have remained critical concerns.

Four of Indonesia's five largest cities—Jakarta, Surabaya, Bandung, and Semarang—are on Java; the other, Medan, is located on Sumatra. These five cities may be considered as metropolitan areas rather than large provincial towns since they contain the major government, financial, and business offices. Other large cities, such as Padang, Yogyakarta, and Palembang, are centres of provincial government and of local trade and have limited international ties or contact with foreigners.

The character of the cities

The cities have individual characters. Jakarta, as the capital and centre of finance, has fine government buildings, broad avenues and large fountains, and an increasing number of multistory hotels and office buildings. Surabaya is a major port and industrial city. Bandung, a former resort area and military centre, has much light industry and a number of universities. Semarang is the administrative capital and commercial hub of central Java. Yogyakarta, which was the capital of the revolutionary government between 1946 and 1949, is the seat of the ruling family of the sultan of Yogyakarta and remains the centre of Javanese culture. It also is the site of a major university, Gadja (Gajah) Mada, and of schools of art, traditional dance, and music and is the centre of the batik cloth industry. In Sumatra, Medan and its port city of Belawan constitute the commercial centre for the rich northern agricultural districts, and Palembang, Sumatra's second largest city, is the port for the oil industry and for a variety of other industries in the south.

The social composition of Indonesia's cities is highly diverse and reflects the heavy flow of migration from rural areas. The most varied of these is Jakarta: while many people may have been born or raised there, they often continue to refer to themselves in terms of their regional heritage—such as Batak, Javanese, or Minangkabau—and it is not uncommon for them to use their local languages at home.

The social and economic character of Indonesian cities is a continuing topic of study. A social hierarchy exists that is roughly composed of an elite group of government officials, military officers, and business leaders with a Western orientation; a growing middle class of civil servants, teachers, and other professionals and skilled workers who are significantly underpaid and must struggle to maintain their economic position; and a larger number of poorly educated unskilled labourers, traders, and other members of the informal economy who strongly identify with their villages and frequently move back and forth to engage in economic pursuits in both areas. This three-tiered hierar-

chy also conforms closely to an economic structure that is based on various government opportunities and on formal and informal business activities.

A transient foreign element of diplomats and company representatives plays a minor role in city structure. The permanent foreign element—mainly of Chinese, Indian, and Arab business families—is more fully integrated, but each group maintains its own contacts and patterns of life. The Indonesians gradually are developing an urban culture. This notion, perhaps more appropriately viewed as urban sophistication, is most conspicuous in Jakarta, with its strong international contacts. Since association with this international culture implies a degree of wealth, it is largely confined to the families of officials, professionals, and prominent businessmen. The lower-income groups, on the other hand, retain their basic ethnic cultures, strengthened by trips to home villages during times of harvest or during the Muslim month of Ramaḍān (a period of fasting and atonement).

Although Indonesia's social structure is decentralized, its administrative structure is highly centralized, with Jakarta the headquarters of the central government. Most taxes, including land and real estate taxes, are collected by the central government, on which city and provincial governments must depend for their revenue. Efforts have been made, however, to decentralize some government functions, particularly with respect to finance and to the management and delivery of various services.

THE PEOPLE

Diversity of the population

The Indonesian motto—"Bhinneka Tunggal Ika" ("Unity in Diversity")—is illustrated by the fact that within the Indonesian population there are more than 300 different ethnic groups and 250 distinct languages and that most of the major world religions are practiced there, in addition to a wide range of indigenous ones. Within this diversity there are certain groupings and concentrations; thus, most of the people are of Malay ancestry, speak languages that have an Austronesian (Malayo-Polynesian) base, and profess Islām. The largest of the subgroups is the Javanese, whose language is also the most dominant.

Indonesia is situated at the meeting point of two geographic races, the Asiatic in the west and the Melanesian in the east. The great majority of Indonesians are related to the peoples of East Asia, although over the centuries there has also been considerable mixing with Arabs, Indians, and Europeans. In the eastern islands, however, most of the people are of Melanesian origin.

Ethnic groups. *The western islands.* The diverse ethnic populations of western Indonesia may be grouped into three broad groups: an inland wet-rice society, coastal peoples, and tribal groups. The first group, the strongly Hinduized wet-rice growers of inland Java and Bali, make up more than two-thirds of the national population. With an ancient, highly sophisticated culture of strong social and agricultural traditions, it includes the Javanese, Sundanese, Madurese, and Balinese peoples. The second group of Islāmic coastal peoples is ethnically heterogeneous, including the Malays from Sumatra and, from southern Celebes, the Makasarese, who are found in all coastal towns but are a stronger influence outside Java. The third group, the tribal peoples, including the Toraja and Dayak, has developed in inland areas where the climate will not support wet-rice cultivation and where shifting cultivation is practiced; these various groups tend to be small and isolated and have developed a wide range of cultures.

There are also a number of other major ethnic groups that do not fit into this cultural categorization. They include the Batak and Minangkabau in Sumatra and the Minahasa in northern Celebes.

The eastern islands. Eastern Indonesia is characterized by the traditional Melanesian cultural division between coastal, or "beach," peoples and interior, or "bush," peoples. The Moluccas reflect this pattern, although their proximity to the western islands makes them a more complex ethnographic and linguistic area. The islands are populated by a number of distinct ethnic groups. Typical of the coastal peoples are the Ambonese, who live along the coasts of Ambon and neighbouring islands, including western Ceram. Many of the people living in the mountainous interior regions are called Alfurs, or Alfuros; some of these groups have been relocated to coastal areas, but—unlike the coastal peoples—they do not engage in fishing activities.

The people of Irian Jaya, the native Papuans, display much more strongly the distinction between coastal and interior groups. Those in the foothills and on the coast have affinities with other Melanesian peoples to the east and south of New Guinea. In addition, Indonesians from the western islands have mixed with indigenous peoples in the coastal trading settlements. The people of the interior, on the other hand, have been isolated and insular for a long period of time. These interior Papuans are highly distinct from other Indonesians in terms of culture, language, and history. Some groups, in fact, continue to avoid contact with the outside world and inhabit nearly inaccessible areas, where they follow a way of life that has changed little in centuries. They live in small clans, and their dialects, customs, and social structure display a degree of complexity that is not found among the indigenous coastal peoples.

The Chinese community

Nonindigenous peoples. The largest nonindigenous group is the Chinese, who account for only about 2 percent of the total population but control perhaps 75 percent of the nation's wealth. Most of the Chinese have lived in Indonesia for generations. The majority of them, the *peranakans*, do not speak Chinese, have Indonesian surnames, and through intermarrying with Indonesians have developed distinct dialects and customs. A smaller segment of the Chinese population, the *totoks*, are clearly Chinese-oriented in terms of language, religion, and custom. Of the total Chinese population, most live in the towns and cities of Java and Sumatra, where they are engaged in trade. The Chinese also form a significant fraction of the population in western Kalimantan, where many are farmers and fishermen, and in the Riau Islands, where a large number are engaged in mining.

Most of the former Dutch and Eurasian residents left Indonesia after independence. Indians, Arabs, and other Europeans are relatively unimportant in numbers, although their influence in business and other elements of Western culture is apparent in the major cities.

Languages. Most of the languages spoken in Indonesia have an Austronesian base. The major exceptions are those of Irian Jaya, where Papuan languages are used, and some of the Moluccas, where the North Halmaheran language family is found. The Austronesian language family is broken into 16 major groups within which languages are closely related though distinctly different. On Java there are three major languages—Javanese, Sundanese, and Madurese—while on Sumatra there are 15, many of which are divided into a number of distinct dialects. Within the Toraja group, a relatively small population in the interior of Celebes, there are eight languages. In eastern Indonesia each island has its own language, which is often not understood on the neighbouring islands.

The national language

The national language, Bahasa Indonesia, evolved from a Malay dialect spoken in the Riau-Jambi area of eastern Sumatra; it has much in common with other Malay dialects that have long served as regional lingua francas. Since it is a relatively simple and widely used language that was not associated with one of the dominant ethnic groups, Bahasa Indonesia has been accepted without serious question and has served as a strong force of national unification. It is now learned by all children in the schools, where the local language is the medium of instruction during the first two years and Bahasa Indonesia is used for the remaining years. In 1972 a uniform revised spelling was agreed to between Indonesia and Malaysia so that communications could be improved and literature more freely exchanged between the two countries.

Religions. The vast majority of the population professes Islām, which in most cases is strongly influenced by Hinduism, Buddhism, and older pagan and animistic beliefs. The Hindu population lives mainly on Bali, and there are Christians scattered throughout the country, more than two-thirds of whom are Protestant and the remainder mainly Roman Catholic. Most Chinese practice Buddhism

and Confucianism. In remote areas some tribal religions are practiced.

Successive layers of religious beliefs reach back to the rituals and magic of the original settlers. Remains of *Homo erectus* (originally called *Pithecanthropus,* or Java man) have been found in central Java. There are remnants of an aboriginal people in southern Sumatra. The present population is descended from a series of waves of migrants from Asia, especially southwestern China. There is little recorded history before the 7th century AD, although there is evidence of earlier cultures (Śrīvijaya) in Sumatra.

The earliest recorded Indonesian history shows extensive religious influences from India; the early Indonesian states that centred on Java or Sumatra evolved through many forms of Hinduism and Theravāda and Mahāyāna Buddhism. During the 9th century AD both Hinduism and Buddhism were practiced as court religions; Śiva and Buddha were looked upon as manifestations of the same spiritual being. The blending of the two religions continued until the 14th century, when Islām was introduced along the coasts by Muslim traders from India. Throughout all the religious changes on the court level, the common people adopted part of each new religion as an additional layer over their basic traditional and animistic beliefs. Although Islām has become the dominant religion, it includes elements of all past beliefs.

The major religions were all introduced on the coast and, except in such open areas as Java and southern Sumatra, penetrated slowly inland. Regions such as central Kalimantan and Irian Jaya, the mountains of northern Sumatra, and the interiors of other mountainous islands remained virtually untouched by outside religions. Ritualistic headhunting in Kalimantan and Celebes and cannibalism in northern Sumatra were practiced until the arrival of Christian missionaries in the late 19th century.

Indonesian variants of Islām

Islām is most strictly practiced in Aceh, western Sumatra, western Java, southeastern Kalimantan, and some of the Lesser Sunda Islands. Away from these strongholds, the people consider themselves to be Muslims, but most do not follow the full ritual of fasting and prayers. On Java only about one-third of the Muslims follow orthodox practices; they are referred to as the *santri.* Members of the old Javanese aristocracy, including a majority of white-collar workers, are termed *priyayi.* A third, syncretistic tradition, called *abangan,* is strongly influenced by traditional and ancestral spirits and is closely associated with the peasants. Ritual ceremonies, *selamatan*s, are held on all special occasions, the head of a bull is buried at the dedication of a new building, and the many rituals connected with birth, death, and marriage are carefully observed by people at all levels. (J.F.McD./T.R.L.)

Demographic trends. Indonesia is the fifth largest nation in the world in terms of population. The distribution and density of the population varies considerably from region to region; the islands of Java, Bali, and Madura account for nearly two-thirds of the total population. Overall, the population has nearly doubled since 1960, with a moderately high rate of growth, but there have been significant regional contrasts in this rate. In Java, for example, population growth has been significantly less than in the outer islands. A sharp decline in fertility rates has also been in progress throughout Indonesia. The two major factors for this decline have been an increase in the age when people marry and the rapid introduction of birth-control methods. Lower fertility has been especially conspicuous in central Java. Mortality rates have also declined substantially since the 1950s, largely because of improved health care, better dietary and nutrition practices, and improvements in housing and water quality. The rates of infant and child mortality have also declined.

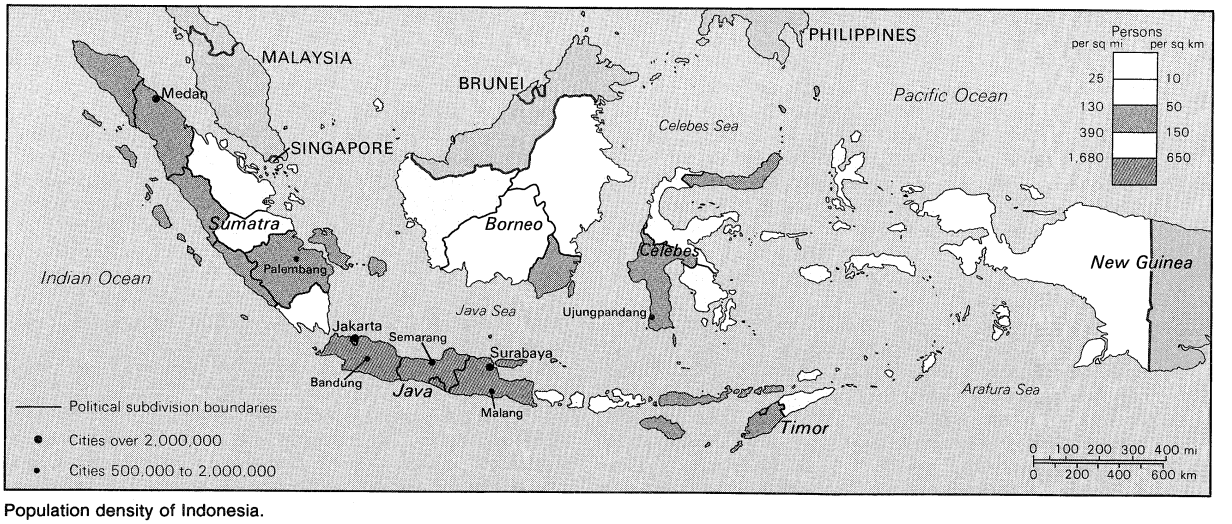
Indonesia's age structure is typical of many developing countries, with nearly 40 percent of the population under 15 years of age. The older component of the population has been increasing, but those 65 years or older account for less than 5 percent of the population. While life expectancies at birth are about 55 years for males and 58 years for females, an Indonesian who has survived the child mortality period can expect to live well beyond age 60.

Two major migration patterns have become discernible. The first involves the growing flow of people into urban areas, particularly Jakarta, which has resulted in an overall increase in the proportion of the population living in cities. Temporary, or circular, migration between rural and urban areas in connection with employment has also become common. The second pattern is of people leaving Java for the outer islands. The central government has facilitated much of this movement (called transmigration) by sponsoring a program of resettling landless Javanese in sparsely populated areas.

THE ECONOMY

Indonesia has played a modest role in the world economy since independence, and its importance has been considerably less than its size, resources, and geographic position would seem to warrant. The country is a major exporter of petroleum, natural gas, and tin. In addition, Indonesia is one of the world's main suppliers of rubber and a less significant producer of a wide range of other commodities, such as coffee, tea, tobacco, copra, spices (cloves and nutmeg), and oil-palm products. Nearly all commodity production comes from large estates. Widespread exploration for deposits of oil and other minerals has resulted in a number of large-scale projects that have contributed substantially to general development funds. The projects have, however, tended to reinforce the general position of Indonesia as a supplier of raw materials to world markets.

The corollary of the primary economy is that the country has remained a major importer of manufactured goods and of the technical skills and knowledge required for development. In spite of the substantial domestic market, there has been relatively little industrial development, and the industrial base is somewhat small, mostly concerned



Population density of Indonesia.

with mineral and forestry production and with food processing. There is little evidence of much growth in indigenous entrepreneurial activity in manufacturing. Domestic economic resources are limited, and there is a heavy dependence on inflows of foreign aid and private capital to finance large-scale development. Nonetheless, there has been a dramatic increase in the industrial sector that has caused agriculture's contribution to the national income to decline; agriculture, however, still constitutes the largest single element of the economy.

Labour

Indonesia's labour force has remained predominantly agricultural. Independent peasant farmers account for the larger part of production—notably of rice and other food crops—and a major proportion of export crops; the estate sector is responsible for the rest of production, which is mainly for export. Organized labour has been weak and still suffers from the repressions of early years in which the main trade union, the All-Indonesia Labour Federation (since 1985 the All-Indonesia Union of Workers), was tied to the Communist Party of Indonesia. Large, politically motivated associations of farmers were loosely knit and did little to promote the welfare of their members. Some Indonesian businessmen have banded together to agitate against the advantages allegedly given to nonnationals, but their organizations are of negligible importance. The government has sought to incorporate functional groups such as those of farmers and fishermen into a quasi-governmental political party, and many organizations and associations (including minority religious groups, such as the Buddhists, and women's organizations) have joined.

Economic mismanagement and the subordination of development to political ideals during the first 16 years of independence led to financial chaos and to a serious deterioration in the capital stock. With a major change of economic direction after Suharto assumed power in the mid-1960s, some measure of stability was regained, and the conditions for an orderly policy of rehabilitation and economic development were established.

Development policy

Since the mid-1960s the government has played a crucial role in development, but the private sector has become more influential. The overall goal has been the creation of a mixed economy. A series of five-year plans has emphasized the government's role in developing the economic infrastructure of the country, notably in agriculture, irrigation, transportation, and communications. Thus, the government, together with foreign aid, has remained a major force in projects where private enterprise has not been forthcoming. These have included the state-owned oil company Pertamina, fertilizer production that supports agriculture, and the cement, chemical, paper, and textile-spinning industries. The emphasis in the public sector increasingly has been on independent, self-financing state enterprises. The remainder of major investments in exploration and exploitation of natural resources or in new manufacturing industry has been left to the private sector, subject to overall government supervision and contracts. Defense, though given low priority, has been a substantial drain on the country's resources.

This consistent approach to development has carried its own risks. Its emphasis has been on soundness, and it has been heavily dependent on foreign aid and foreign technical assistance. One problem with such an orthodox emphasis on the role of private enterprise has been that it has led to an increasingly inequitable distribution of income. It has not been possible to implement the policies with sufficient force because powerful military interests and a vast inefficient bureaucracy distort the economic structure.

A major restraint on development has been a swollen and ill-paid bureaucracy. The government's ability to organize and implement development projects has also been limited by lack of experience and the unreliability and inadequacy of statistical and other information. A large inflow of foreign technical assistance has been necessary to help devise economic programs and projects, often as a condition of external capital aid. Low pay and poor working conditions have fostered corruption that has distorted development and imposed a substantial burden.

Resources. Indonesia has a large, mostly unprospected variety of mineral deposits, including those of petroleum, tin, manganese, copper, nickel, bauxite, and coal. Tin deposits are found on the islands of Bangka, Singkep, and Billiton (Belitung) and off the southwestern shore of Kalimantan. Nickel occurs on Celebes, Halmahera, and other islands of the Moluccas and in Irian Jaya. Manganese deposits are located in central Java and on Sumatra, Kalimantan, Celebes, and Timor. There are copper deposits in the Jayawijaya Mountains of Irian Jaya. The Riau Islands, Bangka and Singkep islands, and Kalimantan have bauxite, and coal occurs on Sumatra and Kalimantan. There are also deposits of iron, sulfur, gold, and silver. Large-scale coal and petroleum deposits provide ready raw material for thermal energy generation. Minerals

Mining and forestry. Mining provides about one-eighth of the gross domestic product. It employs only a limited work force, but through exports and taxation it contributes substantially to foreign-exchange earnings and development. In the past petroleum and tin were the most important, although coal, bauxite, gold, and other minerals have been mined. Indonesia now produces copper, nickel, manganese, and other commodities.

Petroleum is produced in Sumatra and Kalimantan and on small-scale offshore sites in the Java Sea. Refinery production since 1968 has been in the hands of Pertamina. Foreign oil companies, however, operate under a production-sharing formula, by which the ownership of the oil resources remains with the government of Indonesia and the foreign companies act as contractors, supplying the necessary capital.

The islands of Bangka, Billiton, and Singkep were mined for tin long before World War II. Production is dominated by state companies. Nickel is produced on Celebes, in Irian Jaya, and in the Moluccas. Bauxite is mined on the Riau Islands and in western Kalimantan and is processed at an aluminum smelter—the first in Southeast Asia—at Kualatanjung in northern Sumatra. Coal is mined in southern and western Sumatra.

Indonesia has some of the world's largest tracts of exploitable tropical forest, and since the 1960s the timber industry has grown rapidly. This rapid exploitation has caused considerable damage, however, and has prompted the Indonesian government to curtail clear-cutting of trees and to implement reforestation programs. There are several small areas of deciduous forest and plantations (mostly teak), but most of the trees are tropical hardwoods.

Agriculture and fishing. The consistent monsoon climate and almost even distribution of rainfall in Indonesia make it possible for the same types of crops to be grown throughout the country. Only about one-tenth of the total land surface, however, is devoted to agriculture. Intensive cultivation is restricted to Java, Bali, Lombok, and certain areas of Sumatra and Celebes. In Java much of the cultivated land is in rice, mainly along the northern coastal and central plains. In the drier section of eastern Java crops such as corn (maize), cassava, sweet potatoes, peanuts (groundnuts), and soybeans dominate the small farms, although such cash crops as tobacco and coffee are also grown on plantations.

Development in Sumatra and in the outer islands is less intensive and consists primarily of estate-raised cash crops. Sumatra accounts for more than half the total area under estate production, which is located mainly on the northeastern coast of the island. Around Medan there are extensive plantations producing tobacco, rubber, palm oil, kapok, tea, cloves, and coffee, none of which is native to the region. Rice, corn, and cassava are grown in the Padang area in the west and around the oil fields near Palembang in the southeast. Cash crops

There has been a shift from rice toward other less demanding basic subsistence crops, such as cassava. Rice has remained the cornerstone of peasant agriculture, however, and increased production of it has been the most important single aim of every five-year plan. Yields have been increased through various Bimingan Massal, or Bimas ("Mass Guidance"), schemes designed to promote the use of high-yielding varieties and fertilizer by increasing the availability of credit.

Rubber has remained the major commodity, although replanting has been a long-term process; and in contrast

to neighbouring Malaysia, the estates have failed to modernize to the same degree. Oil palms have been increasing in importance, however, and the ailing sugar industry has been rehabilitated to increase production for domestic consumption.

Fisheries have been developed on a small scale. With the aid of Japan, there has been an increase in both production and exports. A large part of the inland catch is from irrigation canals; marine fish account for most of the total catch.

Industry. Import substitution and support for the agricultural sector have been the two major aims of industrial policy. Import substitution, however, has been geared to commodities such as food, textiles, fertilizers, and cement, and no attempt at a massive, broadly based industrialization policy has been attempted.

The largest industries, many of which are state-owned, are those that process agricultural and mineral products. The estate groups and companies normally control their own facilities, and Pertamina controls petroleum refining. A significant proportion of manufacturing, however, is done by medium- and small-scale privately owned enterprises, much of which consists of cottage and home industries that supply consumer goods. Much of the small-scale industry is owned by the Chinese community. These small-scale workshops manufacture such consumer goods and general products as furniture, household equipment, textiles, and printed matter. The main centre of private industry is western Java, although considerable development has also taken place in Jakarta.

The textile industry

The major industry based on imported raw materials is the textile industry. The spinning mills are largely either state-owned or in the hands of foreign concerns, while the weaving and finishing factories, which are centred in Bandung, are largely small-scale and privately owned by local entrepreneurs. Batik production—an Indonesian method of hand-printing textiles—is concentrated in central Java. Although production of batik remains a major cottage industry, there are also a number of larger-scale operations.

Finance. Generally, the aims of the government's credit and fiscal policies have been to provide the conditions for private incentive within the context of financial orthodoxy. Subsidized credit and interest rates, however, have been used in accordance with general government priorities. Consumption and trade credit have been generated largely within the trading system, but there is also a range of private banking and moneylending facilities. Apart from these, the private financial sector has been weak and has played only a modest role in mobilizing domestic resources. A strong effort has been made to deregulate the financial system, and foreign investment and aid, subject to development conditions, have been welcomed. The foreign-exchange system in Indonesia has been greatly simplified, and incentives have been provided for foreign investment.

Major public (government-owned) banks include Bank Indonesia, the central bank; Bank Ekspor Impor Indonesia, the export-import bank; Bank Rakyat Indonesia, which specializes in rural credit; Bank Negara Indonesia, which specializes in industrial credit; Bank Bumi Daya, for the estates and forestry; Bank Pembangunan Indonesia, the development bank; Bank Dagang Negara, the state foreign-exchange bank; and Bank Tabungan Negara, the state savings bank. Each bank is diversified and operates independently. Foreign banks also operate in Indonesia. Nonbanking financial institutions are restricted. Indonesia has also opened a stock exchange to attract investment and capital.

Commerce and trade. A complex and reasonably well-developed commercial sector has been formed, based on the marketing and exporting of agricultural produce and on supplying consumer goods and services to the domestic market. It has been dominated by the Chinese community, although indigenous participation and unofficial army activity at the lower levels have grown.

Exports and imports

In general, Indonesia exports raw materials in return for manufactured goods. Oil, timber products, and rubber are Indonesia's major exports. Other agricultural exports include coffee, copra, tea, pepper, tobacco, and oil-palm products. In the long term non-oil minerals will probably be of increasing significance. In addition to manufactures, imports consist largely of capital equipment, consumer goods, and foodstuffs and also of such raw materials as cotton, rice, and other food crops.

Indonesia's most important trading partners are Japan, the United States, and Singapore. Other major trading partners include Germany, South Korea, Taiwan, The Netherlands, and Australia. (E.I.U./T.R.L.)

Transportation. Because Indonesia is an island country, sea transport plays a key role in the movement of raw materials and agricultural products from their source to markets. The physical nature of the country has favoured the development of strong sea links for freight and strong air links for passengers. Many parts of Indonesia have not been adequately served by the transport network, a factor that has critically hampered economic development.

Roads and railways. On the islands road transport is dominant. The only islands with adequate land transportation networks are Java, Madura, and Sumatra. On Java, where existing rail and road organization is good and is capable of being expanded to meet growing needs, emphasis has been placed on road transport because of the short distances involved. Road traffic can be increased rapidly as roads are improved and trucks imported, while railroad traffic continues to be hampered by the poor condition of rolling stock. There is, however, an important role for railways for both freight and passengers, because the high population density of Java places a limitation on new road construction.

The Indonesian State Railway (Perusahaan Jawatan Kereta Api, or PJKA) operates on Java, Madura, and Sumatra. Geographic features and commodity composition have reduced the competitive position of the railroad. There has been little demand for long-distance bulk movement, normally the mainstay of railroad operation but in Indonesia handled by shipping.

Most of the paved roads are on Java and Madura, where the network of highways is adequate to meet traffic needs in most areas. Much of the remaining paved mileage is on Sumatra and Bali. Western and central Kalimantan and Celebes have some good roads, but on Irian Jaya and the Moluccas there are few road interconnections between major settled areas.

Water and air transport. Most of the major population centres are close to the sea, where they can be served and linked by coastal and interisland shipping services. The adjacent seas are relatively calm because Indonesia is outside the belt of typhoons and high winds, and, even where docking facilities are not available, it is usually possible for ships to anchor and discharge and load from lighters and other craft.

There are numerous ports, some of which have facilities and water depths that allow ships of more than 500 tons to load and unload at quayside. The major dry-cargo ports are Tanjungperiuk (Tanjungpriok), the outport of Jakarta; Tanjungperak, the outport of Surabaya; and Belawan, the outport of Medan. Palembang in southern Sumatra is the major petroleum port. Other major ports include Semarang and Cirebon on Java, Telukbayur (the outport of Padang) on Sumatra, Manado on Celebes, Ambon in the Moluccas, Jayapura on Irian Jaya, and Banjarmasin on Kalimantan. The main commodities carried are petroleum and petroleum products, rice, copra, cement, flour, fertilizer, coconut oil, salt, rubber, asphalt, logs, and lumber.

International air services

International air services are confined to Jakarta in Java, Medan in Sumatra, and Denpasar in Bali. Major cities in Sumatra have limited service to Malaysia, Yogyakarta has limited service to Japan, and Jayapura in Irian Jaya has limited service to Papua New Guinea. Scheduled services within the country are provided by several companies, the most important of which are Garuda, the national airline, and Merpati Nusantara, which is partially subsidized by the government. There are also several nonscheduled airlines.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The Republic of Indonesia was proclaimed in 1945. Its jurisdiction included the present area from

Sabang in Sumatra to Merauke in Irian Jaya, or the entire area of the former Dutch (or Netherlands) East Indies. The Netherlands retained possession of a large part of this region, however, and a provisional capital was established in Yogyakarta, which was the stronghold of the revolution.

With the close of the struggle for independence in 1949, the Republic of the United States of Indonesia was established. The federal system did not last, however, and in 1950 the federated governments unanimously decided to return to a republican form of government. After some difficulties the Republic of Indonesia returned to the constitution of 1945 by presidential decree.

Constitutional framework. Executive power lies in the president, who is assisted by a vice president. Both are elected every five years by the People's Consultative Assembly (Majelis Permusyawaratan Rakyat; MPR), and consequently are responsible, although not subordinate, to that body. The ministers and heads of departments are appointed and dismissed by the president, who is also responsible for the supreme command of the army, navy, and air force. The president has the authority to issue regulations, to implement acts, and to make agreements with foreign countries.

The president is assisted by the Supreme Advisory Council (Dewan Pertimbangan Agung). A maximum of 45 members are appointed by the president for a term of five years. They are prominent figures from all fields with wide experience at the regional as well as the national level. The Supreme Audit Board (Badan Pemeriksa Keuangan) appointed by the president and having 5 members, controls state finance and makes regular reports to the House of People's Representatives (Dewan Perwakilan Rakyat; DPR).

Besides holding executive power, the president is the leader of the legislative branch, the DPR, and in case of emergency may issue governmental regulations with the consent of the DPR as a substitute for legislative acts. If such governmental regulation does not get the consent of the DPR, it is considered revoked.

The People's Consultative Assembly

The People's Consultative Assembly is the highest authority in the state, with the primary responsibilities of electing the president and vice president and determining the constitution and the broad lines of governmental policy. It consists of 1,000 members, including representatives from the DPR, regional delegates, and representatives of political parties and functional groups such as farmers, businessmen, the armed forces, and students, appointed by the president on the basis of nominations from those respective groups. The term of office of the MPR is five years, and the assembly sits at least once every five years.

The DPR consists of 500 members, 400 of whom are elected on a proportional system and 100 of whom are appointed by the respective groups as representatives of political parties and functional groups. The body sits once a year, and its members serve a term of five years.

The Regional Council of Representatives (Dewan Perwakilan Rakyat Daerah) forms the government for each region, the regulations and composition of which are determined along lines similar to those of the DPR.

The president is advised and assisted by a cabinet of ministers from the various government departments. Ministries include those for broad areas such as economic, financial, and industrial affairs; foreign affairs; defense; social welfare; supervision of development; research and technology; and administrative reform. Each minister is assisted by a secretary-general, one or more director generals, an inspector general (if appropriate), and a staff of special assistants.

Local government. Indonesia is divided into 24 *provinsi* (provinces), the two *daerah istimewa* (special autonomous districts) of Aceh and Yogyakarta, each of which is headed by a governor, and the *daerah khusus ibukata* (metropolitan district) of Jakarta Raya. There are three provinces on Java, seven on Sumatra, four on Kalimantan, four on Celebes, four on Bali and the Lesser Sunda Islands, or east Indonesia, and one each for the Moluccas and Irian Jaya.

The 246 second-order divisions, *kabupaten* (regencies), are headed by a *bupati,* and the 3,349 third-order divisions, *kecamatan* (districts), are headed by a *camat.* In addition, there are 54 cities that have obtained autonomous status and have been recognized as *kotamadya* (municipalities), which are headed by a mayor (or *walikota*).

These regional units are all headed by officials of the central government. The kampong, or village, is headed by a *ketua,* and a desa, or group of villages, is headed by a *lurah;* these officials are elected locally and provide the link between the people and the central government on the district level. Regional and local government is highly dependent on the central government, which controls most appointments and collects the majority of the revenues.

The political process. The election law states that all citizens who have reached the minimum age of 17 or who have married may vote in general elections. All those who have reached the age of 21 may stand for elections. Voting is direct and by secret ballot.

The history of national elections

The first election after independence was held in 1955. Almost 170 political parties and factions contested, and 4 major parties obtained the majority of the votes. The election was carried out with little disturbance, but the resulting government was gradually set aside during the closing years of the regime of Sukarno—Indonesia's first national figure and first president, from 1949 to 1967—as the concept of a "Guided Democracy" took hold. At one point there were almost 100 ministries, each competing to build a more impressive edifice. The structure collapsed with an attempted coup d'état in 1965, which led to the downfall of Sukarno.

After a period of stabilization and restructuring in which the armed forces played a major role, the second election of the DPR was held in 1971. Contesting this election were nine political parties and the Joint Secretariat of Functional Groups (Sekretariat Bersama Golongan Karya; Sekber Golkar, or Golkar), a government-sponsored organization of nonaffiliated functional groups. These groups include nonparty associations of peasants, fishermen, civil servants, cooperatives, religious groups, students, the armed forces, and veterans that are allowed to contest the elections on the same level as political parties. After the 1965 disturbances the Golkar took on a stronger role, and the various groups combined into Golkar to present a united front for the 1971 elections, strongly supported by both the government and the military. It is, however, impossible to understand the political working on all levels in Indonesia without being aware of the concept of *mufakat,* or "consensus," arrived at on the basis of extensive consultations (*musyawarah*) aimed at reaching unanimous agreement. Decisions are seldom arbitrary or made by one person but are the result of extensive discussions. This is the traditional approach to all problems. Many parties have combined to form two officially recognized parties, the United Development Party (Partai Persatuan Pembangunan; PPP), and the Indonesian Democratic Party (Partai Demokrasi Indonesia; PDI).

Justice. The judicial system consists of a Supreme Court (Mahkamah Agung) in Jakarta, which is the final court of appeal; high courts located in principal cities on Java, Sumatra, Celebes, Kalimantan, Bali, the Moluccas, and Irian Jaya, which deal with appeals from district courts; and more than 250 district courts.

There are four judicial spheres (for general, religious, military, and administrative matters), each with its own courts. The religious, military, and administrative courts deal with special cases or particular groups of people, while the general deal with normal cases, both civil and criminal.

There is one codified criminal law for all of Indonesia; the Dutch codified civil code is applied to foreigners. For Indonesians the civil law is the uncodified *hukum adat,* or "local customary law," which varies from one district or ethnic group to another.

Armed forces. The Armed Forces of the Republic of Indonesia (Angkatan Bersenjata Republik Indonesia) is not a colonial inheritance but was founded as a national armed force soon after independence. It consists of the army, navy, air force, and the state police. The navy has a small number of cruisers, submarines, destroyers, and frigates, as well as many smaller craft. Military service is not compulsory, except for university students.

Education. At independence educational opportunities

for Indonesians were limited even on the primary and secondary levels and were practically nonexistent on the university level. Since the 1940s the government has placed great emphasis on mass education, and the majority of the children now enter primary schools. Most people are literate.

Responsibility for education is centred in the Ministry of Education and Culture; the ministries of Religious Affairs and Agriculture also have extensive educational programs, and most other ministries have training and upgrading programs in specific areas.

The educational system

The educational system involves six years of primary education, followed by three years of junior and three years of senior secondary schools. Each level is divided into general, vocational, technical, and agricultural curricula.

Higher education includes many public universities—including one in each province—institutes and teacher-training colleges, as well as other institutes and academies controlled by various government ministries, and numerous private institutions of higher learning. Because of the great problem of maintaining adequate staff and standards in such a widespread system, the Ministry of Education and Culture has established five consortia, consisting of representatives from the main universities, to deal with agriculture, science and technology, medicine, social sciences, and education. Major universities include the Bogor Agricultural University, the Bandung Institute of Technology, the University of Indonesia in Jakarta, Gadjah Mada University in Yogyakarta, and Airlangga University in Surabaya.

Health and welfare. The government has made efforts in socialized medicine through the establishment of comprehensive district medical centres. The most complete district centres combine existing clinics with maternal and child-health centres and provide services for family planning, school health, nutrition, communicable-disease control, health statistics, environmental health, health education, dental health, and public-health nursing. These centres also supervise the community and village health centres (*puskesmas*), which are the primary health providers in rural areas. In an effort to improve access to health care for people whose health is most at risk, however, an integrated health-service post concept (*posyandu*) has emerged. These posts are more widely available than the village health centres and offer a variety of services to women and children in particular, ranging from immunizations and nutrition counseling to family planning.

Health conditions in Indonesia are closely related to problems of diet. The major communicable diseases are well under control, although outbreaks of cholera, malaria, tuberculosis, and leprosy occur. The diet of most people consists of rice, vegetables, and a small amount of fish. One of the most serious problems is the shortage of medical and paramedical personnel, mainly nurses and midwives. An indigenous midwife (*dukun beranak*), often with limited training, assists at most of the births in Indonesia; extensive training programs have been set up to bring the *dukun beranak* toward the standards of qualified midwives. Medical training is offered at some state schools and a number of private schools.

The concept of family planning runs counter to traditional views, and there was much resistance to such programs when they were introduced. A massive attempt has been made to provide information on family planning to women of childbearing age, with clinics that are run by the Ministry of Health. This program has achieved considerable success, particularly in Java and Bali, and has come to be considered a model in Asia.

Housing. In rural areas the floors of dwellings consist of pounded earth or concrete, or else of raised wood floors, while wooden framing supports walls of woven bamboo matting, and the roofs are of dried palm fibre or tiles. In urban areas floors are of cement or tile, the framing of the dwellings is of teak or meranti wood, the walls are of brick and plaster, and the roofs of tile or shingle. Although most of the population is nonurban, the major housing problems are in the cities, where new arrivals crowd into squalid slums. In their desire to escape the restraints of the traditional rural life and seek the opportunities of the cities, most immigrants find living conditions that are less attractive than those of the country.

Jakarta is the most modern city in Indonesia and also the one with the greatest problems. It lacks a dependable supply of electricity, gas, and water, an adequate telephone system, a waste-disposal system, and adequate school and health facilities. There is a severe housing shortage, with the gap being filled by substandard, temporary housing that does not require building permits. Subsidized housing is provided largely by employers, including government ministries, for a limited number of key employees, although efforts by the government are being made to provide more low-cost housing to a wider group of people.

CULTURAL LIFE

The diverse cultural heritage

Indonesia exhibits a rich diversity of cultural forms that range from those of the old Malay, which are preserved mainly in the remote interiors of Sumatra and Borneo, through the traditional Javanese and Balinese forms, which are heavily influenced by the Hindu stories of the Mahābhārata and the Rāmāyaṇa, to the modern culture that has evolved from this complex heritage. The American art historian Claire Holt, in *Art in Indonesia* (1967), has divided cultural life into three overlapping spheres: "the Heritage," which includes the statues and monuments of the ancient cultures; "Living Tradition," which covers the traditional theatre using shadow plays (*wayang kulit*), puppets (*wayang golek*), or human actors (*wayang orang* or *wayang wong*) and the use of new media to express traditional concepts as in the painting and sculpture of Bali; and "Modern Art," which encompasses new forms of painting, sculpture, drama, and dance. For much of the population, particularly in the rural areas, "Living Tradition" is a valid term; for the cultural heritage centred around traditional, highly stylized, and semi-ritualistic forms, such as the shadow play, strongly influences all aspects of their lives.

The ancient culture. From the 8th through the 10th century AD, extensive temple complexes (*tjandis*; *candis*) were built in central Java, most of which are now buried or in ruins. The government is actively engaged in restoration.

The remains of the first of the great central Javanese monuments, the Śaivite temple of the Diyeng (Dieng) Plateau, date to the early 8th century. The Śailendra dynasty, which ruled Java and Sumatra (8th–9th century), built the great Mahāyāna Buddhist monuments, including that of Borobuḍur, around AD 800. Late in the 9th century the kings of Mataram built the Hindu monuments around Prambanan. Lara Yonggrang (Lara Jonggrang) Temple, commonly called Prambanan Temple, is the best-preserved of a series of Hindu temple complexes in the region. It consists of six main temples; three large ones along the west, dedicated to Śiva, Vishnu, and Brahmā, contain fine statues. Of the three smaller temples along the east, the middle one contains a statue of Nandi, the bull of Śiva. The main temples are heavily ornamented with stone carvings of the gods and other heavenly beings, and there is a series of relief panels depicting the Rāmāyaṇa story.

Borobuḍur is often considered the most significant monument in the Southern Hemisphere and one of the finest Buddhist monuments in the world. It stands on a hill about 20 miles northwest of Yogyakarta and rises to a height of 115 feet from its base, which measures 403 feet square. The monument consists of a lower structure of six square terraces and an upper structure of three circular terraces, combining the ancient symbols of the circle for the heavens and the square for the earth. In the centre of each side of the square terraces is a staircase leading to the next level. The inner wall on each level has niches containing statues of Buddha, whose life is depicted in the bas-reliefs that cover both inner walls and balustrades. The circular terraces are not decorated and contain 72 bell-shaped stupas, each containing a statue of Buddha. In the centre of the upper terrace is the main stupa, which stands 23 feet high. It was opened in 1842, but no statues or relics were found.

Between the 10th and 16th centuries, the centre of power shifted to eastern Java. Literature in ancient Javanese

(*kawi*) flourished during this period, and a number of large temple complexes were constructed, none of which, however, approached the grandeur of Borobudur or Prambanan. The most imposing complex is Panataran Temple near Blitar, which was constructed at the peak of the Majapahit period in the 14th century. With the ascendancy of Islām in the 16th century, the temples fell into ruins, and the main continuity of Hindu influence shifted to Bali.

Traditional arts. The "Living Tradition" is best represented by the various ways in which the Indian legends of the Rāmāyaṇa and Mahābhārata permeate society. It is in evidence in many Asian countries, but nowhere stronger than in the highly populated rural areas of Java and Bali. The same situations and poses in the carvings on Prambanan are seen in the wayang ("play") performances in contemporary villages and cities. The Indonesian word *wayang* (*wajang*) refers to the flat leather puppet used in shadow plays; in its broader sense it has come to mean the performance itself, whether it involves leather or wooden puppets or human actors.

The wayang

In puppet performances the narration and dialogue are recited by a storyteller (*dalang*), who manipulates the puppets and is the artist of the performance. When leather puppets are used, the audience can sit either in front of or behind a screen, thus viewing either the puppets or their shadows. When people perform these dramas, they often wear masks (*wayang topeng*). The performances are accompanied by a gamelan orchestra, which consists almost entirely of percussion instruments such as gongs, the xylophone-like gender and gambang, a two-stringed instrument called a rebab, and a bamboo flute (*suling*).

Bali long has been of special interest culturally because its Hindu traditions have been preserved, undisturbed by the spread of Islām. Sculpture, wood carving, and painting continue to evolve in an environment that encourages the development of the most colourful and exotic forms. Although continually changing, these forms remain true to the basic traditions and religious beliefs of the people.

There are probably as many distinct dance styles in Indonesia as there are languages and dialects. The most advanced are found on Java, where one of the best-known troupes is the *kraton* (sultan's palace) of Yogyakarta, and on Bali, where many villages have their own dance troupes. The dances have common roots and characteristics, including many kneeling and crouching postures; little running, leaping, or spinning; much use of the hands, fingers, and eyes; and, with the exception of some Balinese dances, a slow tempo. Elaborate and traditional costumes are customary, with emphasis on the headdress, which may include flowers, horns, feathers, or incense sticks. Both men and women participate, the men's dances being more varied and vigorous. Although the stylized dances, such as the Balinese legong or the Javanese serimpi, are the best known, traditional dance styles are found throughout the country. They include the candle and umbrella dances of central Sumatra, the hobbyhorse dances of Java, the trance dances of many regions, and the more ancient tribal dances of the interior of Kalimantan and the eastern islands.

Decorative arts include carvings in stone, wood, bone, and ivory, woven and dyed fabrics, and metalwork. Although some of these owe much to other parts of Southeast Asia, the various sections of Indonesia have produced individual styles. The commemorative and symbolic motifs of both the Dong-Son culture of Indochina and the late Chou dynasty of China have strongly influenced Indonesian art. Textile design is regarded as the most varied and attractive artistic achievement.

The art of batik

Batik making, practiced almost exclusively on Java, involves a complex wax-resistance process in which all parts of a cloth that are not to be dyed are coated on both sides with wax before the cloth is dipped into the dye. Using a penlike wax holder called a *canting,* it is possible to create intricate and elaborate designs. It is a time-consuming process, and those batik fabrics that are made entirely by hand take several weeks to complete. Much modern batik is made using copper stamps (*caps*) to apply the wax, thereby greatly speeding up the process and lowering the cost.

On woven fabric, which is made everywhere from Sumatra through the eastern islands, the most characteristic element is the key-shaped figure combined with other geometric figures. The rhombus (an equilateral parallelogram usually having oblique angles) frequently occurs together with straight lines, equilateral triangles, squares, or circles, which permits an enormous number of variations, including stylized representations of human beings and animals. Each island or region has its characteristic patterns, which serve to identify the area in which the cloth is made.

The art of weaving is highly developed. It includes the famous ikat method, in which the thread is dyed selectively before weaving by binding fibres around groups of threads so that they will not take up colour when the thread is dipped in the dyebath. This process may be applied to the warp, which is most common and is found in Sumatra, Borneo, and Sumba. Weft ikat is found mainly in south Sumatra, and the complex process of double ikat is still carried on in Tenganan in Bali, where such cloth has great ceremonial significance.

The National Museum in Jakarta has an extensive collection of Indonesian carvings, textiles, and artifacts; in addition, it contains models of traditional houses and villages from various parts of the country. The Jakarta Museum displays historic material of the city. There are a number of other museums throughout the country, the most notable of which are the Radya Pustaka in Surakarta (Solo, or Sala) in central Java, the Museum Bali in Denpasar, and the Ratna Warta Fine Arts Museum in Ubud in Bali. The Presidential Palace in Bogor, which has a fine collection of Indonesian art, is located adjacent to the Botanical Gardens.

Modern art forms. In contrast to these traditional forms of cultural expression, modern art forms are not as well developed, although several trends are discernible. A generation of Indonesian choreographers has emerged since independence educated at the performing arts academies. These people are familiar with Western classical and modern dance, and they have adapted traditional dance works for modern audiences. A specific example is the performance style called the *sendratari,* which is essentially a form of traditional dance drama that utilizes modern movements and costumes. This particular event is performed in the Prambanan temple complex.

In the late 19th century a landscape painting style evolved in Indonesia that imitated Western works. Since the 1950s, however, a genuine Indonesian form of expression, derived from Western techniques, has grown and flourished. The artists creating these works, which are done in oil and batik, are concentrated largely in Jakarta, Yogyakarta, and Bali. A tradition of Indonesian contemporary writing has also been developing, spurred in large part by the adoption of the modern Indonesian language. (J.F.McD./T.R.L.)

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

History

THE ARCHIPELAGO AND ITS EARLY HISTORICAL RECORDS

The Indonesian archipelago stretches for more than 3,000 miles east to west and is the largest island complex in the world. The sea has inevitably influenced Indonesian history. Not surprisingly, the boat became a pervasive metaphor in literary and oral tradition and in the arts in Indonesia. Monsoon winds, blowing north and south of the equator, have facilitated communication within the archipelago and with the rest of maritime Asia; the warm rainfall has nourished rich vegetation. In early times the timber and spices of Java and the eastern islands were known afar, as were also the resins from the exceptionally wet equatorial jungle in the western islands of Sumatra and Borneo. Not long after the beginning of the Christian era, goods were already being shipped overseas, and navigable rivers brought the Indonesian hinterland into touch with distant markets.

Easy overseas communication did not, however, result in the formation of territorially large kingdoms. The many estuaries of Sumatra and Borneo, facing the inland seas,

Political fragmentation

possessed an abundance of nutritious seafood that made possible a settled mode of life, and the contacts between the people of one estuary and their neighbours were more important to them than those they could make with overseas lands. Indonesian maritime history is the story of the efforts of local groups, endowed with more or less comparable resources, to protect their separate identities. The same local interests prevailed on the island of Java, where the lava-enriched soil, watered by gently flowing rivers, encouraged wet-rice production and a patchwork pattern of settled areas in the river valleys separated by mountains and jungle. Long before records begin, many of these coastal and riverine groups were evolving an elementary form of hierarchy, accompanied by the craftsmen's tokens of rank. No single group was large enough to overrun and occupy neighbouring territories; its energies were absorbed rather by an ever more intensive exploitation of its own natural resources. Those living on or close to the sea knew that geographic isolation was out of the question but regarded their maritime environment as a means of enhancing their well-being through imports or new skills. Looking outward, far from inculcating a sense of belonging to larger communities, encouraged the pursuit of local interests. Thus, the structure of Indonesian written and oral sources suggests to historians that the origins of kingdoms on the coasts of the Java Sea were associated with the success of local heroes in turning the arrival of foreign trading treasure to their advantage.

Indonesian place-names have frequently remained unchanged since the beginning of documented history. In these often nearby places, each leader saw himself at the centre of the world that mattered to him, which was not, until later, the archipelago or even a single island but his own strip of coast or river valley. Some centres achieved local hegemony but never to the extent of extinguishing permanently the pretensions of rival centres. Thus, the early history of Indonesia is compounded of many regional histories that only gradually impinge on each other.

Indonesia's languages

The historical fragmentation of the archipelago, sustained by its rich climate and accentuated rather than offset by easy access to the outside world, is reflected in its languages. Scholars have debated the location of the areas outside Indonesia from which the speakers of the Austronesian (Malayo-Polynesian) languages originally came: the Asian mainland and the Pacific islands have been proposed.

What is significant for the historian, however, is that the speakers of these languages almost certainly drifted into the region in small groups over long periods of time and did not suddenly assume a common identity when they reached the coasts and rivers of the archipelago. On the contrary, they remained scattered groups, sometimes coexisting with descendants of earlier Pleistocene populations, who, in their turn, had also learned to make economic use of their environment over an immense span of cultural time. The perhaps 200 languages within the Western, or Indonesian, branch of the Austronesian family are an index of the manner in which the peoples of the Indonesian archipelago submitted to the realities of their landscape.

The historian, examining stone or metal inscriptions, which, together with surviving copies of early religious texts, are the important sources of documentary information, must remember that the evidence is always concerned with specific places. Comprehensive narrative histories of extensive areas cannot be written. The reality behind interregional relationships must often remain a riddle. The historian's task is the study of cultural history in widely scattered groups of society rather than narrative accounts of still very indistinct kingdoms; it is the investigation of beliefs shared by the ruling classes and the peasantry and of the points of contact between them. The ideas of men of rank were articulated in architecture and literature, reflecting varying degrees of exposure to influences outside the archipelago, but all groups of the population subscribed to basic assumptions concerning dependence of people on the goodwill of the gods.

INDONESIAN "HINDUISM"

It may one day be shown by students of prehistory that Indonesians were sailing to other parts of Asia long ago. Records of foreign trade, however, begin only in the early centuries AD. A study of the Roman historian Pliny the Elder's *Natural History* suggests that, in the 1st century AD, Indonesian outriggers were engaged in trade with the east coast of Africa. Indonesian settlements may have existed at that time in Madagascar, an island with distinct Indonesian cultural traits. The geographer Ptolemy, in the following century, incorporated information from Indian merchants in his *Guide to Geography* concerning "Iabadiou," presumably referring to Java, and "Malaiou," which, with its variants, may refer to Malayu in southeastern Sumatra.

Foreign contacts

Regular voyages between Indonesia and China did not begin before the 5th century AD. Chinese literature in the 5th and 6th centuries refers to western Indonesian tree produce, including camphor from northern Sumatra, and also to two resins that seem to have been added to the seaborne trade in western Asian resins and were known in China as "Persian resins from the south ocean." Indonesian shippers were probably exploiting the economic difficulties southern China was suffering at the time because it had been cut off from the ancient Central Asian trade route. Certain small estuary kingdoms were beginning to prosper as international entrepôts. Their location is unknown, though Palembang's commercial prominence in the 7th century suggests that the Malays of southeastern Sumatra had been active in the "Persian" trade with southern China.

Hindu religious conceptions. The cultural effects of these commercial exchanges, usually described as "Hinduization," have been discussed for many years. It is now held that Hinduism was brought to Indonesia not by traders, as was formerly thought, but by Brahmans who taught the Śaivite message of personal immortality. Sanskrit inscriptions, attributed to the 5th and 6th centuries, have been found in eastern Kalimantan, a considerable distance from the international trade route, and also in western Java. They reveal that Indian literati, or their Indonesian disciples, were honoured in some royal courts. The rulers were prominent *rakas*, heads of groups of villages in areas where irrigation and other needs had brought into being intervillage relationships and supravillage authority. The inscriptions, and also Chinese sources, indicate that some rulers were involved in warfare and must have been seeking to extend their influence. The Śaivite Brahmans supervised the worship of Śiva's phallic symbol, the lingam (*liṅga*), in order to tap the god's favours on behalf of their royal patrons. These Brahmans were representatives of an increasingly influential devotional movement (bhakti) in contemporary Indian Hinduism; they probably also taught their patrons how to achieve a personal relationship with the god through "austerity, strength, and self-restraint," in the words of one inscription from Borneo. The rulers, therefore, were encouraged to attribute their worldly successes to Śiva's grace; the grace was obtained through devotional exercises lovingly offered to Śiva and probably regarded as the guarantee of a superior status in the life after death. These Śaivite cults, marks of a privileged spiritual life, would have been a source of prestige and royal authority.

Indonesian religious conceptions. The question must be asked, however, to what extent such religious ideas were comprehensible to those who first heard them. Indonesians, who had been accustomed to constructing terraced mountainlike temples—symbolizing holy mountains—for the burial and worship of the dead, would not have been perplexed by the Brahmans' doctrine that Śiva also dwelt on a holy mountain. Natural stones, already placed on mountain terraces for the ritual of megalithic worship, would have been easily identified with Śiva's natural stone lingam, the most prestigious of all lingams. Indonesians, who were already concerned with the passage rites and welfare of the dead, and who considered the elaborate rituals of metalworking as a metaphor for spiritual transmutation and liberation of the soul, would have paid particular attention to Hindu devotional techniques for achieving immortality in Śiva's abode. The meditative ascetic of Hinduism may have been preceded in Indonesia by the trance-inducing shaman (priest-healer). Again, the

Mountain temples

notion that water was a purifying agent because it had been purified by Śiva's creative energy on his mountain-top would have been intelligible to mountain-worshiping Indonesians, especially if they already endowed the water flowing from their own gods' mountain peaks with divinely fertilizing qualities.

Indonesian religious conceptions must certainly have supplied the perspectives of those who first listened to the Brahmans. Confidence in the Brahmans, honoured especially as teachers (gurus), would have depended on their demonstrating means of achieving religious goals already recognized as important in the indigenous system of beliefs. The Brahmans' role was probably prepared during earlier visits by Buddhist missionaries, who also shared the Indian concern for religious salvation.

But Indonesian circumstances and motivation underlay the adoption of Indian forms. The use of Hindu terminology in the inscriptions represents no more than Indonesian attempts to find suitable metaphoric expressions from the sacred Sanskrit literature for describing their own realities. Sanskrit literature, imported from India on manuscripts or by feats of memory, would have been especially culled when courtly literati were seeking to describe those rulers who had achieved an intensive personal relationship with Śiva. One must not be deceived by the accumulating acquaintance with Indian civilization reflected in Indonesian inscriptions and Javanese literature. The Indonesians, like others in early Southeast Asia, had no difficulty in identifying themselves with the universal values of "Hindu" civilization represented by the sacred literature. Indian literary and legal works were to provide useful guidelines for Indonesian creative writing, but they did not bring about a thoroughgoing "Hinduization" of the archipelago any more than Indian Brahmans were responsible for the formation of the early kingdoms of the archipelago.

In the final analysis, therefore, India should be regarded as an arsenal of religious skills, the use of which was subordinated to the ends of the Indonesians. Expanding communication meant that increasing numbers of Indonesians became interested in Indian thought. The first reasonably well-documented period of maritime Malay history provides further evidence of the Indonesian adaptation of Indian religious conceptions.

THE MALAY KINGDOM OF ŚRĪVIJAYA-PALEMBANG

The kingdom of Śrīvijaya is first mentioned in the writings of the Chinese Buddhist pilgrim I-ching, who visited it in 671 after a voyage of less than 20 days from Canton. He was on the first stage of his journey to the great teaching centre of Nālandā in northeastern India. The ruler of Śrīvijaya assisted I-ching on his journey. Archaeological surveys undertaken since the late 1970s immediately to the west of Palembang city—an area now being overtaken by suburban development—have revealed such a quantity of materials as to make it practically certain that this was Śrīvijaya's heartland in the 7th and subsequent three centuries. Surface remains of more than a thousand shards of Chinese ceramics, two-thirds of which are datable from the 8th to the 10th century, have been recovered from several sites. Shards from the 11th to the 14th century found elsewhere in the neighbourhood may represent shifts of political and commercial activity in the Palembang area. Shards found on nearby Seguntang Hill (Bukit Seguntang), on the other hand, span all these centuries. A piece of Romano-Indian rouletted ware, attributable to the early centuries AD, has been dug up in Palembang near the river; the same ware has been found in Java near Jakarta. Moreover, new stone statues have been found, and the sheer bulk of Buddhist and Hindu statuary now recovered from the Musi River basin has suggested to at least one art historian that the basin must have contained the site of a polity near the sea that enjoyed considerable international contacts. Only Palembang suggests itself as the site in question. Finally, stupa remains have been unearthed at the foot of Seguntang Hill. These discoveries reinforce the textual evidence that Palembang was the heartland of this empire.

Archaeological evidence

Buddhism in Palembang. Śrīvijaya-Palembang's importance has been established by Arab and Chinese historical sources spanning a long period of time. Its own records, in the form of Old Malay inscriptions, are limited almost entirely to the second half of the 7th century (682–686). The inscriptions reveal that the ruler was served by a hierarchy of officials and that he possessed wealth. The period when the inscriptions were written was an agitated one. Battles are mentioned, and the ruler had to reckon with disaffection and intrigues at his capital. Indeed, the main theme of the inscriptions is a curse on those who broke a loyalty oath administered by drinking holy water. The penalty for disloyalty was death, but those who obeyed the ruler were promised eternal bliss.

I-ching recommended Palembang, with more than a thousand monks, as an excellent centre to begin studying Buddhist texts. The 7th-century inscriptions, however, are concerned with less scholarly features of Buddhism. They deal with Tantric aids to magical power (see below), in the form of yantra symbols, which were distributed by the ruler as favours to faithful servants. Some of his adversaries disposed of them, too. Especially interesting as evidence of the influence of Buddhism within the context of royal power is the Talang Tuwo inscription of 684, which records the king's prayer that a park he has endowed may give merit to all living beings. The language and style of this inscription, incorporating Indian Tantric conceptions, make it clear that the ruler was presenting himself as a bodhisattva—one who was to become a Buddha himself—teaching the several stages toward supreme enlightenment. Here is the first instance in the archipelago's history of a ruler's assumption of the role of religious leader.

The inscriptions show that the teachings of the Tantric school of Mahāyāna Buddhism, with its magical procedures for achieving supernatural ends, had reached Palembang before the end of the 7th century. Tantric Buddhism came into prominence in India only in the 7th century, and the synchronism of its appearance in Palembang reflects not only the regularity of shipping contacts between Sumatra and India but, more importantly, the Malays' quick perception of the contribution of Tantric Buddhism as a source of personal spiritual power. The word for "curse" in the inscriptions is Malay, and it is reasonable to suppose that the Malays grafted Tantric techniques onto indigenous magical procedures. The vitality of Malay religion is probably also reflected in the prestige of the sacred Seguntang Hill near Palembang, which was visited by those in search of spiritual power. Seguntang Hill would not suddenly have become such a centre as a result of traffic in Tantric conceptions during the 7th century. In other words, the disturbances reflected in the inscriptions are less likely to have been the growing pains of a rising kingdom than the efforts of an already important kingdom to achieve, or perhaps recover, hegemony in southern Sumatra.

The maritime influence. Special circumstances affecting Śrīvijaya-Palembang toward the end of the 7th century are consistent with this conclusion. In the centuries before the Chinese undertook long voyages overseas, they relied on foreign shipping for their imports, and foreign merchants, trading with China, required a safe base in Indonesia before sailing on to China. This seaborne trade, regarded in China as "tributary" trade with the "emperors' barbarian vassals," had developed during the 5th and 6th centuries but languished in the second half of the 6th century as a result of the civil war in China that preceded the rise of the Sui and T'ang dynasties. Chinese records for the first half of the 7th century mention several small harbour kingdoms in the region, especially in northeastern Sumatra, that were pretending to be Chinese vassals. The rulers of Palembang, hoping for a revival of trade under the new T'ang dynasty, must have been anxious to monopolize the China trade and eliminate their rivals. They succeeded in doing this. Before I-ching left Southeast Asia in 695, Śrīvijaya was in control of the Strait of Malacca; the ruler's determination to control all harbours in the region that might compete in the China trade explains his militancy, as shown in the Old Malay inscriptions.

Importance of the China trade

The subsequent power of the maharajas of Śrīvijaya depended on their alliance with those who possessed warships. The fact that Arab accounts make no mention of

piracy in the islands at the southern end of the Strait of Malacca suggests that the seafaring inhabitants of these islands identified their interests with those of the maharajas, refraining from molesting merchant ships and cooperating in controlling Śrīvijaya's potential competitors in northern Sumatra. The maharajas offered their loyal subjects wealth, posts of honour, and—according to the inscriptions—supernatural rewards. But the grouping of maritime Malays in this geographically fragmented region survived only as long as the Palembang entrepôt was prosperous and its ruler offered enough largess to hold the elements together. His bounty, however, depended on the survival of the Chinese tributary trading system, which needed a great entrepôt in western Indonesia. Early Malay history is, to an important extent, the history of a Sino-Malay alliance. The maharajas benefited from the China trade, while the emperors could permit themselves the conceit that the maharajas were reliable imperial agents.

The Palembang rulers' exact span of territorial influence is unknown. The Banka Strait and the offshore islands at the southern entrance of the Strait of Malacca would have been essential to their maritime power. According to the 7th-century inscriptions, the rulers also had influence in southern Sumatra on the Sunda Strait. Elsewhere in the hinterland, including what became known as Malayu in the Hari River basin, their authority would have been exercised by alliances with local chiefs or by force and always with decreasing effect the further these areas were from Palembang.

Decline of Palembang

Malay unity under the leadership of the maharajas was inevitably undermined when, as early as the 10th century, Chinese private ships began to sail to centres of production in the archipelago, with the result that the Chinese market no longer depended on a single Indonesian entrepôt. Toward the end of the 11th century, Śrīvijaya-Palembang ceased to be the chief estuary kingdom in Sumatra. Hegemony had passed, for unknown reasons, to the neighbouring estuary town of Jambi, which was probably controlled by the great Minangkabau country of Malayu in the interior. With the decline of the tributary trade with China, a number of harbours in the region became centres of international trade. Malayu-Jambi never had the opportunity to build up naval resources as Śrīvijaya-Palembang had done, and in the 13th century a Javanese prince took advantage of the power vacuum.

CENTRAL JAVA FROM THE 8TH TO THE 10TH CENTURY

Eastern Javanese inscriptions throw little light on happenings before the 10th century, but the evidence from south-central Java, and especially from the Kedu Plain in the 8th and 9th centuries, is more abundant. This period in central Java is associated with the Śailendra princes and their rivals. An Old Malay inscription from north-central Java, attributed to the 7th century, establishes that the Śailendras were of Indonesian origin and not, as was once suspected, from mainland Southeast Asia. In the middle of the 9th century the ruler of Śrīvijaya-Palembang was a Śailendra who boasted of his Javanese ancestors; the name Śailendra also appears on the undated face of an inscription on the isthmus of the Malay Peninsula; the other face of the inscription—dated 775—is in honour of the ruler of Śrīvijaya.



Sites associated with early Indonesian history.

In spite of ambiguous references to Śailendra connections overseas, there is no solid evidence that the territories of the central Javanese rulers at this time extended far beyond central Java, including its north coast. Yet the agricultural wealth of this small kingdom sustained vast religious undertakings; the monuments of the Kedu Plain are the most famous in Indonesia. The Borobuḍur temple complex, in honour of Mahāyāna Buddhism, contains 2,000,000 cubic feet (57,000 cubic metres) of stone and includes 27,000 square feet (2,600 square metres) of stone bas-relief. Its construction extended from the late 8th century to the fourth or fifth decade of the 9th. Śiva's great temple at Prambanan, though not associated with the Śailendra family, is less than 50 miles (80 kilometres) away, and an inscription dating to 856 marks what may be its foundation stone. The two monuments, which have much in common, help to explain the religious impulses in earlier Javanese history.

Borobuḍur

Borobuḍur is a terraced temple surmounted by stupas, or stone towers; the terraces resemble Indonesian burial foundations, indicating that Borobuḍur was regarded as the symbol of the final resting place of its founder, a Śailendra, who was united after his death with the Buddha. The Prambanan temple complex is also associated with a dead king. The inscription of 856 mentions a royal funeral ceremony and shows that the dead king had joined Śiva, just as the founder of the Borobuḍur monument had joined the Buddha. Divine attributes, however, had been ascribed to kings during their lifetimes. A Mahāyāna inscription of this period shows that a ruler was said to have the purifying powers of a bodhisattva, the status assumed by the ruler of Śrīvijaya in the 7th century; a 9th-century Śaivite inscription from the Kedu Plain describes a ruler as being "a portion of Śiva."

The divine qualities of these kings, whether of Mahāyāna or of Śaivite persuasion, had important implications in Javanese history and probably in the history of all parts of the archipelago that professed the forms of Indian religion. The ruler was now and henceforth seen as one who had achieved union with the supreme god in his lifetime. Kingship was divine only because the king's soul was the host of the supreme god and because all the king's actions were bound to be the god's actions. He was not a god-king; he was the god. No godlike action was more important than extending the means of personal salvation to others, always in the form of union with the god. The bas-relief of the Borobuḍur monument, illustrating Mahāyāna texts and especially the *Gaṇḍavyūha*—the tale of the tireless pilgrim in search of enlightenment—is a gigantic exposition of the Mahāyāna path to salvation taken by the king; it may be thought of as a yantra, or instrument to promote meditation and ultimate union with the Buddha. But Borobuḍur can also be identified as a circle, or mandala, of supreme mystical power that signified the Void of the Vairocana Buddha according to the Vajrayāna persuasion of Tantric Buddhism. The mandala was intended to protect the Śailendra realm for all time. The pedagogical symbolism of the Prambanan temple complex is revealed in its iconography, dominated by the image of the four-armed Śiva, the Great Teacher—the customary Indonesian representation of the supreme deity. Prambanan affirms the Śaivite path to salvation; the path is indicated in the inscription of 856, which implies that the king had practiced asceticism, the form of worship most acceptable to Śiva. Śaivism in Java as well as Mahāyāna Buddhism had become hospitable to Tantric influences. An almost contemporary inscription from the Ratubaka Plateau, which is not far from the Prambanan complex, alludes to special rites for awakening Śiva's divine energy through the medium of a ritual consort.

These royal tombs taught the means of salvation. The

The significance of royal tombs

royal role on earth was similar. The kings, not the religious elite, bore the responsibility of ensuring that all could worship the gods, whether under Indian or Indonesian names. Every god in the land was either a manifestation of Śiva or a subordinate member of Śiva's pantheon, and worship therefore implied homage to the king, who was part of the god. The growing together, as a result of Tantric influences, of Śaivism and Mahāyāna Buddhism meant that, over the centuries, the divine character of the king became continually elaborated. His responsibility was the compassionate one of maintaining his kingdom as a holy land. The bodhisattva-king was moved by pity, as were all bodhisattvas, while the Śiva-like king, as an inscription of the 9th century indicates, was also honoured for his compassion. Compassion was expressed by providing an environment wherein religion could flourish. Keeping the peace, protecting the numerous holy sites, encouraging religious learning, and above all performing purification rituals to render the land acceptable to the gods were different aspects of a single mission: the teaching of the religious significance of life on earth. The lonely status of the ruler did not separate him from the religious aspirations of his subjects; Prambanan provides a recognition of the community of interest between ruler and ruled. The 856 inscription states that a tank of purifying water, filled by a diverted river, was made available as a pilgrimage centre for spiritual blessings. Hermitages had been built at the Prambanan complex, and the inscription states that they were "to be beautiful in order to be imitated."

The great monuments of the 9th century suggest something of the cultural ambience within which events took place. One new development in central Java was that capable local rulers, called *raka,* were gradually able, when opportunities arose, to fragment the lands of some *raka* and absorb the lands of others. At the same time, they established lines of communication between themselves and the villages in order to guarantee revenue and preserve a balance between their own demands and the interests of the independent and prosperous agricultural communities. When a ruler manifested divine qualities, he would attract those who were confident that they were earning religious merit when they supported him. Local princes from all over the Kedu Plain constructed small shrines around the main Prambanan temple in a manner reminiscent of a congregation gathered around a religious leader. The inscription of 856 states that they built "cheerfully."

EASTERN JAVA AND THE ARCHIPELAGO FROM 1019 TO 1292

After the beginning of the 10th century, inscriptions and monuments in central Java cease. For more than 500 years little is known of developments in central Java, and nothing of what happened in western Java or in the eastern hook of the island. The evidence for these years comes almost exclusively from the Brantas River valley and the adjacent valleys of eastern Java. This abrupt shift in the historian's focus of attention has never been satisfactorily explained.

Government and politics. Eastern Java did not form a natural political unit. No single town emerged that was so exceptionally endowed in local resources as to become a permanent capital; instead, the residencies of defeated kings were abandoned, and the sites of some of them are unknown. The problems of government in these conditions are illustrated by the events of the 11th century. In 1016 the overlord's city was destroyed in what an inscription of 1041 (called the "Calcutta" inscription) described as "the destruction of the world," and the kingdom fell apart. The most recent explanation of the episode is that a Javanese vassal had rebelled. The kingdom was restored by the dead king's son-in-law Airlangga (Erlangga), a half-Balinese prince. From 1017 to 1019 he lived with hermits, probably practicing asceticism. In 1019 he was hailed as ruler of the small principality of Pasuruan near the Brantas delta, but he could not take the military offensive until 1028 and his final success was not before 1035. His victories gradually established his claims to divine power. Airlangga dispatched his last enemy by provoking an uprising against him in the manner taught by Kauṭilya, the master of Indian statecraft who recommended the use of subversion against an enemy. In his "Calcutta" inscription Airlangga expressed the hope that all in the land would now be able to lead religious lives.

Rule of Airlangga

He then undid the results of his achievement. Foreseeing that two of his sons might quarrel, he divided his kingdom so that one son should rule over the southern part, known as Panjalu, Kaḍiri, or Daha, and the other over the northern part, Janggala. The consequences of this decision are mourned in a 14th-century poem, the *Nāgarakeṛtāgama.* Airlangga's sons refused to honour their father's intentions. Fighting broke out, and the Kaḍiri rulers were unable to establish their uneasy domination over the kingdom until the early 12th century.

The chain of command between the capital and the villages—and the number of officials involved—had grown since the central Java period. The ideal of a greater Javanese unity, protected by a divine king, was probably cherished most by the villagers, since they especially would benefit from peace and safe internal communications. Inscriptions sometimes acknowledge the king's gratitude for villagers' assistance in times of need. The villages were prosperous centres of local government. As a result of increasing contacts with the royal court, village society had now become more stratified, with elaborate signs of status. But local lords could make difficulties for the villages by tampering with the flow of the river or exacting heavy tolls from traders. In comparison with these vexations, the royal right to the villagers' services and part of their produce was probably not resented. No document was more respected than the inscription that recorded a village's privileges.

The king's chief secular responsibility was to safeguard his subjects' lands, including the estates of the temples and monasteries that were so conspicuous a feature of the Javanese landscape. When the king wanted to build a temple on wet-rice land he was expected to buy the land, not confiscate it. At court he was assisted by a small group of high officials, among whom his heir seems to have been the most important. Officials were rewarded with appanages from royal lands, for the king, like his noble vassals, was also a regional lord. The council of officials passed on royal decisions to subordinates. Officials made a circuit of the country and visited village elders. Royal rule was probably not harsh; the protests that have been preserved were probably prompted by unusually weak government. A reasonable relationship between ruler and villagers may be seen in a Balinese inscription of 1025 that records a king's sale of his hunting land to a village after the villagers had complained of their lack of land. Village elders sat with the officers of royal law in order to guarantee fair trials and verdicts reflecting the consensus of local opinion. Customary law was incorporated in the royal statutes. Aggrieved individuals could appeal to the king for redress; groups of villages sought his assistance for large-scale irrigation works. The villages paid taxes to the ruler, who thus enjoyed an economic advantage over other regional lords. Everything depended on the ruler's energy and a general agreement that his government served the interests of all.

The Kaḍiri princes of the 12th century ruled over a land that was never free from rebellion. In 1222 Kertajaya was defeated by an adventurer, Angrok, and a new capital was located at Kutaraja, later renamed Singhasāri, near to the harbours of east Java. The changed economic circumstances in the archipelago as a whole must now be taken into account, since they have an important bearing on the internal history of Java in the 13th and 14th centuries.

The empire of Kertanagara. Long before the 12th century, Chinese shipping had become capable of distant voyages, and Chinese merchants sailed directly to the numerous producing centres in the archipelago. The eastern Javanese ports became more prosperous than ever before. A smaller entrepôt trade also developed on the coasts of Sumatra and Borneo and in the offshore islands at the southern entrance to the Strait of Malacca. Heaps of Chinese ceramics of the 12th to 14th century provide remarkable testimony to an important trading centre at Kota Cina near modern Medan on the northeast coast of Sumatra. In consequence, the Minangkabau princes in the

Eastern Java's prosperity

hinterland of central Sumatra, heirs to the pretensions of the great overlords of Śrīvijaya-Palembang, were deprived of the opportunity of developing their port of Jambi as a rich and powerful trading centre. A power vacuum existed in the seas of western Indonesia, and the Javanese kings aspired to fill it.

Java had probably long been regarded as the centre of a brilliant civilization. Old Javanese became the language of the inscriptions of the island of Bali in the 11th century, and in many parts of the archipelago the contacts of trade must have spread Java's reputation as an island of scholars. A study of the grafting of Tantric ritual onto a megalithic shrine at Bongkisam in Sarawak, some time after the 9th century, provides a glimpse of cultural diffusion at work on the maritime fringes of Indonesia. Javanese cultural influence in other islands almost certainly preceded political domination.

Disunity in the Malay world and the cultural fame of Java are not sufficient to explain why the Javanese king Kertanagara (reigned 1268–92) chose to impose his authority on Malayu in southern Sumatra in 1275. It has been suggested that the king's concern was to protect Indonesia from the threat of the Mongol ruler Kublai Khan by organizing a religious alliance. But Kertanagara probably imposed his political authority as well, though his demands would have been limited to expressions of homage and tribute.

The king's activities overseas were almost certainly intended to enhance his prestige in Java itself, where he was never free from enemies. His political priorities are reflected in a Sanskrit inscription of 1289, attached to an image of the king in the guise of the wrathful Akṣobhya Buddha, claiming that he had restored unity to Java; his overseas exploits are not mentioned.

The divine king

The precise doctrinal contents of Kertanagara's Tantric cult are unknown. In his lifetime and after his death his supporters revered him as a Śiva-Buddha. They believed that he had tapped within himself demonic forces that enabled him to destroy the demons who sought to divide Java. The 14th-century poet Prapañcā, author of the *Nāgarakeṛtāgama* and a worshiper of Kertanagara, on one occasion refers to the king as the "Vairocana Buddha" and associates him with a ritual consort, who is, however, the consort of Akṣobhya Buddha. Prapañcā also admires the king's scholarly zeal and especially his assiduous performance of religious exercises for the good of mankind. The role of the royal ascetic had long been a familiar feature of Javanese kingship. The king who had been buried in the 9th-century mausoleum of Prambanan was identified with Śiva, the teacher of asceticism. Early in the 13th century King Angrok, according to a later chronicle, regarded himself as the Bhaṭāra Guru and therefore as Śiva, the patron of ascetics. Śaivite and Mahāyāna priests were under royal supervision from at least as early as the 10th century, and the Tantric concept of a Śiva-Buddha, taught by Kertanagara, would not have been regarded as extraordinary. Javanese religious speculation had come to interpret Śaivism and the Mahāyāna as identical programs for personal salvation, with complementary gods. Union with divinity, to be achieved here and now, was the goal of all ascetics, including the king, who was regarded as the paragon of ascetic skill. Kertanagara's religious status, as well as his political problems and policies, are by no means eccentric features in early Javanese history; particular circumstances, stemming from Chinese participation in maritime trade in the archipelago, enabled him to exercise his divine power beyond Java itself. In the 14th century the homage of overseas rulers to the Javanese king was taken for granted.

THE MAJAPAHIT ERA

In 1289 Kertanagara maltreated Kublai Khan's envoy, who had been sent to demand the Javanese king's submission. The Mongol emperor organized a punitive expedition in 1292, but Kertanagara had been killed by a Kaḍiri rebel, Jayakatwang, before the invaders landed. Jayakatwang in his turn was quickly overthrown by Kertanagara's son-in-law, later known as Kertarajasa, who used the Mongols to his own advantage and then forced them to withdraw in confusion. The capital city was now established at Majapahit. For some years the new ruler and his son, who regarded themselves as successors of Kertanagara, had to suppress rebellions in Java; not until 1319 was Majapahit's authority firmly established in Java with the assistance of the renowned soldier Gajah Mada. Gajah Mada was the chief officer of state during the reign of Kertanagara's daughter (*c.* 1329–50), and in these years Javanese influence was restored in Bali, Sumatra, and Borneo. Kertanagara's great-grandson, Hayam Wuruk, became king in 1350 under the name of Rajasanagara.

Hayam Wuruk's reign

Hayam Wuruk's reign (1350–89) is remembered in the archipelago as the most glorious period in Javanese history. Prapañcā's poem, the *Nāgarakeṛtāgama,* written in 1365 and surviving in a manuscript found in Lombok at the end of the 19th century, provides a rare glimpse of the kingdom from a contemporary point of view. The poem, originally called the *Deśa warṇana,* or "The Description of the Country," describes itself as a "literary temple" and endeavours to show how royal divinity permeates the world, cleansing it of impurities and enabling all to fulfill their obligations to the gods and therefore to the holy land—the now undivided kingdom of Java. The poem resembles an act of worship rather than a chronicle. The poet does not conceal his intention of venerating the king, and, in the tradition of Javanese poetry, he may have begun it under the stimulus of pious meditation intended to bring him into contact with divine influences embodied in the king.

The core territories of Hayam Wuruk's polity were probably considerably more extensive than those of his predecessors. Important territorial rulers, bound to the royal family by marriage, were brought under surveillance by incorporation in the court administration. A network of royal religious foundations was focused on the capital. But the question remains whether a genuinely more centralized and enduring structure of government was introduced or whether the unity of the realm and the ruler's authority still depended on the ruler's personal prestige. Prapañcā, at least, does not ascribe to him an unrealistic degree of authority, even though his poem is an undisguised representation of the attributes of royal divinity and effects of divine rule in Java. Subordinate officials traveled around the kingdom, asserting the royal authority in such matters as taxes and the control of religious foundations. A sign of the king's prestige was his decision to undertake a land survey to ensure that his subjects' privileges were being maintained. In the absence of an elaborate system of administration, the authority of the government was strengthened by the ubiquity of its representatives, and no one set a more strenuous example than the king himself. According to Prapañcā, "the prince was not for long in the royal residence," and much of the poem is an account of royal progresses. In this way Hayam Wuruk was able to assert his influence in restless areas, enforce homage from territorial lords, reassure village elders by his visits, verify land rights, collect tribute, worship at Mahāyāna, Śaivite, and ancient Javanese holy sites, and visit holy men in the countryside for his own spiritual enlightenment. His indefatigable traveling, at least in the earlier years of his reign, meant that many of his subjects had the opportunity of coming into the presence of one whom they regarded as the receptacle of divinity.

The New Year ceremony

One of the most interesting sections of the *Nāgarakeṛtāgama* concerns the annual New Year ceremony, when the purifying powers of the king were reinforced by the administration of holy water. The ceremony, attended by scholarly Indian visitors, enables the poet to assert that the only famous countries were Java and India because both contained many religious experts. At no time in the year was the king's religious role more emphatically recognized than at the New Year, when the notables of the kingdom, the envoys of vassals, and village leaders came to Majapahit to pay homage and be reminded of their duties. The ceremony ended with speeches to the visitors on the need to keep the peace and maintain the rice fields. The king explained that only when the capital was supported by the countryside was it safe from attack by "foreign islands."

Since the poem venerates the king, it is not surprising

that more than 80 places in the archipelago are described as vassal territories and that the mainland kingdoms, with the exception of Vietnam, are said to be protected by the king. Prapañcā, believing that the king's glory extends in all directions, delineates in detail the actual limits of relevant space from a 14th-century Javanese point of view. No fewer than 25 places in Sumatra are mentioned, and the Spice Islands, whose product was a source of royal wealth, are well represented. On the other hand, northern Celebes (Sulawesi) and the Philippines are not mentioned.

During Hayam Wuruk's lifetime Javanese overseas prestige was undoubtedly considerable, though the king demanded no more than homage and tribute from his more important vassals, such as the ruler of Malayu in Sumatra. In 1377, when a new Malayu ruler dared to seek investiture from the founder of the Ming dynasty in China, Hayam Wuruk's envoys in Nanking convinced the emperor that Malayu was not an independent country. Javanese influence in the archipelago, however, depended on the ruler's authority in Java itself. When Hayam Wuruk died in 1389, the Palembang ruler in southeastern Sumatra saw his opportunity for repudiating his vassal status. He had noted the Ming dynasty's restoration of the long-abandoned tributary trading system and its prohibition of Chinese voyages to Southeast Asia and supposed that foreign traders would again need the sort of entrepôt facilities in western Indonesia that Śrīvijaya-Palembang had provided centuries earlier. He may even have announced himself as a bodhisattva and heir of the maharajas of Śrīvijaya. The Javanese expelled him from Palembang, whence he fled to Singapore and then to Malacca on the Malay Peninsula.

ISLĀMIC INFLUENCE IN INDONESIA

Muslim kingdoms of northern Sumatra. Foreign Muslims had traded in Indonesia and China for many centuries; a Muslim tombstone in eastern Java bears a date corresponding to 1082. But substantial evidence of Islām in Indonesia begins only in northern Sumatra at the end of the 13th century. Two small Muslim trading kingdoms existed by that time at Samudra-Pasai and Perlak. A royal tomb at Samudra, of 1297, is inscribed entirely in Arabic. By the 15th century the beachheads of Islām in Indonesia had multiplied with the emergence of several harbour kingdoms, ruled by local Muslim princes, on the north coast of Java and elsewhere along the main trading route as far east as Ternate and Tidore in the Moluccas.

The establishment of the first Muslim centres in Indonesia was probably a result of commercial circumstances. By the 13th century, in the absence of a strong and stable entrepôt in western Indonesia, foreign traders were drawn to harbours on the northern Sumatran shores of the Bay of Bengal, distant from the dangerous pirate lairs at the southern end of the Strait of Malacca. Northern Sumatra had a hinterland rich in gold and forest produce, and pepper was being cultivated at the beginning of the 15th century. It was accessible to all archipelago merchants who wanted to meet ships from the Indian Ocean. By the end of the 14th century, Samudra-Pasai had become a wealthy commercial centre, giving way in the early 15th century to the better protected harbour of Malacca on the southwest coast of the Malay Peninsula. Javanese middlemen, converging on Malacca, ensured its importance for foreigners and Indonesians alike.

Relocation of centres of trade

Pasai's economic and political fame depended almost entirely on foreigners. Muslim traders and teachers were probably associated with its administration from the beginning and were bound to introduce the religious institutions that made foreign Muslims feel at home. The first Muslim beachheads in Indonesia, and especially Pasai, were to a considerable extent genuine Muslim creations that commanded the loyalty of the local population and encouraged scholarly activities. There were similar new harbour kingdoms on the northern coast of Java. Tomé Pires, author of the *Suma Oriental,* writing not long after 1511, stresses the obscure ethnic origins of the founders of Cheribon, Demak, Japara, and Gresik. These Javanese kingdoms existed to serve the commerce with the extensive Muslim world and especially with Malacca, an importer of Javanese rice. The rulers of Malacca, though of prestigious Palembang origin, had accepted Islām precisely in order to attract Muslim and Javanese traders to their port.

New men could now be expected to contribute impulses to Indonesian life. The northern Sumatran and Javanese coasts seem hitherto to have been on the fringe of the Śaivite-Mahāyāna cultures of southern Sumatra and eastern Java. For the first time in Indonesian history, the possibility existed that the inhabitants of formerly peripheral regions would begin to influence the course of events, inspired by Islām's assertion of the equality of all believers and supported by very profitable communications with the Muslim world throughout Asia.

But Indonesian history is the history of many distinct and often greatly separated regions. The history of early Indonesian Islām is no exception. What happened in the 15th and 16th centuries cannot be explained simply in terms of the influence of new ideas. The political ambitions of many regional princes intervened, and a variety of often rapidly changing and sometimes disturbed situations developed. The historian looks in vain for a uniform pattern of early Muslim life in the archipelago.

Aceh as a Muslim state

Aceh (Acheh), which succeeded Pasai in the 16th century as the leading harbour kingdom in northern Sumatra, became a self-consciously Muslim state, though a persuasive case has been made for the persistence as late as the 17th century of "Hindu" notions of divine kingship familiar in Java. Aceh had contacts with Muslim India and its own heterodox school of Muslim mysticism; its sultans sought an alliance with the Ottoman Turks against the Portuguese, who had conquered Malacca in 1511. The Malay princes of Malacca installed Muslim vassals on the east coast of Sumatra in the 15th century, but when Malacca was captured by the Portuguese the princes transferred their capital southward to Johore and gradually became involved in a conflict not only with the Portuguese but also with the Achinese for control of the Strait of Malacca. Aceh, for its part, was unable to impose its faith on the Batak highlanders in the interior. The single and notable gain for Islām in Sumatra was in the Minangkabau country, where Śaivite-Mahāyāna Tantric cults had flourished in the 14th century. Islām's penetration of Minangkabau by way of the Achinese west coast of Sumatra was far advanced by the beginning of the 17th century. Minangkabau, a land of enterprising and mobile traders, was later to exercise a significant influence in the affairs of the archipelago.

Muslims in Java. The Sumatran beachheads of Islām had commercial ties with other parts of the region, but they were not closely involved in events outside their immediate neighbourhoods. In Java, on the other hand, where the distance between the Muslim coastal fringe and the interior was negligible, a tense situation developed. The Muslims did not overthrow the kingdom of Majapahit (see above). Majapahit, weakened by feuds within its royal family and increasingly denied the benefits of overseas commerce, merely withered away and disappeared in the early 16th century. The passing of its hegemony left a power vacuum in Java that set in train a conflict between Islām and the aristocratic traditions of the interior.

In later centuries, the Javanese inland elite chose to bridge over the events of the 15th and 16th centuries and see a continuity between Majapahit and Mataram, the great kingdom of 17th-century Java. This vision of the past, however, conceals a very troubled period in Javanese history. The militant mood of coastal Islām may be seen in the enforced imposition of the new faith on western Java and also on Palembang in southern Sumatra. Similarly, the impact of Islām may be gauged by the fury of the 17th-century Mataram kings against the princes and Muslim notables of the northern coast.

The conflict seems to have begun with the determination of the Demak coastal rulers in the first half of the 16th century to rule over a great Javanese kingdom. The coastal princes, especially as their harbours grew richer and their dynasties older and more confident, came to see themselves not only as Muslim leaders but as Javanese princes. Their pretensions are reflected in Tomé Pires' statement that they cultivated the "knightly" habits of the ancient aristocracy. But when Demak sought to expand inland, bringing with it Islām, its armies were halted in the

mid-16th century by Pajang. Some years later, Mataram, another principality in central Java, came to the fore. The climax of the conflict was in the first half of the 17th century, when Agung, ruler of Mataram, took the offensive and destroyed the coastal states and with them the basis of Javanese overseas trade.

Links with Javanese tradition

It is unlikely that this bitter struggle was fought only for religious reasons. Islām came to Indonesia from India, perhaps from southern India, and the mood of heterodox mystic Ṣūfī sects of Islām was probably not foreign to the Javanese ascetics. Ṣūfī "saint" (*walī*) and Javanese guru eventually would have understood and respected each other's yearning for personal union with God. The Javanese tradition, in which small groups of disciples were initiated by a teacher into higher wisdom, was paralleled by the Ṣūfī teaching methods. For Muslim theologian and Javanese scholar alike the concern was always less with the nature of God than with skills for communicating with him. Arabic texts tended eventually to be recited as meditative aids, just as the Tantric mantras once had been.

The earliest Javanese disciples of Islām were, however, not the thoughtful representatives of earlier religious systems in Java but humble men of the coast who had been left outside the traditional teachings of the courts and the anchorites. These men doubtless saw in Islām a simple message of hope, offering them not only a congenial personal faith but also opportunities of secular advancement in a trading society where rank was not as important as fervour. Early Muslim literature has a theme of the wandering adventurer who comes from obscure origins, makes good, and seeks the consolations of Islām. For Muslim disciples such as these the times offered boundless means for achieving success, either in trade or in the service of ambitious princes. These princes, parvenu aristocrats and also the product of Islām, needed guardians of their conscience, courtly advisers, and, above all, military commanders. For the new elite the progress of coastal Islām brought both spiritual and material gain.

All of this must have been greatly disturbing to those in the interior who had been nurtured in older traditions and saw no reason for abandoning their Śaivite-Mahāyāna values. For the aristocrats of the interior, the memories of Majapahit's hierarchical system of government under a godlike king represented standards of civilized behaviour that must be asserted at all cost against the forces of confusion released by the coastal population. Contacts between wandering dervishes and the peasants, at a time of acute distress caused by warfare, and the pretensions of Muslim court officials, some of whom claimed a privileged religious status without precedent in Javanese history, must have seemed to threaten the foundations of society. The ruler of the interior kingdom of Pajang is depicted in the Javanese chronicles as an ascetic and as the son and grandson of ascetics. He was, in this respect, a true Javanese king. When, several generations later, the ruler of Mataram destroyed the coastal states he was seeking to destroy the forces that disunited Java. This was in the tradition of earlier Javanese kings. His conquests were as much a part of his mission as Kertanagara's had been in the 13th century.

Thereafter Islām was permitted to survive only on Javanese royal terms. Its innovating effects were postponed until the end of the 19th century. It was now one of several religious activities and therefore tolerable in Javanese eyes. Muslim officials in the court of Mataram became well-rewarded and obedient servants of the ruler. In time, scholars returned to the study of the earlier genre of Javanese literature, including texts that taught the nature of government according to the values of the "Hindu-Javanese" world. In the countryside, Islām remained influential in time of social distress, preaching to aggrieved peasants of the coming of the Messiah. As a literary influence Islām survived in the form of mystical texts and poems, romantic tales, and also in borrowings by later inland-court historians of material from the "Universal Histories" (*Sĕrat Kaṇḍa*) of the coastal culture. The borrowings are testimony of the impact of what had happened in the 15th and 16th centuries, which later historians could reinterpret but not ignore.

The history of 16th-century Java is still not fully understood, but Portuguese intervention seems to have been unimportant. The Portuguese survived chiefly as private traders, and, by the end of the century, the level of Muslim Indonesian trade with the Middle East, and thence with Europe, was greater than it had ever been. In the neighbourhood of the Strait of Malacca, Aceh and Johore were struggling for overlordship, and the scene in Java was being prepared for the final phase in the struggle between coastal Islām and the inland aristocracy. The outcome might have been the emergence of greater Indonesian unities under cover of Javanese claims to leadership. The situation was altered by the appearance of the Dutch at the end of the century. (O.W.W.)

GROWTH OF THE DUTCH EMPIRE

Dutch East India Company. The fall of Malacca to the Portuguese in 1511 is often taken as a turning point in Indonesian history. Over the next century Portuguese efforts were to be directed to securing control of the trade of the Spice Islands. At the end of the 16th century, Dutch and British interests in the region gave rise to a series of voyages: those of James Lancaster in 1591, Cornelis de Houtman in 1595 and again in 1598, Jacob van Neck in 1598, Lancaster again in 1601, and others. In 1602 the Dutch East India Company (formal name United East India Company [Vereenigde Oost-Indische Compagnie; VOC]) received its charter, two years after the formation of the English East India Company; it began to attempt to exclude European competitors from the Indies, control the trade carried on by indigenous Asian traders, and establish its own commercial monopoly.

Establishment of the VOC

Monopoly itself was not, of course, an innovation. Aceh, for example, had controlled trade on the northwest and east coasts of Sumatra. The company's monopoly, however, was more extensive and came to form the basis of the Dutch territorial empire. For these reasons many historians have tended to see 1511 or 1600 as the beginning of a period of European domination lasting until the 20th century.

Since the 1930s, however, some historians have criticized the view of Indonesian history that judges Europeans to have been the major factor in shaping the history of the Indies from the 17th century onward. By contrast, they have stressed an essential continuity of Indonesian history and have argued that the VOC at first made little change in traditional political or commercial patterns. Traditional Asian commerce, according to one view, was a noncapitalistic peddling trade, financed by patrician classes in Asian countries and conducted by innumerable small traders who collected spices and pepper in the Indies for disposal in the port cities of Asia. In this view the VOC was seen, in effect, as merely another merchant prince, gradually inserting itself into the existing trade patterns of the Spice Islands and accommodating itself to them. As Batavia became the headquarters from which it established factories in the Spice Islands and elsewhere, the company gradually became a territorial power but was, at first, only one power among others and not yet ruler of the Indies. Only during the 19th century did new economic forces, the product of industrial capitalism, burst upon the Indies and submerge them under a new wave of European imperialism.

The theory is an overstatement. If the coming of the Europeans did not represent a sharp break in the continuity of Indonesian history, it did, at least, initiate changes that, in the long run, were to be of enormous importance. The VOC itself represented a new type of power in the Indies: it formed a single organization, traded across a vast area, possessed superior military force, and, in time, employed a bureaucracy of servants to look after its concerns in the Indies. In sum, it could impose its will upon other rulers and force them to accept its trading conditions. Under the governor-generalship of Jan Pieterszoon Coen and his successors, particularly Anthony van Diemen (1636–45) and Joan Maetsuyker (1653–78), the company laid the foundations of the Dutch commercial empire and became the paramount power of the archipelago.

The process was gradual. Historians now have tended to bypass the indigenous-external antithesis and to refer

instead to a general commercial expansion throughout Southeast Asia from the 15th to 17th century—an "age of commerce" involving native as well as foreign participants. During this period the VOC did go far toward establishing its commercial control in the Indies. It captured Malacca from the Portuguese (1641), confined the British, after a period of fierce rivalry, to a factory at Bencoolen in southwestern Sumatra, and established a network of factories in the eastern islands. Though it may have wished to limit its activities to trade, the company was soon drawn into local politics in Java and elsewhere, and, in becoming the arbiter in dynastic disputes or in conflicts between rival rulers, it inevitably emerged as the main political entity in the islands.

Sultan Agung

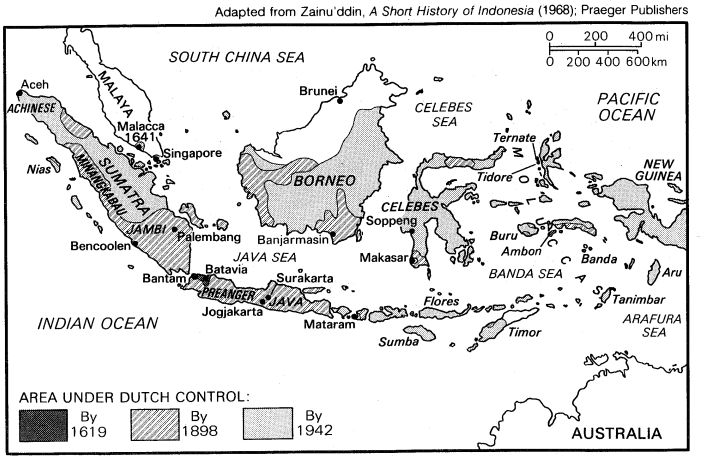
In the 1620s Sultan Agung, ruler of the central Javanese kingdom of Mataram and representative of the old and highly sophisticated Javanese civilization, sought to extend his power over Bantam in western Java. This brought him into conflict with the Dutch, and he laid siege to the Dutch fortress at Batavia. Though Agung's forces were eventually compelled to withdraw, the result of the confrontation was inconclusive and left both the Dutch and Javanese warily respectful of each other's strength. But during the following century internal dissensions in Mataram led to increasing Dutch involvement, and in the early 18th century a series of wars of succession among pretenders to the throne of Mataram hastened the process. In return for its services in 1674 to Amangkurat I, Sultan Agung's successor, and to his successor, Amangkurat II, shortly afterward, the VOC received the cession of the Preanger regions of western Java.

This was the first of a series of major territorial advances. In 1704 Dutch forces assisted in replacing Amangkurat III with his uncle, Pakubuwono I, in return for which further territory was ceded. In this way almost all of Java gradually passed under Dutch control, and by 1755 only a remnant of the kingdom of Mataram remained. This was divided into two principalities, Jogjakarta and Surakarta, which survived until the end of Dutch rule. In an attempt to control the pepper trade in Sumatra, the VOC established footholds in western Sumatra and in Jambi and Palembang over the course of the 17th century, and it interfered in local conflicts in support of rulers who favoured it; but the main Dutch expansion there did not take place until the 19th century.

In acquiring territorial responsibilities, the company did not at first establish a close administrative system of its own in the areas that passed under its direct control. In effect, the VOC replaced the sovereign of the royal court and, in so doing, inherited the existing structure of authority. An indigenous aristocracy administered the collection of tribute on behalf of the company, and only gradually was this system converted into a formalized bureaucracy. The VOC, like the royal court before it, drew revenue in the form of produce from the peasantry within its domain.

To implement its commercial monopoly, the VOC established company factories (trading posts) for the collection of produce, pressured individual rulers to do business solely with the company, controlled the sources of supply of particular products (clove production, for example, was limited to Ambon, nutmeg and mace to the Banda Islands) and, in the 18th century, pushed through a system of so-called forced deliveries and contingencies. Contingencies constituted a form of tax payable in kind in areas under the direct control of the company; forced deliveries were produce that native cultivators were compelled to grow and sell to the company at a set price. There was little difference between the devices. In theory, forced deliveries were thought of as a form of trade in which goods were exchanged, but they were, in fact, as the British scholar J.S. Furnivall has described it, "tribute disguised as trade," while contingencies were "tribute undisguised." In effect, the whole system of company trade was designed to extract produce from the Indies for disposal on a European market, but without stimulating any fundamental technological change in the area's economy. The profits belonged to the company, not to the producers. The indigenous traders of the region were pushed aside by the VOC as it gained control of more and more of the export trade of the archipelago. The growth of Batavia resulted, for example, in the decline of the north coast ports of Java, through which much of the spice trade had been channeled since before the 15th century. In this way the traditional pattern of trade was checked and distorted.

Adapted from Zainu'ddin, *A Short History of Indonesia* (1968); Praeger Publishers



The expansion of Dutch control in Indonesia.

Decline of the VOC

During the 18th century, the VOC ran into financial difficulties from a variety of causes: the breach of the company's monopoly by "smuggling," the growing administrative costs as the company came to shoulder greater responsibilities of government, and the corruption of the company's servants. In 1799 the Dutch government of the Batavian Republic wound up the affairs of the company.

The French and British in Java, 1806–15. The fall of the Netherlands to France during the Revolutionary Wars and the dissolution of the company led in due course to significant changes in the administration of the Indies. Under Napoleon's "Kingdom of Holland," one of his marshals, Herman Willem Daendels, was appointed as governor-general. Daendels strengthened Javanese defenses, raised new forces, built new roads within Java, and improved the internal administration of the island. He attempted to formalize the position of the Javanese regents, subordinating them to Dutch prefects and emphasizing their character as civil servants of a central government rather than as semi-independent local rulers.

In 1811 Java fell to a British East India Company force under Lord Minto, governor-general of India, who, after the surrender, appointed Thomas Stamford Raffles as lieutenant governor. Raffles approached his task in the conviction that British administrative principles, modeled in part on those developed in Bengal, could liberate the Javanese from the tyranny of Dutch methods; he believed that liberal economic principles, by ending compulsory cultivation, could simultaneously expand Javanese agricultural production, improve revenue, and make the island a market for British goods. Along with his doctrinaire liberalism, he brought to his task a respect for Javanese society. Before his appointment, he had been a student of Malay literature and culture, and during his period in Batavia he encouraged the study of the society he found about him. Raffles rediscovered the ruins of the great Buddhist temple Borobudur in central Java and published his *History of Java* in 1817, a year after his return to England.

Raffles carried further the administrative centralization begun by Daendels and planned to group the regencies of Java into 16 residencies. By declaring all lands the property of the government and by requiring cultivators to pay a land rent for its use, he proposed to end the compulsory production system. This, he believed, would free the peasants from servility to their "feudal" rulers and from the burden of forced deliveries to the Dutch and allow them to expand their production under the stimulus of ordinary economic motives. Unfortunately, Raffles oversimplified the complexities of traditional land tenure. He misread the position of the regents, whom he at first wrongly believed to be a class of feudal landholders rather than an official aristocracy. (The regents, in fact, had no proprietary rights in the land of their subjects.) He was concerned to replace

Failure of Raffles' reforms

what he saw as a tribute system, paid in the form of forced deliveries, by the payment of a fixed and regular rent that would leave the landholders more free to enjoy the fruits of their enterprise than they had been in the past. But despite a series of adjustments in his original plan, Raffles failed to devise an effective means of applying his theories before the return of Java to Dutch hands as part of the general settlement following the defeat of Napoleon.

DUTCH RULE FROM 1815 TO C. 1920

Before the 19th century, Indonesian societies had experienced considerable pressure from Europeans, but they had not been submerged by Western influences. The political order of Mataram had been eroded, and the first steps had been taken toward administrative centralization in Java. In the outer islands, local rulers had been forced to submit in some measure to the will of Batavia. The trading patterns of the archipelago had been changed and constricted. Nevertheless, these were superficial developments when seen against the continuing coherence and stability of Indonesian societies. They were superficial, also, compared with the Western impact still to come.

When the Dutch returned to Indonesia after the Napoleonic Wars, their main concern was to make the colony at least self-supporting. During the interregnum, both exports and revenue had declined sharply, despite Raffles' hopes for his land-rent system. The costs of government in Java were rising as a result of the growing complexity of administration. In restoring their authority, the Dutch retained the main outlines of the system of residencies, regencies, and lower administrative divisions, though they did not, at first, follow exactly the attempts of Daendels and Raffles to turn the regents into salaried officials, specifically responsible to the residents. Rather, they saw the regent as the "younger brother" of the resident. This difference in theory was perhaps of slight practical effect, since the tendency in lower levels of territorial administration continued in the direction of an increasingly centralized control. Several factors contributed to the trend: one was the need to deal with a series of disturbances, particularly in Java and western Sumatra, but also on a lesser scale in Celebes, Borneo, and the Moluccas; a second was the new economic policy, adopted in 1830, which placed new economic responsibilities on local officials.

The Java War

The Java War of 1825–30 sprang from a number of causes. In part, it was the product of the disappointed ambitions of its leader, Prince Diponegoro, who had been passed over for the succession to the throne of Jogjakarta. In part, it sprang from resentment among the aristocratic landholders of Jogjakarta, whose contracts for the lease of their lands to Europeans had been canceled by the governor-general. There was support, too, from Islāmic leaders. And there were also, no doubt, hidden factors of the kind often to be found in cases of agrarian protest in Java—factors such as the messianic expectation of the coming of a Just Ruler who would restore the harmony of the kingdom. From these varied causes there sprang a revolt that, through the skillful use of guerrilla tactics, continued to challenge Dutch authority for five years, until the Dutch treacherously seized Diponegoro during truce negotiations and exiled him to Celebes.

About the same time, the Dutch in western Sumatra were drawn into the so-called Padri War (named for Pedir, a town in Aceh through which Muslim pilgrims usually returned home). Basically, this was a religious struggle between revivalist Islāmic leaders in Minangkabau and the adat (customary law) leaders of the community. Under Tuanku Imam Bonjol, the Padri forces resisted Dutch pressure from the early 1820s until 1837. The effect of this involvement was inevitably to strengthen the Dutch administrative commitment in western Sumatra.

The Cultivation System. The formation in 1824 of the Netherlands Trading Society (Nederlandsche Handel-Maatschappij; NHM), a company embracing all merchants engaged in the Indies trade and supported by the Netherlands government with the king as its chief shareholder, did not produce the hoped-for commercial expansion. In 1830, however, a newly appointed governor-general, Johannes van den Bosch, devised a new method by which the government could tap the resources of the Indies. This was the so-called Cultivation (or Culture) System (Cultuurstelsel).

The Cultivation System provided that a village set aside a fifth of its cultivable land for the production of export crops. These crops were to be delivered to the government in lieu of tax. Land rent was to continue at the same time as a complementary part of the system and as a measure of the amount to be produced by each village. Thus, if a village, through the growing of export crops on a fifth of its land, returned an amount in excess of the land rent for which it had been assessed, it would be free of land rent and would be reimbursed to the extent of the excess; on the other hand, if a village produced less than the assessed amount of land rent, it would have to make up the difference.

Economic and social effects of the system

From the government's point of view, the Cultivation System was an overwhelming success. Exports soared, rising from 13 million guldens in 1830 to 74 million a decade later. The products were disposed of through the Netherlands Trading Society, and between 1840 and 1880 their sale brought to the Dutch treasury an annual average of 18 million guldens, approximately a third of the Dutch budget. The effects of the system for the Javanese were, however, of more dubious value. Though its founder believed that, by stimulating agricultural production, the Cultivation System would ultimately benefit the people of Java as well as the home government, it came to be considered, in later years, both by Dutch critics and by outside observers, as a particularly harsh and burdensome policy. Van den Bosch's expectations were not entirely false. The policy did extend village production in certain areas, and the population of Java increased from 6 million to 9.5 million during the full operation of the system. The range of exports from Java was extended. Indigo and sugar were the first items to be made the subject of compulsory cultivation; coffee, tea, tobacco, and pepper were subsequently added. Nevertheless, the system placed a heavy burden on the cultivators and tended to accentuate social and economic differences within rural society. Dominant peasants, members of a rural elite, were able to manipulate the system to their advantage. And while the Cultivation System brought the Indies into contact with a wider overseas market, the Indies government stood between producer and market, and the annual surplus added to Dutch, not Javanese, prosperity. The system did nothing to stimulate technological change or economic development for the Javanese people. An increasing commercial role was played not by the indigenous population but by the Chinese who fitted into colonial rule as a separate caste, engaged in tax farming, moneylending, and small trading.

There were other consequences. The Cultivation System accentuated the differences between Java and the outer islands, and in Java it led to a considerable tightening of the administrative system. The regent became the kingpin of the system, responsible to the resident for the delivery of crops from his regency. In some cases, regents, secure in the knowledge they were backed by Dutch power, imposed additional burdens upon their subjects—a development that received trenchant criticism in the novel *Max Havelaar* (1860), written under the pseudonym Multatuli by the Dutch writer Eduard Douwes Dekker, a former official of the Indies government. But the long-term effect of the new functions imposed on regents was to reduce their independence and to hasten the process, started by Daendels, by which a loosely structured administrative aristocracy was gradually converted into a salaried civil service. Regents were no longer able to draw their revenues from their subjects, and the lines of authority were clearly drawn. Regents, aided by a junior Dutch official (the *controleur*), became clearly responsible to the Dutch residents above. By 1860 the administrative divisions of Java had been firmly established, and the service that staffed them had acquired the character it was essentially to preserve for the remainder of the colonial period.

In the 1860s the Cultivation System came under attack not only from humanitarian quarters but also from private business interests in The Netherlands. The latter appealed

The Liberal Policy

to liberal economic principles in support of their right to share in the riches of the Indies; and their pressure was effective. Though the Cultivation System was not abolished and continued for a number of years to make its contribution to the Dutch treasury, the decision was taken to encourage also the entry of private capital. The Liberal Policy, as it was called, was effectively inaugurated in 1870 by the adoption of an agrarian law that provided that European investors could acquire land under long-term leasehold, either from Indonesian landholders, or in the case of unoccupied land, from the government. Certain safeguards were provided for the Indonesian landholder: the provision that Europeans lease, rather than purchase, land was intended to prevent the alienation of Indonesian land, and the government was charged with the responsibility also of preventing Europeans from leasing land needed for the subsistence of village populations.

Within this framework Dutch capital began to flow to the Indies on a scale that was to transform the character of the Indonesian economy and society. During the next 60 years there was a 10-fold increase in the value of exports (from 107 million guldens to 1.16 billion). There was a change, also, in kinds of products exported. Such exports as coffee, sugar, tea, and tobacco continued to expand; but such industrial raw materials as rubber, copra, tin, and oil soon came to dominate the export economy. These remarkable developments were, in large measure, the product of a totally different system of production. Under the company, during the interregnum, and, later, under the Dutch crown working through the Cultivation System, export crops were grown by Indonesian cultivators on their own land. Under the Liberal Policy, however, the new crops were the subject of estate production. Much economic expansion took place in Sumatra rather than Java, and Sumatra's east-coast residency became the seat of a vast new plantation economy. The estates were company-owned, and the economic developments of the late 19th century were indeed the product of corporate, rather than individual, enterprise.

Dutch territorial expansion. Rapid economic development was accompanied by territorial expansion. Though the Dutch had established their control effectively over Java by the mid-18th century and though they had gradually expanded their original holdings in Sumatra over the course of the 19th, their control over the rest of the archipelago was patchy and incomplete. It was exercised, in the main, through agreements with local rulers rather than through direct control over territory. In the closing years of the 19th century and the early years of the 20th rapid moves were made to round out the Dutch empire and extend it effectively over the whole of the Indies. In northern Sumatra, warfare with the people of Aceh lasted with varying degrees of intensity from 1873 to 1908 and brought the northern tip of Sumatra under Dutch control. In Celebes and the Moluccas, where the Dutch had for long exercised a general authority, a new instrument—the Short Declaration (in contrast to the earlier Long Contract)—bound local rulers to accept the control of Batavia. Dutch authority was extended in this way over Bone and Luwu in the Celebes, over central Borneo, over Bali and the Lesser Sunda Islands, and over Ternate, Ceram, and Buru in the Moluccas. Footholds were established also over parts of West New Guinea. Communications were developed—roads and railways in Java and Sumatra and expanded shipping services to link Java to the outer islands—to serve the needs of the new plantation economy. Between 1870 and 1910 the Dutch had thus effectively completed the process of converting the Indies into a unified colonial dependency and, indeed, of laying the foundations of the future Indonesian republic.

The "new imperialism" of the late 19th century may be seen as part of a worldwide movement whereby the industrial countries of western Europe partitioned among themselves the hitherto undeveloped areas of the globe. In Africa and in the South Pacific, in Burma (Myanmar), Indochina, and Malaya, as well as in Indonesia, a new "forward movement" was taking place that stood in dramatic contrast to the earlier patterns of commercial empire. If the European presence created a watershed in Indonesian history, it is to be discerned about 1870 rather than in 1600.

Impact of the "new imperialism"

The social impact of these developments upon Indonesian society was tremendous. The economic and political expansion brought a new Dutch population to the Indies: civil servants to staff the growing services of government, managers to run the new estates, and clerks to staff the import-export houses and other businesses. These came to form a European enclave within the major cities and accentuated the lines of social division in what was increasingly a caste society divided along racial lines. The Dutch, however, were never a purely expatriate community whose members were anxious to retire as soon as possible to The Netherlands. Many of them regarded the Indies as their home. Their sense of belonging was very different, for example, from that of the British in India, and it was to give an added bitterness to the later struggle to retain the colony after World War II. From the Indonesian point of view, the growing cities became the home of a new urban way of life and stimulated social change. A new elite emerged under the influence of the expanding Western impact. So did a new class of unskilled and semiskilled workers who found employment as domestic servants or as labourers in the light industries that began to develop. Rural society, though more sheltered, was also altered by the currents of change. Although the agrarian law and the later labour legislation had provisions to protect existing customary rights over land and to guarantee fairness of contracts for labourers, the mere fact of contract employment on the estates affected the village society from which workers were drawn and played its part in hastening the growth of a rootless and disoriented population, divorced increasingly from the shelter of traditional village society but not absorbed into the new urban culture.

The Ethical Policy. Liberals confidently assumed that, just as freedom of enterprise would maximize welfare at home, so the application of European capital to the task of developing colonial resources would gradually improve the lot of colonial peoples. By the end of the 19th century, the 30 years of the Liberal Policy in Indonesia did not appear to have achieved that miracle. By the end of the century, growing criticism of the Dutch record in the Indies was given particularly influential expression by C.T. van Deventer, a Liberal Democratic member of the States General, who argued that The Netherlands had been draining wealth from the Indies and had incurred thereby a "Debt of Honour" that should be repaid. His suggestion was that The Netherlands should turn from its strictly laissez-faire policy in the Indies and pursue instead a positive welfare program supported by funds from the metropolitan treasury. In 1901 a change of government in The Netherlands provided the opportunity for a new departure in policy along the lines suggested by van Deventer. According to the Ethical Policy, as it was called, financial assistance from The Netherlands was to be devoted to the extension of health and education services and to the provision of agricultural extension services designed to stimulate the growth of the village economy.

Van Deventer's reforms

The Ethical Policy was seen by its most fervent supporters as a noble experiment designed to transform Indonesian society, to enable a new elite to share in the riches of Western civilization, and to bring the colony into the modern world. Its ultimate goals were, of course, not clearly defined. Van Deventer looked to the emergence of a Westernized elite who would be "indebted to the Netherlands for its prosperity and higher Culture" and who would gratefully recognize the fact. Others hoped for the growth, by "cultural synthesis," of a new East Indian society based on blending of elements of Indonesian and Western cultures and able to enjoy a large measure of autonomy within the framework of the Dutch empire.

Despite these rather grandiose visions, the achievements of the Ethical Policy were much more modest. It neither checked declining living standards nor promoted an agrarian revolution. It did, however, provide agricultural assistance and advice, but this was directed to the improvement of techniques of irrigation and cultivation within the existing wet-rice technology of Java. Its effect, therefore, was to confirm the gulf between the European economy

of the estates, mines, oil wells, and large-scale commerce and the traditional, largely subsistence, Indonesian economy of wet-rice or shifting cultivation. In education, a little was done to provide a greater degree of opportunity at primary, secondary, and even tertiary levels, but at the end of the 1930s only a handful of high school graduates was produced locally, and the literacy rate was calculated at just over 6 percent.

The goals of the Ethical Policy were set too high, and the devices adopted to implement them were too modest. Given the inertia of traditional societies, it was not to be expected that a new order would be created as easily as the proponents of the policy had hoped. Nevertheless, during the years of its operation, the Indies did see the release of tremendous forces of social change. These resulted, however, not from the conscious plans of the Ethical Policy but from the undirected force of Western economic development. Java's population, which had risen from about 6 million to almost 30 million over the course of the 19th century, increased to more than 40 million by 1920. The population increase, together with urbanization, the penetration of a money economy to the village level, and the labour demands of Western enterprise combined to disrupt traditional patterns. Where the Ethical Policy was most effective, despite the limitations of its educational achievement, was in producing a small educated elite who could give expression to the frustration of the masses in a society torn loose from its traditional moorings. Western currents of thought had their impact also within Islāmic circles, where modernist ideas sought to reconcile the demands of Islām and the needs of the 20th century. It is against this background that a self-conscious nationalist movement began to develop.

TOWARD INDEPENDENCE

The rise of nationalism. Indonesian nationalism in the 20th century must be distinguished from earlier movements of protest: the Padri War, the Java War, and the many smaller examples of sporadic agrarian unrest had been "prenationalistic" movements, the products of local grievances. By contrast, the nationalism of the early 20th century was the product of the new imperialism and was part of wider currents of unrest affecting many parts of Africa and Asia. In Indonesia nationalism was concerned not merely with resistance to Dutch rule but with new perceptions of nationhood, embracing the ethnic diversity of the archipelago and looking to the restructuring of traditional patterns of authority in order to enable the creation of Indonesia as a modern state. It derived in part from specific discontents, the economic discriminations of colonial rule, the psychological hurt arising from the slights of social discrimination, and a new awareness of the all-pervading nature of Dutch authority. Important, too, was the emergence of the new elite, educated but lacking adequate employment opportunities to match that education, Westernized but retaining still its ties with traditional society.

Formation of the nationalist movement

The formation in 1908 of Budi Utomo ("High Endeavour") is often taken as the beginning of organized nationalism. Founded by Wahidin Sudirohusodo, a retired Javanese doctor, Budi Utomo was an elitist society, the aims of which—though cultural rather than political—included a concern to secure an accommodation between traditional culture and the modern world. Numerically more important was Sarekat Islām ("Islāmic Association"), founded in 1912. Under its charismatic chairman, Omar Said Tjokroaminoto, the organization expanded rapidly, claiming a membership of 2,500,000 by 1919. Later research suggests that the real figure was likely to have been no more than 400,000, but even with this greatly reduced estimate Sarekat Islām was clearly much larger than any other movement of the time. In 1912 the Indies Party (Indische Partij)—primarily a Eurasian party—was founded by E.F.E. Douwes Dekker; banned a year later, it was succeeded by another Eurasian party, Insulinde. In 1914 the Dutchman Hendricus Sneevliet founded the Indies Social Democratic Association, which became a communist party in 1920 and adopted the name Indonesian Communist Party (Partai Komunis Indonesia; PKI) in 1924.

By the end of World War I there was, thus, a variety of organizations in existence, broadly nationalist in aim, though differing in their tactics and immediate goals and in the sharpness of their perceptions of independent nationhood. In the absence of firm party discipline, it was common for individuals to belong simultaneously to more than one organization and, in particular, the presence of Indonesian Social Democratic Association members in Sarekat Islām enabled them to work as a "bloc within" the larger movement. The idea that the time was not yet ripe for communist parties to assume independent leadership of colonial nationalism later led the Comintern to formulate the strategy of cooperation with anti-imperialist "bourgeois" parties.

At the end of World War I the Dutch, in an effort to give substance to their promise to associate the Indonesian community more closely with government, created the People's Council (Volksraad). Composed of a mixture of appointed and elected representatives of the three racial divisions defined by the government—Dutch, Indonesian, and "foreign Asiatic"—the People's Council provided opportunities for debate and criticism but no real control over the government of the Indies. Some nationalist leaders were prepared to accept seats in the assembly, but others refused, insisting that concessions could be obtained only through uncompromising struggle.

In 1921 the tension within Sarekat Islām between its more conservative leaders and the communists came to a head in a discipline resolution that insisted that members of Sarekat Islām belong to no other party; this, in effect, expelled the communist "bloc within," and there followed a fierce rivalry between the two for control of the grassroots membership of the organization. The PKI, once it had committed itself to independent action, began to move toward a policy of unilateral opposition to the colonial regime. Without the support of the Comintern, and even without complete unanimity within its own ranks, it launched a revolt in Java at the end of 1926 and in western Sumatra at the beginning of 1927. These movements, which had elements of traditional protest as well as of genuine communist insurrection, were easily crushed by the Indies government, and communist activity was effectively ended for the remainder of the colonial period.

The Bandung "study club"

The defeat of the communist revolt and the earlier decline of Sarekat Islām left the way open for a new nationalist organization, and in 1926 a "general study club" was founded in Bandung, with a newly graduated engineer, Sukarno, as its secretary. The club began to reshape the idea of nationalism in a manner calculated to appeal to Indonesia's new urban elite. After the failure of the ideologically based movements of Islām and communism, nationalist thinking was directed to the idea simply of a struggle for independence, without any precommitment to a particular political or social order afterward. Such a goal, it was believed, could appeal to all, including Muslims and communists, who could, at least, support a common struggle for independence, even if they differed fundamentally about what was to follow. Nationalism, in this sense, became the idea that the young Sukarno used as the basis of his attempt to unify the several streams of anticolonial feeling. The ideas of the Bandung Study Club were reinforced by currents of thought emanating from Indonesian students in Holland. Their organization, reorganized in 1922 under the name of the Indonesian Union (Perhimpunan Indonesia), became a centre of radical nationalist thought and, in the mid-1920s, students returning from Holland joined forces with like-minded groups at home.

The new nationalism required a new organization for its expression, and in July 1927 the Indonesian Nationalist Association, later the Indonesian Nationalist Party (Partai Nasional Indonesia; PNI), was formed under the chairmanship of Sukarno. The PNI was based on the idea of noncooperation with the Indies government and was thus distinguished from those groups, such as Sarekat Islām, that were prepared to accept People's Council membership. Sukarno, however, while seeking to create a basis of mass support for the PNI, also attempted with some success to work together with more moderate leaders and

succeeded in forming a broadly based, if rather precarious, association of nationalist organizations.

Arrest of Sukarno

At the end of 1929 Sukarno was arrested with some of his colleagues and was tried, convicted, and sentenced to four years' imprisonment. He was released at the end of 1931, but by then the united movement he had helped to create had begun to disintegrate. The PNI dissolved itself and reformed as Partindo. A number of other groups came to join in a new organization, the Indonesian National Education Club, known as the New PNI. While Partindo saw itself as a mass party on the lines of the old PNI, the New PNI, under the leadership of Mohammad Hatta and Sutan Sjahrir, aimed at training cadres who could maintain a continuing leadership of the movement and who could thus prevent it from being so easily immobilized by the arrest of its leaders.

In 1933 Sukarno was arrested again and exiled to Flores and later to Bencoolen (Bengkulu) in southern Sumatra. Repressive action followed against other party leaders including Hatta and Sjahrir, who were also exiled. In the later 1930s nationalist leaders were forced to cooperate with the Dutch, and such moderate parties as Parindra accepted People's Council membership. In 1937 a more radical party, Gerindo, was formed, but it considered support of The Netherlands against the threat of Nazism more important than the question of independence.

War in Europe and the Pacific changed the situation. The fall of the Indies to the Japanese onslaught early in 1942 broke the continuity of Dutch rule and provided a completely new environment for nationalist activity.

Japanese occupation. Japanese military authorities in Java, having interned Dutch administrative personnel, found it necessary to use Indonesians in many administrative positions, thus giving them opportunities that had been denied them under the Dutch. In order to secure popular acceptance of their rule, the Japanese sought also to enlist the support of both nationalist and Islāmic leaders. Under this policy Sukarno and Hatta both accepted positions in the military administration.

Though initially welcomed as liberators, the Japanese gradually established themselves as harsh overlords. Their policies fluctuated according to the exigencies of the war, but, in general, their primary object was to make the Indies serve Japanese war needs. Nationalist leaders, however, felt able to trade support for political concessions. Sukarno was able to convince the administration that Indonesian support could only be mobilized through an organization that would represent genuine Indonesian aspirations. In March 1943 such an organization, Putera (Pusat Tenaga Rakjat; "Centre of the People's Power") was inaugurated under his chairmanship. While the new organization enabled Sukarno to establish himself more clearly as the leader of the nation and while it enabled him to develop more effective lines of communication to the people, it also placed upon him the responsibility of trying to sustain Indonesian support for Japan through, among other things, the *romusha* (forced-labour) program. Later in the year Indonesian opinion was given a further forum in a Central Advisory Council and a series of local councils. At a different level, Indonesian youths were able to acquire a sense of corporate identity through membership in the several youth organizations established by the Japanese. Of great importance, also, was the creation in October 1943 of a volunteer defense force composed of and officered by Indonesians trained by the Japanese. The Sukarela Tentara Pembela Tanah Air (Peta) was to become the core of the republic's army during the revolution.

In March 1944 the Japanese, feeling that Putera had served Indonesian rather than Japanese interests, replaced it with a "people's loyalty organization" (Djawa Hokokai), which was kept under much closer control. Six months later the Japanese premier announced the Japanese intention to prepare the Indies for self-government. In August 1945, on the eve of the Japanese surrender, Sukarno and Hatta were summoned to Saigon, Vietnam, where Terauchi Hisaichi, commander of Southeast Asia, promised an immediate transfer of independence.

On their return to Djakarta (Jakarta; formerly Batavia), Sukarno and Hatta were under pressure to declare independence unilaterally. This pressure reached its climax in the kidnapping of the two men, for a day, by some of Djakarta's youth leaders. After the news of the Japanese surrender had been confirmed, Sukarno proclaimed independence on the morning of Aug. 17, 1945.

Proclamation of independence

The revolution. The proclamation touched off a series of risings across Java that convinced the British troops entrusted with receiving the surrender of Japanese forces that the self-proclaimed republic was to be taken seriously. At the level of central government, the constitution adopted by republican leaders was presidential in form, but a widely representative Central Indonesian National Committee became, in effect, an ad hoc parliament. Sukarno, as president, agreed to follow parliamentary conventions by making his cabinets dependent upon their ability to command the committee's confidence.

The spontaneous character of the Indonesian Revolution was demonstrated by a number of incidents, notably in the struggle for Bandung in late 1945 and early 1946 and in the Battle of Surabaya in November 1945, in which Indonesian fighters resisted superior British forces for three weeks. Though the Dutch had expected to reassert their control over their colony without question and though they were able to play upon outer-island fears of the Java-based republic, they eventually were compelled to negotiate with republican representatives led by Sjahrir, who by then was prime minister. The Linggadjati Agreement (1946–47), by which the Dutch agreed to transfer sovereignty in due course to a federal Indonesia, appeared to offer a solution to the conflict. (The Dutch claimed that a federation was necessary because of the diversity of the Indies and the difference between heavily populated Java and the more sparsely populated outer islands.) Differing interpretations, however, made the agreement a dead letter from the beginning. In July 1947 the Dutch, in an attempt to settle matters by force, initiated what they termed a police action against the republic. Its effect was to evoke UN intervention in the form of a Good Offices Committee, and it ended in the precarious Renville Agreement of January 1948. In December 1948 a second police action was launched.

Meanwhile, the government of the republic faced some domestic opposition. In 1946 a left-wing plot was organized by followers of Ibrahim Datuk Tan Malaka, who opposed the policy of negotiation with the Dutch. This so-called July 3rd Affair was easily crushed. In September 1948 a more serious challenge in the form of a communist revolt (the Madiun Affair) was also defeated.

The second police action aroused American concern. It also closed Indonesian ranks firmly behind the republic. In these circumstances, The Netherlands, at a roundtable conference at The Hague, finally agreed in August 1949 to transfer sovereignty over its colony (with the exception of western New Guinea) to an independent United States of Indonesia in December 1949; a decision about the ultimate fate of western New Guinea (Irian Barat, now Irian Jaya) was to be the subject of future negotiation.

INDEPENDENT INDONESIA TO 1965

The years of constitutional democracy. The initial federal constitution of 1949 was replaced in 1950 by a unitary but still provisional constitution. It was parliamentary in character and assigned an essentially figurehead role to the president. From the revolutionary period, Indonesia had inherited a multiparty system. The main parties after independence were the major Muslim party, Masyumi (Masjumi); the Muslim theologians' party, Nahdatul Ulama (NU), which seceded from Masyumi in 1952; the Nationalist Party (PNI); the Communist Party (PKI); the "national communist" party, Murba; the lesser Muslim parties, Perti and Partai Sarekat Islām Indonesia (PSII); and the Socialist Party (PSI). Until the first elections were held, in 1955, Parliament was filled by appointment under a gentlemen's agreement between parties as to their probable electoral strengths. The elections of 1955, a remarkable and technically successful experiment in the exercise of political choice by a largely nonliterate population, confirmed the position of Masyumi, NU, PNI, and PKI as the country's four leading political parties.

Party alignments

With the exception of the PKI, the parties did not represent clearly opposing interests or programs, though some broad bases of support could be seen. The PNI was particularly strong in the ranks of the civil service, while Masyumi tended to find its support in market towns and among the trading classes; NU was stronger in rural areas. The PSI, an influential party until it was virtually eliminated in the elections, had strong support in the higher ranks of the army and bureaucracy. Also important was the regional distribution of party strengths. The PNI, NU, and PKI were essentially Java-based parties, while Masyumi drew most of its strength from outside Java, particularly in western Sumatra and southwestern Celebes (Sulawesi). Its support within Java was to be found mainly in Jawa Barat (West Java) province, the home of the Sundanese and not of the ethnic Javanese. This unevenness in party strengths meant that political rivalry in the early years of independence tended to have a regional flavour, a fact that was of importance when regional resistance to the centre reached the point of open revolt in 1958. In simplified terms, it is possible to see, in the regional distribution of party strengths, a broad opposition between the hierarchical, rice-based society of Java and the more strongly Muslim areas where commerce rather than agriculture has been the basis of historical development. Any interpretation of political conflicts in Indonesia must take account also of the extent to which the parties and their suborganizations reflected major cultural streams (*aliran*) in Indonesian society rather than interests, classes, or even regions. In addition to Masyumi's suspicion of the Javanese parties, the division within Java between *santri* (devout Muslims) and *abangan* (reflecting an earlier, pre-Muslim syncretism) is important in understanding the rivalry of the NU and Masyumi, on the one hand, and the PNI and the PKI, on the other.

Rapid government turnover

In the early and mid-1950s there was a rapid succession of governments—Hatta (December 1949–August 1950), Mohammad Natsir (September 1950–March 1951), Sukiman Wirjosandjojo (April 1951–February 1952), Wilopo (April 1952–June 1953), Ali Sastroamidjojo (July 1953–July 1955), Burhanuddin Harahap (August 1955–March 1956), and Ali's second government (March 1956–March 1957). This instability created a growing disillusionment with the fruits of independence and a sense of contrast between the heroism of the revolution and the self-seeking party rivalry that followed it. In particular, conflict between the export-producing outer islands and the heavily populated island of Java was becoming more marked. In December 1956 these factors of discontent led to movements of regional dissidence, supported by local military commanders, in western Sumatra, the Minahasa Peninsula of northern Celebes, and elsewhere.

Introduction of Guided Democracy. Against this background, Sukarno, resentful of his circumscribed position as figurehead president, began to move toward a greater interference with constitutional processes. In February 1957 he announced his own "Concept" for Indonesia. Criticizing Western liberal democracy as unsuited to Indonesian circumstances, he called for a political system of "democracy with guidance" based on indigenous procedures. The Indonesian way of deciding important questions, he argued, was by way of prolonged deliberation (*musyawarah*) designed to achieve a consensus (*mufakat*). This was the procedure at the village level, and it should be the model for the nation. He proposed a government based on the four main parties plus a national council representing not merely political parties but functional groups—workers, peasants, intelligentsia, national entrepreneurs, religious organizations, armed services, youth organizations, women's organizations, etc.—in which, under presidential guidance, a national consensus could express itself.

The next two years were a period of almost continuous crisis. The resignation of the second Ali government was followed by a proclamation of a "state of war and seige" and the formation of a nonpartisan government under Djuanda. At the end of 1957, in a series of direct actions across the country, Dutch property was seized as part of a campaign for the recovery of Irian Barat, and the government in due course took over responsibility for the running of these enterprises. The army itself was drawn into the management of estates, and military entrepreneurs came, in time, to play a continuing economic role.

Rebellion in western Sumatra

Early in the following year leaders from western Sumatra launched a direct challenge to Jakarta in the form of an alternative government of the republic, the Revolutionary Government of the Republic of Indonesia. The rebellion, supported by some senior Masyumi leaders, was backed also by the military commander of Sulawesi Utara (North Celebes) province.

The central government acted swiftly and successfully to suppress the rebellion. In this changed situation, with the regions defeated, the parties discredited, and the army's prestige enhanced by its success against the rebels, Sukarno once more took up the idea of Guided Democracy. With the support of the army chief of staff, General A.H. Nasution, he proposed a return to the 1945 constitution, a presidential type of government within which he believed it would be possible to implement the principles of deliberation and consensus. When the Constituent Assembly (elected in 1955 to draft a permanent constitution) failed to agree to this proposal, Sukarno introduced it by presidential decree on July 5, 1959.

Sukarno's policies. Under the 1945 constitution, Sukarno possessed executive responsibility as well as ceremonial functions as head of state. He quickly created a new government with Djuanda Kartawidjaja, now first minister, at its head. Pending elections under a new electoral law, he appointed members in accordance with the functional representation principle to the bodies for which the constitution provided: the People's Consultative Assembly (Majelis Permusyawaratan Rakyat; MPR) and the Supreme Advisory Council (Dewan Pertimbangan Agung; DPA). In 1960, when parliament rejected the government's budget, he replaced it with a provisional nominated parliament.

Sukarno's central purpose was the preservation of national unity and the restoration of a sense of national identity, goals he pursued through an increasingly flamboyant style. Sukarno's concern with symbols of greatness—expressed in grandiose buildings, national monuments, and evocative slogans and in such occasions as the Fourth Asian Games (1962), to which Indonesia was host—was not accompanied by an attempt to come to grips with the nation's economic problems. The damage done to the economy by the seizure of Dutch enterprises in 1957 and the wasteful extravagances of his later search for grandeur was justified in his eyes as integral to the task of making Indonesians proud of themselves and of their independence. Nevertheless, he was careless of the economic consequences of his policies. He appeared to show no recognition of foreign indebtedness, declining exports, or the inflation that reached new rates of acceleration in the early 1960s.

Sukarno's protection of the communists

Sukarno's power during the years of Guided Democracy depended in great measure on the preservation of a balance between the army and the PKI. Sukarno consistently protected the PKI from moves made against it by the army, and the period was one of growth in the communists' prestige. He opposed military attempts to prohibit its congresses and to suppress its newspapers. He banned movements opposing the party and advanced PKI leaders to positions within the national leadership. To many observers he appeared to be preparing the way for the communists to come to power. To others he appeared merely to be redressing a balance that was in constant danger of being tilted against the PKI.

In foreign policy, Indonesia adopted a neutralist stance. At the Asian-African Conference in 1955 (the Bandung Conference) the country staked a claim to Third World leadership. By the early 1960s, however, Indonesia was moving to a new international position. In ideological terms, Sukarno had sketched the world, as he saw it, in terms of a conflict between Nefos and Oldefos (New Emerging and Old Established Forces). In this analysis was embodied his continuing hostility to the West.

In 1962 Indonesia's campaign to recover Irian Barat, which the Dutch had retained in 1949, achieved final success. An agreement was reached with The Netherlands

for the transfer of the territory to Indonesia after a period of temporary UN administration, though with provision for the inhabitants of the territory to make an "Act of Free Choice" before the end of 1969. (This was eventually effected by representative councils, which confirmed Irian Barat's continuance as part of Indonesia.) The resolution of this issue was followed, however, by the development of Indonesia's opposition to the formation of Malaysia and its commitment, after an erratic series of changes of mood, to a policy of "confrontation" toward the new Malaysian federation in September 1963. The confrontation policy was followed two years later by Indonesia's sudden withdrawal from the UN in January 1965 in reaction to the seating of Malaysia on the Security Council.

INDONESIA SINCE 1965

The coup and after. On the night of Sept. 30, 1965, a group of army conspirators kidnapped and murdered six army generals. A seventh, Nasution, escaped. The following morning the 30th September Movement announced that it had seized power to forestall a coup against the president by a council of generals. In the meantime General Suharto, commander of the army's strategic reserve, began to gather the reins of power into his own hands. By evening he had seized the initiative from the conspirators.

The PKI maintained that the coup attempt was an internal affair of the army. The army leadership insisted that it was part of a PKI plot to seize power, and in the following months communists across Java and in Bali were slaughtered, with estimates of those killed ranging from 80,000 to more than 1,000,000. With the destruction of the PKI, one of the elements of balance that had supported the Sukarno regime was eliminated, and the president himself, though he had much strong support, came under increasing pressure. In March 1966, against a background of student action, the army forced him to delegate extensive powers to Suharto, now chief of staff of the army. With these powers, Suharto banned the PKI and moved gradually to consolidate his position as the effective head of government. In March 1967 the MPR appointed Suharto acting president, and in March 1968 he was appointed to the presidency in his own right. Sukarno was kept under house arrest until his death on June 21, 1970.

Downfall of Sukarno

Suharto's New Order. Suharto was concerned to reverse many of Sukarno's policies. The confrontation with Malaysia was quickly ended, and Indonesia rejoined the United Nations. In addition, Indonesia was a major participant in the creation of the Association of Southeast Asian Nations (ASEAN) in 1967. Domestically, the support of the army enabled Suharto to achieve a political stability that had been lacking under Sukarno. But the major policies initiated by Suharto's New Order had to do with economic rehabilitation. Successful negotiations secured a rescheduling of Indonesia's foreign debts and attracted aid through an intergovernmental group of donor countries. The complex regulations governing economic activity were simplified. In 1967 a new foreign investment law provided a framework for new private capital investment.

Economic development. The results of the new economic policies were soon apparent. The inflation rate was quickly reduced and the national currency, the rupiah, was stabilized. Manufacturing expanded rapidly. Oil production increased, thanks partly to exploration by a number of new foreign companies operating through Pertamina, the gigantic state oil corporation. (Pertamina's position as the centerpiece of Indonesia's economic expansion ended in 1975, however, when because of indebtedness it had to be rescued by the government.) Military entrepreneurs played a significant part in these developments. In the mid-1980s the decline in oil prices led to a shift in economic direction toward encouraging private-sector investment and emphasizing manufactured exports as a replacement for the traditional reliance on oil and commodity exports.

These new policies had their critics, both inside and outside the country. To some it seemed that the republic was becoming economically dependent on Western capital and, in particular, on large transnational corporations, that direct foreign investment had created an Indonesian comprador class that battened on foreign companies, and that new wealth had exaggerated existing inequalities rather than removing them. Alternatively, it was argued that long-term improvement depended on the economic growth that would flow from policies designed to encourage large-scale investment rather than small-scale, labour-intensive developments.

Achievements of the New Order

The economic achievements of New Order policies were spectacular, and during the 1970s and '80s they transformed the developmental patterns of the archipelago. New investment was especially obvious outside Java. Historically the political centre and the economic hub of the Indies, Java seemed to retain that position within the modern republic, commanding about three-quarters of all new investment projects (excluding oil exploration) from the late 1960s to the early 1980s. The expansion of manufacturing over that period was also concentrated in Java. This apparent dominance, however, was undermined by the density of Java's population. In per capita terms its share of foreign investment was outstripped by some of the outer provinces. Sumatera Utara, the home of the great plantation expansion of the late 19th century, added oil and natural gas exploration and mining to its estate agriculture. Mining and oil had an even greater impact on the development of Aceh, Riau, and Kalimantan Timur, as well as Irian Jaya. In per capita terms Kalimantan Timur, with timber and oil, natural gas, and coal, attracted high levels of both foreign and domestic investment and became one of the most rapidly developing provinces of the republic. By contrast, the Nusa Tenggara Barat, Nusa Tenggara Timur (West and East Lesser Sunda), and Timor Timur were the poorest and least developed provinces in both absolute and per capita terms. Successive five-year plans emphasized the importance of redressing regional disparities and spreading economic growth more evenly.

Changes in Indonesian society. The economic successes of the Suharto regime were accompanied by some shifts in the balance of Indonesian society. Observers hesitated to apply the term "middle class" either to the emerging elite of the late colonial period or to the small traders of coastal ports or market towns of Indonesia. With its suggestion of substantial commercial activity, the term seemed not to fit the elite, which, under the Dutch, appeared rather as an administrative or bureaucratic class linked—both in Java and elsewhere—to an earlier aristocracy. In addition, indigenous traders lacked the wealth associated with a bourgeoisie.

The picture was further complicated by the special position of the Chinese in rural and urban trade. The increased Chinese immigration in the 20th century confirmed the distinction between *peranakan* and *totok* communities (*i.e.,* between those who had been in Indonesia for a number of generations and had adopted Indonesian customs and language and more recent arrivals who retained their language and remained more conscious of being Chinese). Unevenly spread across the archipelago and an ethnic minority playing a distinct economic role, the Chinese were likely to attract Indonesian hostility. After independence—and in spite of occasional outbreaks of anti-Chinese feeling—they continued to expand their participation both in retail trade and in large-scale commerce and finance.

Social change accelerated under the New Order. Along with the decline of the position of traditional aristocracies went the growth of a new bureaucracy, the rise of the army both in politics and administration and in commercial activity, the establishment of an Indonesian business class, and the presence of Chinese business interests, the latter sometimes in association with civilian or military Indonesian entrepreneurs. These developments suggested that a new—albeit extremely diverse—middle class was emerging, defined in part by economic function, by access to political power, and by a life-style of conspicuous consumption. Whether it was one class or several and whether the term should cover, simultaneously, wealthy capitalists and small rural traders, senior bureaucrats and low-level clerks, and military officers and civil professionals continued to be a matter for debate.

Rise of a middle class

These developments tended to confirm, rather than to modify, the structure of power in Suharto's Indonesia.

Political developments. Politically, the New Order con-

tinued to be a stable regime partly because of economic development across the archipelago but mainly because it was underpinned from the beginning by military power. It would be incorrect to describe it as a military regime, and Suharto, in the early years of his presidency, was concerned to observe constitutional forms. His initial government had strong civilian components in the persons of Sultan Hamengkubuwono IX of Yogyakarta and Adam Malik (both of whom later served as vice president). But military strength, allied closely with bureaucracy, was apparent nonetheless, and the government developed clear authoritarian characteristics.

Suharto acted to control and discipline, and ultimately to rationalize, the political parties. In 1973 the four Muslim parties (Parmusi, formed as a successor to Masyumi, together with the NU, PSII, and Perti) were amalgamated to form the United Development Party (Partai Persatuan Pembangunan; PPP), and the five non-Muslim parties (PNI, Parkindo, Katholik, IPKI, and Murba) were amalgamated to form the Indonesian Democratic Party (Partai Demokrasi Indonesia; PDI). More formidable than either was the government-sponsored organization, the Joint Secretariat of Functional Groups (Sekretariat Bersama Golongan Karya; Sekber Golkar, or Golkar). In theory, Golkar was a nonpartisan organization representing, like Sukarno's functional groups, the elements of which the nation was composed; in practice, it was a government party, and its sweeping electoral successes owned much to pressure brought to bear on voters by government agencies. In 1971 it secured 236 seats out of 360 in the DPR, and its dominance was confirmed in subsequent elections in 1977, 1982, and 1987. Important also as a measure of political control was the government's imposition of the Pancasila, or the Five Principles (belief in one God, nationalism, humanitarianism, democracy, and social justice) as the national ideology.

Suharto's reelection to office

Between 1971 and 1988 parliamentary elections were followed by the unopposed reelection of Suharto for successive presidential terms. These results were not achieved without effort. Suharto's economic policies and, in particular, the attempt to spread development more evenly across the archipelago, contributed to reducing the strong regional feelings of the 1950s, though there remained perceptions that the regime was dominated by Java. A special case was Irian Jaya, where the government has had to contend with the resistance of the Free Papua Movement (Organisasi Papua Merdeka; OPM), even after the 1969 Act of Free Choice, with actual outbreaks in 1977, 1978, and 1981. Encouragement of immigration from Java and elsewhere and extended educational opportunities intended to make Irian Jaya more fully a part of Indonesia were seen as examples of cultural imperialism. The exploitation of the resources of the province—oil, natural gas, copper, and timber—was also a source of resentment.

Unlike Irian Jaya, which had always been claimed by Indonesia as a part of the republic, the Portuguese colony of Timor had not been the subject of any such claim until political changes in Portugal threw the future of the colony into doubt. In 1975–76 Indonesia forcibly incorporated East Timor (subsequently Timor Timur province) into the republic in a fashion that evoked domestic as well as foreign criticism and left the government facing a continuing, and particularly harsh, pacification operation against Fretilin, the resistance movement struggling for an independent East Timor. By the late 1980s Indonesia appeared to be in control but with enormous tasks of reconstruction to undertake.

In addition to these areas of specific resistance there has been some Islāmic opposition to the regime. Developments in Muslim thought tended increasingly to blur the old stereotyped distinction between modernist and traditionalist, or fundamentalist, thinking. Though dealing essentially with theological issues, the effect of these shifts in emphasis was felt as a movement of Islāmic renewal both within and outside the PPP. Focused initially on dislike of the essentially secular ideology of Pancasila, they came to represent a more general ambivalence. There were also some intellectual and student criticisms of the corruption built into the structure of the economy, which was seen as reaching into the highest levels of government. There were examples of open discontent, as when students chose the visit of the Japanese prime minister Tanaka Kakuei in 1974 to initiate demonstrations against Suharto and against the role of foreign capital in Indonesia; the demonstrations developed into open rioting in Jakarta. In 1978, before the reelection of Suharto for a third term, the government felt it necessary to close sections of the press and to arrest student leaders.

Political lines of division within Indonesian society have not been easy to define, but insofar as they have contained elements of rivalry between centre and regions, opposition between Muslims and non-Muslims, tensions between different strands of Islām, and, in Java, the division between *santri* and *abangan,* as well as between rich and poor, they have reflected divisions of long standing.

International relations. Indonesia's domestic stability has been accompanied by moderation in external policies. The country's standing as a Third World leader was enhanced in 1985 when it hosted a second Asian-African Conference to commemorate the one held in 1955. In conjunction with the government of Papua New Guinea, Indonesia sought to contain incidents on the border between Irian Jaya and Papua New Guinea. In 1989 it reached agreement with Australia on the exploitation of seabed resources between the two countries. More generally, Indonesia participated increasingly in the affairs of the region. Through ASEAN it took a firm stand against Vietnam's invasion and occupation of Kampuchea (Cambodia) and in 1989–90 played a major part in exploring the possibilities of a negotiated resolution of the Indochina problem. (J.D.L.)

Participation in ASEAN

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 937, 969, and 976, and the *Index.*

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

BIBLIOGRAPHY

Physical and human geography: General overviews are provided by CHARLES A. FISHER, *South-East Asia: A Social, Economic, and Political Geography,* 2nd ed. (1966), a classic work with a dated though very useful perspective on Indonesia that details economic, social, and political as well as physical variations; G.J. MISSEN, *Viewpoint on Indonesia: A Geographical Study* (1972), an older but well-written historical economic geography that traces agriculture, both traditional and commercial, from the Dutch period until independence and discusses the problems of the economy and urbanization; DONALD W. FRYER and JAMES C. JACKSON, *Indonesia* (1977), a broad survey of the physical environment and an analysis of the country's complex history, with an exploration of the dichotomy between indigenous agriculture and the modern economy, especially oil; HASSAN SHADILY (ed.), *Ensiklopedi Indonesia,* 7 vol. (1980–84); FREDERICA M. BUNGE (ed.), *Indonesia: A Country Study,* 4th ed. (1983); and the relevant section in *The Far East and Australasia* (annual).

R.W. VAN BEMMELEN, *The Geology of Indonesia,* 2 vol. in 3 (1949, reprinted 1970), is an exhaustive survey of the country's geology and natural resources, with detailed information on individual islands. Also useful are WARREN HAMILTON, *Tectonics of the Indonesian Region* (1979); and E.C.J. MOHR, F.A. VON BAREN, and J. VAN SCHUYLENBORGH, *Tropical Soils: A Comprehensive Study of Their Genesis,* 3rd rev. and enlarged ed. (1973), with most examples drawn from Indonesia. Detailed regional studies include ANTHONY J. WHITTEN, MUSLIMIN MUSTAFA, and GREGORY S. HENDERSON, *The Ecology of Sulawesi* (1987); and ANTHONY J. WHITTEN *et al., The Ecology of Sumatra* (1987).

W.F. WERTHEIM, *Indonesian Society in Transition: A Study of Social Change,* 2nd rev. ed. (1959, reprinted 1980); and RUTH T. MCVEY (ed.), *Indonesia* (1963), provide introductions to Indonesian society. N. ISKANDAR, *Some Monographic Studies on the Population of Indonesia* (1970), contains a review of the 1961 census and projections; see also WIDJOJO NITISASTRO, *Population Trends in Indonesia* (1970), which covers the period 1775–1961 and predicts trends through 1991. UNIVERSITAS INDONESIA, *The Population of Indonesia* (1974), is the World Population Year monograph commissioned by the UN Committee for International Coordination of National Research in Demography; and WERNER RUTZ, *Cities and Towns in Indonesia* (1987; originally published in German, 1985), is a study of urbanization based on the 1980 census. CHRISTINE DRAKE, *National Integration in Indonesia: Patterns and Policies* (1989),

examines integration and cohesiveness in a country with a varied social fabric; and VICTOR T. KING (ed.), *Essays on Borneo Societies* (1978), discusses the Dayak peoples.

HAL HILL, *Foreign Investment and Industrialization in Indonesia* (1988), is a good analysis of historical and current economic development. HAL HILL (ed.), *Unity and Diversity: Regional Economic Development in Indonesia Since 1970* (1989), focuses on problems and progress in regional development, with chapters on each province. COLIN MACANDREWS (ed.), *Central Government and Local Development in Indonesia* (1986), discusses the evolution of development policies under the New Order government, with attention to increasing trends toward decentralized decison making and to how the government works and how development policies are implemented at the national and local levels. GRAEME J. HUGO *et al., The Demographic Dimension in Indonesian Development* (1987), analyzes in detail population growth and trends, fertility and mortality, mobility and urbanization, and the growth of the labour force. WOLF DONNER, *Land Use and Environment in Indonesia* (1987), describes the geographic and demographic setting and analyzes the social and environmental "catastrophe" confronting Java; it also analyzes the environmental side effects of nonagricultural land use development. THOMAS R. LEINBACH and CHIA LIN SIEN, *South-East Asian Transport: Issues in Development* (1989), offers a developmental approach to the evolution of transport in the region; the Indonesian chapter traces investment and planning and discusses air, road, rail, and sea sectors. A useful work on Indonesian government and politics is BENEDICT ANDERSON and AUDREY KAHIN (eds.), *Interpreting Indonesian Politics* (1982).

Cultural material is presented in MIGUEL COVARRUBIAS, *Island of Bali* (1937, reissued 1986); RUTH T. MCVEY (ed.), *Indonesia* (1963), a scholarly reference work; NIELS A. DOUWES DEKKER, *Tanah Air Kita: A Book on the Country and People of Indonesia,* 5th rev. ed. (1965?), a popular pictorial review of the islands but one that contains an accurate cultural description; FRITS A. WAGNER, *Indonesia: The Art of an Island Group,* rev. ed., trans. from Dutch (1967); CLAIRE HOLT, *Art in Indonesia: Continuities and Change* (1967); and JACQUES DUMARÇAY, *Borobudur,* trans. from French (1978, reprinted 1983). (T.R.L.)

History: General treatments of Indonesian history in the context of the broader history of Southeast Asia include D.G.E. HALL, *A History of South-East Asia,* 4th ed. (1981); PAUL WHEATLEY, *Nāgara and Commandery: Origins of the Southeast Asian Urban Traditions* (1983); KENNETH R. HALL, *Maritime Trade and State Development in Early Southeast Asia* (1985); and DAVID STEINBERG (ed.), *In Search of Southeast Asia: A Modern History,* rev. ed. (1987). Histories of the country alone may be found in BERNARD H.M. VLEKKE, *Nusantara: A History of Indonesia,* rev. ed. (1959); BENJAMIN HIGGINS and JEAN HIGGINS, *Indonesia: The Crisis of the Millstones* (1963); AILSA ZAINU'DDIN, *A Short History of Indonesia,* 2nd ed. (1980); M.C. RICKLEFS, *A History of Modern Indonesia, c. 1300 to the Present* (1981); and LOUIS FISCHER, *The Story of Indonesia* (1959). J.D. LEGGE, *Indonesia,* 3rd ed. (1980), examines some historiographical problems.

F.D.K. BOSCH, *Selected Studies in Indonesian Archaeology* (1961), contains selected translations of some of Bosch's distinguished contributions to the study of Indonesian culture. G. COEDÈS, *The Indianized States of Southeast Asia* (1968, reissued 1975; originally published in French, 1944), has sections dealing with Indonesia that introduce basic information on early Indonesian history and discuss its implications in a judicious manner. N.J. KROM, *Hindoe-javaansche Geschiedenis,* 2nd ed. (1931), is the first and a very detailed critical account of information on early Indonesian history; though parts of the work are now dated, it remains the basic work on the subject. THEODOOR G. PIGEAUD (ed.), *Java in the 14th Century: A Study in Cultural History,* 3rd ed. rev., 5 vol. (1960–63), a translation of the 14th-century *Nāgarakeṛtāgama,* accompanied by an extensive commentary, is indispensable reading for the study of Java, especially in the 13th and 14th centuries. SOEDJATMOKO *et al.* (eds.), *An Introduction to Indonesian Historiography* (1965), is an important survey and discussion of the skills and contributions of a variety of scholarly disciplines in the field of Indonesian history. The attention being given to the study of early Java is reflected in JAN WISSMAN CHRISTIE, "Raja and Rama: the Classical State in early Java," in LORRAINE GESICK (ed.), *Centers, Symbols, and Hierarchies: Essays on the Classical States of Southeast Asia* (1983), pp. 9–44. On the continuing study of Śrīvijaya, consult O.W. WOLTERS, "Studying Śrīvijaya," *Journal of the Malaysian Branch of the Royal Asiatic Society,* 52(2):1–32 (December 1979); and PIERRE-YVES MANGUIN (comp.), *Bibliography for Sriwijayan Studies* (1989).

For a discussion of trade patterns of the early period of European contact, see J.C. VAN LEUR, *Indonesian Trade and Society: Essays in Asian Social and Economic History,* 2nd ed., trans. from Dutch (1960); M.A.P. MEILINK-ROELOFSZ, *Asian Trade and European Influence in the Indonesian Archipelago Between 1500 and About 1630* (1962); and ANTHONY REID, *Southeast Asia in the Age of Commerce, 1450–1680* (1988–). M.C. RICKLEFS, *Jogjakarta Under Sultan Mangkubumi, 1749–1792: A History of the Division of Java* (1974), examines 18th-century Javanese politics against the background of the Dutch presence. CHRISTINE DOBBIN, *Islamic Revivalism in a Changing Peasant Economy: Central Sumatra, 1784–1847* (1983); and HEATHER SUTHERLAND, *The Making of a Bureaucratic Elite: The Colonial Transformation of the Javanese Priyayi* (1979), examine social change in the 19th century in Sumatra and Java, respectively. JEAN GELMAN TAYLOR, *The Social World of Batavia: European and Eurasian in Dutch Asia* (1983), explores European adaptation to the local scene in the 18th and 19th centuries. CLIVE DAY, *The Policy and Administration of the Dutch in Java* (1904, reprinted as *The Dutch in Java,* 1972), remains an interesting treatment of the Cultivation System and Liberal Policy. The best survey in English of Dutch economic policies in the 19th and 20th centuries is still J.S. FURNIVALL, *Netherlands India* (1939, reissued 1983). See also G.C. ALLEN and AUDREY G. DONNITHORNE, *Western Enterprise in Indonesia and Malaya: A Study in Economic Development* (1954, reprinted 1968); and the collection of Dutch economic writings published as *Indonesian Economics: The Concept of Dualism in Theory and Policy,* 2nd ed. (1966). ROBERT VAN NIEL, *The Emergence of the Modern Indonesian Elite* (1960, reissued 1984), studies the theory and operation of the Ethical Policy. DELIAR NOER, *The Modernist Muslim Movement in Indonesia, 1900–1942* (1973), surveys Islāmic thought in the late colonial period.

GEORGE MCTURNAN KAHIN, *Nationalism and Revolution in Indonesia* (1952, reissued 1970), remains the standard study of the rise of nationalism and the struggle for independence. For a more recent survey of the revolution, see ANTHONY REID, *The Indonesian National Revolution, 1945–1950* (1974, reprinted 1986). RUTH T. MCVEY, *The Rise of Indonesian Communism* (1965), is an authoritative history of the Indonesian Communist Party to the revolts of 1926–27. BERNHARD DAHM, *Sukarno and the Struggle for Indonesian Independence* (1969; originally published in German, 1966), explores the development of Sukarno's thinking up to 1945. The Japanese occupation is examined in HARRY J. BENDA, *The Crescent and the Rising Sun: Indonesian Islam Under the Japanese Occupation, 1942–1945* (1958, reissued 1983); and BENEDICT R. O'G. ANDERSON, *Some Aspects of Indonesian Politics Under the Japanese Occupation: 1944–1945* (1961). Anderson's *Java in a Time of Revolution: Occupation and Resistance, 1944–1946* (1972), gives a close study of the opening period of revolution. The standard account of the early years of independence is HERBERT FEITH, *The Decline of Constitutional Democracy in Indonesia* (1962). DANIEL S. LEV, *The Transition to Guided Democracy* (1966), carries the story from 1957 to 1959; and REX MORTIMER, *Indonesian Communism Under Sukarno* (1974), continues from 1959 to 1965. J.D. LEGGE, *Sukarno: A Political Biography,* 2nd ed. (1984), also covers the period. J.A.C. MACKIE, *Konfrontasi: The Indonesia-Malaysia Dispute, 1963–1966* (1974), examines closely a significant episode of Indonesian foreign policy. The essays in CLAIRE HOLT (ed.), *Culture and Politics in Indonesia* (1972), provide illuminating treatment of aspects of modern Indonesian history and culture. The role of the army is examined in HAROLD CROUCH, *The Army and Politics in Indonesia,* rev. ed. (1988). The New Order is discussed by KARL D. JACKSON and LUCIAN W. PYE (eds.), *Political Power and Communications in Indonesia* (1978). RICHARD ROBISON, *Indonesia: The Rise of Capital* (1986), gives a radical interpretation of New Order economic policy. DAVID REEVE, *Golkar of Indonesia: An Alternative to the Party System* (1985), examines the background to the political ideas of the New Order. CHARLES A. COPPEL, *Indonesian Chinese in Crisis* (1983), is a study of the modern Chinese community. The journals *Indonesia* (semiannual) and *The South East Asian Review* (semiannual) contain useful articles on current scholarship. (O.W.W./J.D.L.)

Industrial Engineering and Production Management

The aim, and to a remarkable degree the achievement, of modern industry is to produce large quantities of well-designed, effective goods at prices that put them within reach of the largest possible numbers of consumers. The capability of doing so originated with the Industrial Revolution in 18th-century England. This period was characterized by several important inventions and by the rapid introduction of machines in the textile industry, which produced significant increases in trade and profits for England. By the early 19th century what came to be known as the "factory system" had started to spread, first to the United States, where it was quickly established, then more slowly to the rest of western Europe, and eventually around the world.

The development of large-scale industrial production drew on many of society's resources: science, engineering, finance, law, and politics. This article is concerned specifically with the application of engineering and scientific methods, scientific research, and management concepts and methods to the problems of industrial production.

Industrial production refers to the methods used in factories for the efficient manufacture of goods. Two important aspects of industrial production are product technology, which is concerned with the design and characteristics of the product itself, and process technology, which includes the design, organization, and operation of the machines and the planning and control systems used in producing manufactured goods. Industrial engineering is the field that has grown up around the study of methods used in both areas. Production management refers to the area whose specialty is the management of the human, capital, and technical resources used in industrial production.

Research and studies conducted in the early part of the 20th century resulted in new knowledge about industrial production and led to the introduction of scientific methods for the management of work. The field of industrial engineering broadened considerably when the powerful concepts and techniques developed by early pioneers were discovered to be applicable not only to manufacturing and production operations but also to many other operating systems. In subsequent decades the computer has drastically altered the field and has made possible substantial new developments in planning and control systems and in the processing of the mass of information. Increasingly powerful computers and software have also made it possible to build large, sophisticated mathematical models and simulations to aid in the understanding of complex operating systems, including the interactions of people, equipment, and information within them.

Paralleling the development of improved methods for achieving efficient production were increasing concerns for the health, safety, and comfort of workers and for the effects of industrial production on the environment. These concerns spurred the development of techniques for analyzing total systems or important parts of operating systems and led to significant developments in such areas as pollution control and the design of equipment to minimize the risks to workers' health.

This article deals with several facets of industrial engineering and production management. It begins with a review of industrial production or, more generally, production systems—the systems that are the focus of most applications of industrial engineering and production management. Included are the general principles and requirements that guide the development of such systems and a brief history of mass production, the phenomenon that might be described as the finest achievement of industrial production.

The discussion of production management emphasizes aspects that are important in almost any setting—managing human resources, selecting equipment and technology, managing systems flows, financing operations, and controlling the use of capital. Fundamental issues of planning and control are also discussed.

Mathematical and computer-based methods and techniques used in the fields of industrial engineering and operations research are also covered. These methods consist of the development of rigorous mathematical models and procedures to test the soundness of alternative decisions, to find ways of improving processes, and to aid in the understanding of the complex interactions that characterize most operating systems.

This article also discusses systems engineering, which focuses on the application of a wide range of concepts and techniques—from several different disciplines—to solve complex engineering problems. The key to successful systems engineering is the proper integration of techniques and the development of a full description and understanding of the problem at hand.

The section of this article on ergonomics and human-factors engineering discusses the disciplines that deal with the design of products, processes, and equipment which can be used comfortably, safely, and effectively by human beings.

Much of the scientific research conducted in developed countries is done by industrial organizations and companies. Indeed, in some fields, such as computer technology and electronics, it is imperative that manufacturers maintain active programs of scientific research aimed at improving current products and production methods and discovering fruitful areas for new developments. With competition in many industries taking place on a global scale, research and development has become an important competitive weapon; and many independent research organizations, as well as those funded by governments or by universities, also contribute to the development of new technologies that can yield commercial benefits. The section of this article on research and development covers the history of industrial research and focuses on the management of different types of research and development operations. (W.K.H.)

The article is divided into the following sections:

Production systems 203
  Basic considerations 203
    Underlying principles
    Types of production systems
    Important considerations
  Mass production 204
    The Industrial Revolution and early developments
    Pioneers of mass production methods
    Manufacturing pioneers
    A summary of mass production concepts
    The mass production of automobiles
    Nonmanufacturing examples of mass production
    Mass production and society
    Economic effects
Production management 207
  The "five M's" 207
  Planning and control 208
  The importance of models and methods 208
Industrial engineering and operations research 208
  Engineering and science as a support to management 208
  The evolving nature of industrial engineering 209

Scientific method and operations research 209
Basic aspects
History
Essential characteristics 210
Systems orientation
The interdisciplinary team
Methodology
Phases of operations research 210
Problem formulation
Model construction
Deriving solutions from models
Testing the model and the solution
Implementing and controlling the solution
Computers and operations research 212
Simulation
Decision analysis and support
New software tools for decision making
Examples of operations research models and applications 213
Resource allocation
Linear programming
Inventory control
Japanese approaches
Replacement and maintenance
Queuing
Job shop sequencing
Manufacturing progress function
Network routing
Competitive problems
Search problems
Frontiers of operations research 217
Strategic problems
The system design problem
The planning problem
Types of organization
Systems engineering 217
The development of systems engineering 218
Mathematical modeling
Operations research and systems engineering
Communications and electronics
Computers and systems engineering
Systems engineering techniques, tools, and procedures 219
Modeling and optimization
Identifying objectives
A design example
User orientation
Tools
Applications of systems engineering 221
The development of radically new systems
Long-term systems development
Applications to government and social problems
Ergonomics and human-factors engineering 222
The human-factors approach to design 222
The man-machine model
A man-machine example
Human factors in large systems
Applications of human-factors engineering 223
Push-button telephone
Space suit
Typewriter keyboard
Social problems
Research and development 224
Introduction and definitions 224
History and importance 224
Types of laboratories 225
Company laboratories
Government laboratories
Independent laboratories
Research associations
University laboratories
The role of government 226
The management of research and development activities 227
Project management and planning techniques 227
Value engineering and cost-benefit analysis
PERT and CPM
Bibliography 228

Production systems

BASIC CONSIDERATIONS

Underlying principles. All production systems, when viewed at the most abstract level, might be said to be "transformation processes"—processes that transform resources into useful goods and services. The transformation process typically uses common resources such as labour, capital (for machinery and equipment, materials, etc.), and space (land, buildings, etc.) to effect a change. Economists call these resources the "factors of production" and usually refer to them as labour, capital, and land. Production managers refer to them as the "five M's": men, machines, methods, materials, and money.

When viewed as a process, a production system may be further characterized by flows (channels of movement) in the process: both the physical flow of materials, work in the intermediate stages of manufacture (work in process),

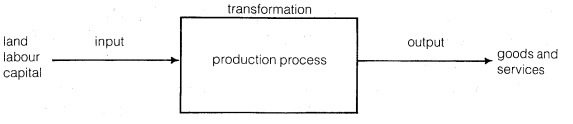


Figure 1: A production system generalized.

and finished goods; and the flow of information and the inevitable paperwork that carry and accompany the physical flow. The physical flows are subject to the constraints of the capacity of the production system, which also limits the system's ability to meet output expectations. Similarly, the capacity of the information-handling channel of the production system may also be an important measure of a system's output. The management of information flows, or the planning and control of the system to achieve acceptable outputs, is an important task of the production manager.

While the capacity of the system is the major factor in determining whether output expectations can be met, the additional consideration of quality must also be seen as a limiting factor. The quality of a product, measured against some objective standard, includes appearance, performance characteristics, durability, serviceability, and other physical characteristics; timeliness of delivery; cost; appropriateness of documentation and supporting materials; and so on. It is an important part of the definition of a system.

Types of production systems. There are three common types of basic production systems: the batch system, the continuous system, and the project system. In the batch system, general-purpose equipment and methods are used to produce small quantities of output (goods or services) with specifications that vary greatly from one batch to the next. A given quantity of a product is moved as a batch through one or more steps, and the total volume emerges simultaneously at the end of the production cycle. Examples include systems for producing specialized machine tools or heavy-duty construction equipment, specialty chemicals, and processed food products, or, in the service sector, the system for processing claims in a large insurance company. Batch production systems are often referred to as job shops.

Basic production systems

In the continuous system, items to be processed flow through a series of steps, or operations, that are common to most other products being processed. Since large volumes of throughput are expected, specially designed equipment and methods are often used so that lower production costs can be achieved. Frequently the tasks handled by workers are divided into relatively small segments that can be quickly mastered and efficiently performed. Examples include systems for assembling automobile engines and automobiles themselves, as well as other consumer products such as televisions, washing machines, and personal computers. Continuous production systems are often referred to as assembly systems or assembly line systems and, as noted below, are common in mass production operations.

The two types of systems mentioned thus far are often found in combination. In the production of integrated circuits for electronic equipment, for example, thousands of circuits are processed as a batch on several large slices of silicon crystal through dozens, or even hundreds, of processing steps. The tiny circuits, each only a few mil-

limetres on a side, are then separated and individually assembled with other circuit elements on a continuous line to produce the final product.

The third type of production system is the project, or "one-shot" system. For a single, one-of-a-kind product, for example, a building, a ship, or the prototype of a product such as an airplane or a large computer, resources are brought together only once. Because of the singular nature of project systems, special methods of management have been developed to contain the costs of production within reasonable levels.

Important considerations. Once the general specifications of a production system have been agreed upon, including precise definitions of needed resources and output expectations, three important decisions remain. First, industrial engineers, production managers, and other specialists must choose and design the technology to be used. Their decisions must include the choice of equipment and tooling, the layout of plant space and facilities, the selection of workers and work procedures, and many other aspects of process design. These choices must be handled carefully; mistakes at this early stage can result in a business losing its competitiveness or the ability to sustain a profitable position in the market.

Next, given a choice of technology, the capacity of the system must be determined. The capacity of the system is designed to be a function of the amount of available capital, the demand forecast for the output of the facility, and many other minor factors. Again, these decisions must be made wisely. Establishing too much capacity, too soon, can burden a company with excess costs and inefficient operations. Too little capacity can make it difficult and expensive to increase output later if the market develops rapidly; this can place a company at a significant cost disadvantage if other competitors, with larger facilities, produce a product at a lower cost or with more consistent quality.

Adjusting production capacity

Finally, given a basic commitment to capacity, decisions must be made on the adaptability of the production volume to meet the inevitable changes in market demand that the firm will experience. Capacity in most production systems is adjusted by hiring or firing workers, by scheduling overtime or cutting back on work hours, by adding or shutting down machines or whole departments or areas of the facility, or by changing the rate of production within reasonable limits. The effectiveness of any one of these adjustment mechanisms depends largely on the technological constraints of the process itself, the economics of the industry, and the nature of the competition. In some industries, adjustment of capacity is a very difficult task. Assembly lines with specialized equipment, for example, are most efficient when run at one speed and cannot be slowed down or run intermittently without severe economic losses. In such cases, careful attention to the fundamental design of the production system is a critical factor in the overall success of the business.

MASS PRODUCTION

The characteristics of production systems are most apparent in mass production operations. The term mass production refers to manufacturing processes that attain high rates of output at low unit cost, with lower costs expected as volume rises. Mass production methods are based on two general principles: (1) the division and specialization of human labour; and (2) the use of tools, machinery, and other equipment, usually automated, in the production of standard, interchangeable parts and products.

The Industrial Revolution and early developments. The principle of the division of labour and the resulting specialization of skills can be found in many human activities, and there are records of its application to manufacturing in ancient Greece. The first unmistakable examples of manufacturing operations carefully designed to reduce production costs by specialized labour and the use of machines appeared in the 18th century in England. They were signaled by five important inventions in the textile industry: (1) John Kay's flying shuttle in 1733, which permitted the weaving of larger widths of cloth and significantly increased weaving speed; (2) Edmund Cartwright's power loom in 1785, which increased weaving speed still further; (3) James Hargreaves' spinning jenny in 1764; (4) Richard Arkwright's water frame in 1769; and (5) Samuel Crompton's spinning mule in 1779. The last three inventions improved the speed and quality of thread-spinning operations. A sixth invention, the steam engine, perfected by James Watt, was the key to further rapid development. After making major improvements in steam engine design in 1765, Watt continued his development and refinement of the engine until, in 1785, he successfully used one in a cotton mill. Once human, animal, and water power could be replaced with a reliable, low-cost source of motive energy, the Industrial Revolution was clearly established, and the next 200 years would witness invention and innovation the likes of which could never have been imagined.

In 1776 Adam Smith, in his *Wealth of Nations,* observed the benefits of the specialization of labour in the manufacture of pins. Although earlier observers had noted this phenomenon, Smith's writings commanded widespread attention and helped foster an awareness of industrial production and broaden its appeal.

The next major advance was made in 1797 when Eli Whitney, inventor of the cotton gin, proposed the manufacture of flintlocks with completely interchangeable parts, in contrast to the older method under which each gun was the individual product of a highly skilled gunsmith and each part was hand-fitted.

Brunel's production line

During the same period similar ideas were being tried out in Europe. In England Marc Brunel, a French-born inventor and engineer, established a production line to manufacture blocks (pulleys) for sailing ships, using the principles of division of labour and standardized parts. Brunel's machine tools were designed and built by Henry Maudslay, who has been called the father of the machine tool industry. Maudslay recognized the importance of precision tools that could produce identical parts; he and his student, Joseph Whitworth, also manufactured interchangeable, standardized metal bolts and nuts.

By the middle of the 19th century the general concepts of division of labour, machine-assisted manufacture, and assembly of standardized parts were well established. Large factories were in operation on both sides of the Atlantic, and some industries, such as textiles and steel, were using processes, machinery, and equipment that would be recognizable even in the late 20th century. The growth of manufacturing was accelerated by the rapid expansion of rail, barge, ship, and road transportation. The new transport companies not only enabled factories to obtain raw materials and to ship finished products over increasingly large distances, but they also created a substantial demand for the output of the new industries.

At this point in the Industrial Revolution, the methods and procedures used to organize human labour, to plan and control the flow of work, and to handle the myriad details on the shop floor were largely informal and were based on historical patterns and precedents. One man changed all of that.

Pioneers of mass production methods. In 1881, at the Midvale Steel Company in the United States, Frederick W. Taylor began studies of the organization of manufacturing operations that subsequently formed the foundation of modern production planning. After carefully studying the smallest parts of simple tasks, such as the shoveling of dry materials, Taylor was able to design methods and tools that permitted workers to produce significantly more with less physical effort. Later, by making detailed stopwatch measurements of the time required to perform each step of manufacture, Taylor brought a quantitative approach to the organization of production functions.

At the same time, Frank B. Gilbreth and his wife, Lillian M. Gilbreth, U.S. industrial engineers, began their pioneering studies of the movements by which people carry out tasks. Using the then new technology of motion pictures, the Gilbreths analyzed the design of motion patterns and work areas with a view to achieving maximum economy of effort. The "time-and-motion" studies of Taylor and the Gilbreths provided important tools for the design of contemporary manufacturing systems.

In 1916 Henri Fayol, who for many years had managed

a large coal mining company in France, began publishing his ideas about the organization and supervision of work, and by 1925 he had enunciated several principles and functions of management. His idea of unity of command, which stated that an employee should receive orders from only one supervisor, helped to clarify the organizational structure of many manufacturing operations.

Manufacturing pioneers. Much of the credit for bringing these early concepts together in a coherent form, and creating the modern, integrated, mass production operation, belongs to the U.S. industrialist Henry Ford and his colleagues at the Ford Motor Company, where in 1913 a moving-belt conveyor was used in the assembly of flywheel magnetos. With it assembly time was cut from 18 minutes per magneto to five minutes. The approach was then applied to automobile body and motor assembly. The design of these production lines was highly analytical and sought the optimum division of tasks among work stations, optimum line speed, optimum work height, and careful synchronization of simultaneous operations.

Introduction of the moving-belt conveyor

The success of Ford's operation led to the adoption of mass production principles by industry in the United States and Europe. The methods made major contributions to the large growth in manufacturing productivity that has characterized the 20th century and produced phenomenal increases in material wealth and improvements in living standards in the industrialized countries. (For a fuller survey of the development of labour and production line manufacture, see WORK AND EMPLOYMENT.)

A summary of mass production concepts. The efficiencies of mass production result from the careful, systematic application of the ideas and concepts outlined above. The following summary lists the basic principles of mass production:

1. The careful division of the total production operation into specialized tasks comprising relatively simple, highly repetitive motion patterns and minimal handling or positioning of the workpiece. This permits the development of human motion patterns that are easily learned and rapidly performed with a minimum of unnecessary motion or mental readjustment.
2. The simplification and standardization of component parts to permit large production runs of parts that are readily fitted to other parts without adjustment. The imposition of other standards (*e.g.*, dimensional tolerances, parts location, material types, stock thickness, common fasteners, packaging material) on all parts of the product further increases the economies that can be achieved.
3. The development and use of specialized machines, materials, and processes. The selection of materials and development of tools and machines for each operation minimizes the amount of human effort required, maximizes the output per unit of capital investment, reduces the number of off-standard units produced, and reduces raw material costs.
4. The systematic engineering and planning of the total production process permit the best balance between human effort and machinery, the most effective division of labour and specialization of skills, and the total integration of the production system to optimize productivity and minimize costs.

Careful, skilled industrial engineering and management are required to achieve the maximum benefits that application of these principles can provide. Planning begins with the original design of the product; raw materials and component parts must be adaptable to production and handling by mass techniques. The entire production process is planned in detail, including the flows of materials and information throughout the process. Production volume must be carefully estimated because the selection of techniques depends upon the volume to be produced and anticipated short-term changes in demand. Volume must be large enough, first, to permit the task to be divided into its sub-elements and assigned to different individuals; second, to justify the substantial capital investment often required for specialized machines and processes; and third, to permit large production runs so that human effort and capital are efficiently employed.

Production planning

The need for detailed advance planning extends beyond the production system itself. The large, continuous flow of product from the factory requires equally well-planned distribution and marketing operations to bring the product to the consumer. Advertising, market research, transportation problems, licensing, and tariffs must all be considered in establishing a mass production operation. Thus, mass production planning implies a complete system plan from raw material to consumer.

In addition to lowering cost, the application of the principles of mass production has led to major improvements in uniformity and quality. The large volume, standardized design, and standardized materials and processes facilitate statistical control and inspection techniques to monitor production and control quality. This leads to assurance that quality levels are achieved without incurring the large costs that would be necessary for detailed inspection of all products.

A major problem of mass production based on continuous or assembly line processes is that the resulting system is inherently inflexible. Since maximum efficiency is desired, tools, machines, and work positions are often quite precisely adapted to details of the parts produced but not necessarily to the workers involved in the process. Changes in product design may render expensive tooling and machinery obsolete and make it difficult to reorganize the tasks of workers. One answer has been to design machinery with built-in flexibility; for relatively little extra cost, tooling can be changed to adapt the machine to accommodate design changes.

Lack of flexibility

Similarly, a production line is usually designed to operate most efficiently at a specified rate (Figure 2). If the required production levels fall below that rate, operators and machines are being inefficiently used; and if the rate goes too high, operators must work overtime, machine maintenance cannot keep up, breakdowns occur, and the costs of production rise. Thus, it is extremely important to anticipate production demands accurately. Planning, an important function of management and engineering design, can alleviate the problems of increased demand by incorporating excess capacity in the facilities that would require the longest time to procure and install. Then, if production loads increase, it is easier to bring the entire system up to the new level. Similarly, if large fluctuations in demand cannot be avoided, flexibility to accommodate these changes economically must be planned into the system.

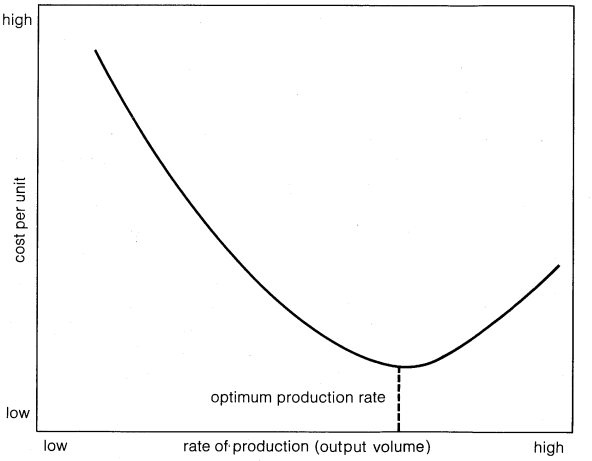


Figure 2: Curve showing relationship of unit cost to production rate.

The functional relationship pictured in Figure 2 applies broadly to the economies of scale associated with mass production as a whole. At any given stage of technological development, the economies obtained by increasing production volume are largest in the initial stages of growth and level off as volumes are further increased. Indeed, if volumes grow too large, unavoidable breakdowns of facilities, failures of coordination, or other strategic factors may cause costs to rise. Advances in technology or changes in other factors can shift the optimum point to higher levels.

Economies of scale

For these reasons planners may limit the maximum size of a single production facility and construct an independent facility if greater production is necessary.

The mass production of automobiles. The traditional example of mass production is the automobile industry, which has continued to refine the basic principles originally laid down by Henry Ford and other pioneers of mass production techniques. Today's automobile is the result of a large number of mass production lines established in a multitude of manufacturing and assembly facilities throughout the world. The assembly plant from which the finished automobile emerges is only the final element of a mass production operation that, for many companies, includes plants in several different countries. Into the final assembly plant flow large subassemblies such as the automobile chassis, the engine, major body components such as doors, panels, upholstered seats, and many electronic, electrical, and hydraulic systems such as brakes, lighting systems, and sound systems. Each of these, in turn, is usually the product of a mass production line in another factory. Stamping plants specialize in producing the formed metal parts that constitute the body of the automobile. Radio assembly plants, in turn, depend upon other assembly plants for components such as transistors and integrated circuits. There are glass plants for windows, transmission plants, tire plants, and many others, each specializing in the mass production of its own product, which is, in turn, fed into the final assembly plant. The control of the flow of material into and out of final assembly plants, including the scheduling of production from feeder plants and the timing of rail and truck shipments, is among the major engineering tasks that make the total mass production system for automobiles work.

In the final assembly line one can see clearly how machinery and human effort in assembly are divided into many specialized skills. The special tooling and machinery developed to handle assembly parts and to aid operators in their tasks can also be observed. At a given point on the line a robot welder—unaided by a human operator—may weld body parts together. At another position the motor is mounted on the chassis by a large machine guided by an operator. In other places body panels and doors are assembled to the chassis, and dashboard instruments and wiring are added by hand with simple tools. Each operator learns his task in detail and uses tools specialized for that task. The total operation is paced by the speed of movement of the conveyor that carries the partially assembled automobiles. The number of operators, machine stations, and flow of materials to the conveyor have all been planned so that the conveyor can maintain an essentially constant speed with each operator and machine functioning near optimum effectiveness.

In Ford's early lines, parts and product were precisely standardized. Only one car model was manufactured, and each unit was identical to every other unit in all aspects, including colour—black. Today's automotive manufacturing engineers have learned to mass produce a highly customized product. The same assembly line may turn out a variety of models with many colours and options. This is achieved by continued insistence on standardization of critical elements such as the methods by which parts are held together internally. Thus, the operator who specializes in assembling doors can handle a variety of models and colours equally well. In addition, the flow of materials to the various line positions is carefully scheduled and controlled so that the specific part required for a given model, colour, or option list arrives at the line at the precise moment that the partially assembled unit requiring the part has arrived along the conveyor. The exquisitely designed production-control systems operating in the automotive and other industries make it possible for the consumer to obtain a greatly enhanced variety of product without sacrificing the cost advantages of mass production techniques.

Product variety from a single production line

Nonmanufacturing examples of mass production. The mass production principles of the division and specialization of labour and the use of standardized parts and processes have been applied to a wide area of productive activity. In agriculture the development of specialized machines for plowing, seeding, cultivating, and harvesting followed by factories for preparing, preserving, and packaging food products has drawn heavily on mass production principles. There are specialized manual tasks supplementing the specialized machines both in the fields and in the processing plants.

In the service industries, such as air transportation, the division and specialization of skills can be observed among ticket agents, pilots, navigators, baggage handlers, flight attendants, maintenance crews, and traffic controllers. All major engineering projects in both design and manufacture generally require a complement of engineering specialties including chemical, mechanical, and electrical engineers and further subdivisions of these professions such as semiconductor engineers, circuit designers, and so forth.

Service industries

Thus, as industry becomes more complex at each level, the division of labour and specialization become necessary. At the same time, the need for coordination and communication between specialized members of the team becomes greater.

Mass production and society. Both the quantity and the variety of material goods in industrialized countries have resulted directly from the application of mass production principles. At the same time the environment and circumstances of those employed by, and associated with, the production of material goods have changed. The benefits that have arisen from the greatly improved productivity made possible by mass production techniques have been shared by employees, investors, and customers. The working environment has greatly changed, however. Similarly the complexities of management have increased substantially, and the investment requirements and risks faced by owners and investors have become much greater.

Before the introduction of mass production techniques, goods were produced by highly skilled craftsmen who often prepared their basic raw materials, carried the product through each of the stages of manufacture, and ended with the finished product. Typically, the craftsman spent several years at apprenticeship learning each aspect of his trade; often he designed and made his own tools. He was identified with his product and his craft, enjoyed a close association with his customers, and had a clear understanding of his contribution and his position in society.

In contrast, the division of labour, the specialization of narrow skills, the detailed engineering specification of how each task is to be carried out, and the assemblage of large numbers of employees in great manufacturing plants have greatly diluted the identification of employees with their productive functions and with their employers. Many surveys in the United States and in the industrialized countries of Europe have shown that workers do not fully understand and appreciate their roles and positions in society. In addition, the division and specialization of labour may lead to such narrowly defined skills and highly repetitive operations, paced by the steady progression of a machine or conveyor line, that tedium and fatigue arise to reduce the sense of satisfaction inherent in productive work.

These physical and psychological factors have been the subject of numerous studies by industrial psychologists and others. Special attention has been paid to work factors which affect the psychological motivation that is a prime determinant of employee productivity. The psychological effects of the repetitive aspects of some mass production tasks have been examined in great detail. Tasks that are precisely paced by the rhythm of machine operation or conveyor-belt movement appear to be particularly fatiguing. For this reason, efforts are made to structure each job so that the operator can vary his pace by working ahead of the conveyor for a period and then slowing down, and by interspersing work breaks with productive periods. Some individuals prefer tasks that are sufficiently repetitive and narrowly skilled that they do not require any substantial amount of mental concentration once the function is mastered. Most fatiguing are those repetitive tasks whose pace is out of the operator's control but which also require moderate mental concentration.

Employee fatigue

With this understanding in mind, work tasks can be structured to produce a minimum of mental and physical

fatigue; this planning is an important part of the design of a successful production operation.

The highly repetitive, tightly paced production operations are usually the most easily automated. Thus, as technology advances, that part of the production operation that is most fatiguing, is least satisfying, and takes minimum advantage of the mental and physical flexibility of human effort is replaced by automatic machinery. Not only is productivity improved, but the remaining functions that require human effort can provide a more satisfying experience.

There is also increasing study of the interaction of workers with the tools and machines that assist their efforts. Working together, engineers and physicians are making quantitative biomechanical studies of how the human body functions in performing physical tasks. These studies, which have led to the improved design of tools and work positions, are part of the broader field of human-factors engineering, which considers the abilities and limitations of people in productive functions and seeks out ways in which machines can be designed to provide the best allocation of function between human effort and machine assistance. These studies are especially important as automated manufacture becomes more common.

Increasing employee job satisfaction

The problem of the loss of employee identification with the job has been of special concern. Progressive industrial organizations work to strengthen this identification in many ways, such as by using job rotation and educational programs to diversify employees' experience and to acquaint them with various aspects of the manufacturing process. This can give each employee a concept of the total manufacturing task and the importance of each employee's specific function within that task. Employee suggestion systems provide further opportunity for the individual to have a direct effect on the productive process; the employee is given other opportunities to help structure the manner in which the job is performed. Thoughtful programs of this type can substantially ameliorate the feeling of anonymity that may otherwise result. Clearly employers must be willing to compromise on the division and specialization of work tasks that technical considerations alone might suggest as desirable. Job content and employee participation must be expanded so that the employee feels significant and retains motivation and identification.

In addition to this increased attention to the structure of the job itself, increases in productivity and resulting increases in wages have reduced working hours and provided employees with opportunities and resources to develop interests outside the workplace. Forward-looking employers, aware of these needs, frequently support these activities through employees' clubs and other means.

Economic effects. Mass production, with its heavy dependence upon mechanized facilities and high levels of production volume, presents great challenges for industrial leadership. The importance of advanced planning and the coordinated control of the large human and capital resources associated with mass production have been described. The day-to-day problems of monitoring the status of a major manufacturing complex are also immense. Up-to-the-minute knowledge of status is essential to effective response when difficulties such as the breakdown of a machine, the shortage of required materials or components, or the absence of important employees occur. Many aids to management have been devised for collecting data, analyzing them, and presenting alternatives for management decision. The electronic computer, with its great capability for collecting, analyzing, and comparing data, is becoming especially important as a management aid, both in the initial planning and simulation of production facilities and in computer-based production and resource-control systems. In fact, the increasing use of mainframe and personal computers as business tools is rapidly making the management of mass production operations a quantitative technology in its own right.

The need for substantial investment is another result of the application of mass production principles. Much of the increase in productivity that has been achieved by mass production is a direct result of the development and use of automatic machinery and processes to supplement human effort. This, in turn, requires the support of a sizable technical staff in advance of production and later substantial capital investment for production facilities. Increased levels of capital, which must often be committed years before production begins, and before the true market for the product can be established, greatly increase the risks that investors must assume and have markedly affected the investment climate in manufacturing industries.

Change in nature of ownership and investment

As capital needs have grown, the nature of ownership and investment in industry has changed dramatically. Economies of scale favour large, high-volume operations that require capital investment levels often well beyond the means of an individual owner. This has been the prime stimulant for corporate ownership of major manufacturing firms. Furthermore, the direct managers of the manufacturing enterprise seldom possess ownership control of the enterprise. This has created a new spectrum of relationships between the owners, the managers, and the employees of large manufacturing firms. Frequently the owners are principally concerned with the profits on their investments and leave the planning and managerial operations to professional managers.

At the same time the large capital needs of growing industries place special emphasis on the ability to acquire the necessary capital resources. Thus, the financial markets become extremely important in determining the general directions in which manufacturing industry will grow. This emphasizes the importance of profit incentives to encourage private investment, which is vital to achieve the productivity advances possible in mass production operations.

Similarly, industrial nations are strongly encouraged to retain and reinvest a significant fraction of their gross national product if national industries are to grow and to compete successfully in international markets. These problems of capital formation have been especially troublesome in introducing mass production in the developing nations.

Other consequences of a mass production economy have become apparent. The increased consumption associated with low-cost production has created problems of conservation of natural resources and the disposal or reconversion of the wastes of production and of goods whose utility is ended. There are technological solutions for the resulting problems of solid wastes and air and water pollution, but the political and economic problems of how the costs will be distributed are difficult to solve. (M.T./W.K.H.)

Production management

THE "FIVE M'S"

Operations management (known in industry as production management) is a responsibility similar in level and scope to other specialties such as marketing and to human resource and financial management. In manufacturing operations, production management includes responsibility for product and process design, planning and control issues involving capacity and quality, and organization and supervision of the workforce.

Role of the production manager

Production management's responsibilities are summarized by the "five M's": men, machines, methods, materials, and money. "Men" refers to the human element in operating systems. Since the vast majority of manufacturing personnel work in the physical production of goods, "people management" is one of the production manager's most important responsibilities.

The production manager must also choose the machines and methods of the company, first selecting the equipment and technology to be used in the manufacture of the product or service and then planning and controlling the methods and procedures for their use. The flexibility of the production process and the ability of workers to adapt to equipment and schedules are important issues in this phase of production management.

The production manager's responsibility for materials includes the management of flow processes—both physical (raw materials) and information (paperwork). The smoothness of resource movement and data flow is determined

largely by the fundamental choices made in the design of the product and in the process to be used.

The manager's concern for money is explained by the importance of financing and asset utilization to most manufacturing organizations. A manager who allows excessive inventories to build up or who achieves level production and steady operation by sacrificing good customer service and timely delivery runs the risk that overinvestment or high current costs will wipe out any temporary competitive advantage that might have been obtained.

PLANNING AND CONTROL

Although the five M's capture the essence of the major tasks of production management, control summarizes its single most important issue. The production manager must plan and control the process of production so that it moves smoothly at the required level of output while meeting cost and quality objectives. Process control has two purposes: first, to ensure that operations are performed according to plan, and second, to continuously monitor and evaluate the production plan to see if modifications can be devised to better meet cost, quality, delivery, flexibility, or other objectives. For example, when demand for a product is high enough to justify continuous production, the production level might need to be adjusted from time to time to address fluctuating demand or changes in a company's market share. This is called the "production-smoothing" problem. When more than one product is involved, complex industrial engineering or operations research procedures are required to analyze the many factors that impinge on the problem.

Inventory control

Inventory control is another important phase of production management. Inventories include raw materials, component parts, work in process, finished goods, packing and packaging materials, and general supplies. Although the effective use of financial resources is generally regarded as beyond the responsibility of production management, many manufacturing firms with large inventories (some accounting for more than 50 percent of total assets) usually hold production managers responsible for inventories. Successful inventory management, which involves the solution of the problem of which items to carry in inventory in various locations, is critical to a company's competitive success. Not carrying an item can result in delays in getting needed parts or supplies, but carrying every item at every location can tie up huge amounts of capital and result in an accumulation of obsolete, unusable stock. Managers generally rely on mathematical models and computer systems developed by industrial engineers and operations researchers to handle the problems of inventory control.

To control labour costs managers must first measure the amount and type of work required to produce a product and then specify well-designed, efficient methods for accomplishing the necessary manufacturing tasks. The concepts of work measurement and time study introduced by Taylor and the Gilbreths, as well as incentive systems to motivate and reward high levels of worker output, are important tools in this area of management. In new operations particularly, it is important to anticipate human resource requirements and to translate them into recruiting and training programs so that a nucleus of appropriately skilled operators is available as production machinery and equipment are installed. Specialized groups responsible for support activities (such as equipment maintenance, plant services and production scheduling, and control activities) also need to be hired, trained, and properly equipped. This type of careful personnel planning reduces the chance that expensive capital equipment will stand idle and that effort, time, and materials will be wasted during start-up and regular operations.

The effective use and control of materials often involves investigations of the causes of scrap and waste; this, in turn, can lead to alternative materials and handling methods to improve the production process. The effective control of machinery and equipment depends on each machine's suitability to its specific task, the degree of its utilization, the extent to which it is kept in optimum running condition, and the degree to which it can be mechanically or electronically controlled.

THE IMPORTANCE OF MODELS AND METHODS

Because of the enormous complexity of typical production operations and the almost infinite number of changes that can be made and the alternatives that can be pursued, a productive body of quantitative methods has been developed to solve production management problems. Most of these techniques have emerged from the fields of industrial engineering, operations research, and systems engineering. Specialists in these fields are increasingly using computers and information processing to solve production problems involving the masses of data associated with large numbers of workers, massive inventories, and huge quantities of work in process that characterize most of today's production operations. Indeed, many mass production operations could not run without the support of these industrial engineers and technical specialists. The important aspects of production control are summarized in the Table.

Industrial engineering and operations research

ENGINEERING AND SCIENCE AS A SUPPORT TO MANAGEMENT

The managers responsible for industrial production require an enormous amount of assistance and support because of the complexity of most production systems, and the additional burden of planning, scheduling, and coordination. Historically, this support was provided by industrial engineers whose major concern was with methods, standards, and the organization of process technology.

Industrial engineering originated with the studies of Taylor, the Gilbreths, and other pioneers of mass production methods. Their work expanded into responsibilities that now include the development of work methods to increase efficiency and eliminate worker fatigue; the redesign and standardization of manufacturing processes and methods for handling and transporting materials; the development of production planning and control procedures; and the determination and maintenance of output standards for

Production-Control Summary

	processes	inventory	inspection	costs
Observation	measuring rate of output; recording idle time or downtime	recording stock levels	inspecting materials and parts	collecting cost data
Analysis	comparing progress with the plan	analyzing demand for stocks in different uses and at different times	estimating process capabilities	computing costs in relation to estimates
Corrective action	expediting	issuing production and procurement orders	initiating full inspection; adjusting processes	adjusting selling price of product
Evaluation	estimating production capacity and maintenance schedules	drawing up replenishment policies and inventory systems	reassessing specifications; improving processes and procedures	evaluating production economics; improving data

workers and machines. Today the field is characterized by an emphasis on mathematical and computer modeling.

THE EVOLVING NATURE OF INDUSTRIAL ENGINEERING

In recent years industrial engineering has broadened significantly as a discipline, and the support it now provides to production and manufacturing managers comes from staff specialists drawn not only from the field of industrial engineering but also from operations research, management science, computer science, and information systems. In the 1970s and 1980s industrial engineering became a more quantitative and computer-based profession, and operations research techniques were adopted as the core of most industrial engineering academic curricula in both the United States and Europe.

Since many of the problems of operations research originate in industrial production systems, it is often difficult to determine where the engineering discipline ends and the more basic scientific discipline begins (operations research is a branch of applied mathematics). Indeed, many academic industrial engineering departments now use the term industrial engineering and operations research or the reverse, further clouding the distinction.

The tools and techniques of operations research are based on mathematical formulas, most of which have been developed since World War II. This section discusses developments that have proved useful in the field of industrial engineering and that form the basis for the contemporary definition of the field. (W.K.H.)

SCIENTIFIC METHOD AND OPERATIONS RESEARCH

Basic aspects. Operations research is the application of scientific method to the management of organized systems such as industrial production systems, government and social programs, and defense systems. It is called operational research in the United Kingdom, where it originated. Operations research attempts to provide those who manage organized systems with an objective and quantitative basis for decision; it is normally carried out by teams of scientists and engineers drawn from a variety of disciplines. Thus, operations research is not a science itself but rather the application of science to the solution of managerial and administrative problems, and it focuses on the performance of organized systems taken as a whole rather than on their parts taken separately. Usually concerned with systems in which human behaviour plays an important part, operations research differs in this respect from systems engineering, which, using a similar approach, tends to concentrate on systems in which human behaviour is not important. Operations research was originally concerned with improving the operations of existing systems rather than developing new ones; the converse was true of systems engineering. This difference, however, has been disappearing as both fields have matured.

Basic aspects

The subject matter of operations research consists of decisions that control the operations of systems. Hence, it is concerned with how managerial decisions are and should be made, how to acquire and process data and information required to make decisions effectively, how to monitor decisions once they are implemented, and how to organize the decision-making and decision-implementation process. Extensive use is made of older disciplines such as logic, mathematics, and statistics, as well as more recent scientific developments such as communications theory, decision theory, cybernetics, organization theory, the behavioral sciences, and general systems theory.

In the 19th century the Industrial Revolution involved mechanization or replacement of human by machine as a source of physical work. Study and improvement of such work formed the basis of the field of industrial engineering. Many contemporary issues are concerned with automation or mechanization of mental work. The primary technologies involved are mechanization of symbol generation (observation by machines such as radar and sonar), mechanization of symbol transmission (communication by telephone, radio, and television), and mechanization of logical manipulation of symbols (data processing and decision making by computer). Operations research applies the scientific method to the study of mental work and provides the knowledge and understanding required to make effective use of personnel and machines to carry it out.

History. In a sense, every effort to apply science to management of organized systems, and to their understanding, was a predecessor of operations research. It began as a separate discipline, however, in 1937 in Britain as a result of the initiative of A.P. Rowe, superintendent of the Bawdsey Research Station, who led British scientists to teach military leaders how to use the then newly developed radar to locate enemy aircraft. By 1939 the Royal Air Force formally commenced efforts to extend the range of radar equipment so as to increase the time between the first warning provided by radar and the attack by enemy aircraft. At first they analyzed physical equipment and communication networks, but later they examined behaviour of the operating personnel and relevant executives. Results of the studies revealed ways of improving the operators' techniques and also revealed unappreciated limitations in the network.

Similar developments took place in the British Army and the Royal Navy, and in both cases radar again was the instigator. In the army, use of operations research had grown out of the initial inability to use radar effectively in controlling the fire of antiaircraft weapons. Since the traditional way of testing equipment did not seem to apply to radar gunsights, scientists found it necessary to test in the field under operating conditions, and the distinguished British physicist and future Nobel Laureate P.M.S. Blackett organized a team to solve the antiaircraft problem. Blackett's Antiaircraft Command Research Group included two physiologists, two mathematical physicists, an astrophysicist, an army officer, a former surveyor, and subsequently a third physiologist, a general physicist, and two mathematicians.

Radar testing problem

By 1942 formal operations research groups had been established in all three of Britain's military services.

Development of operations research paralleling that in Britain took place in Australia, Canada, France, and, most significantly for future developments, in the United States, which was the beneficiary of a number of contacts with British researchers. Sir Robert Watson-Watt, who with A.P. Rowe launched the first two operational studies of radar in 1937 and who claims to have given the discipline its name, visited the United States in 1942 and urged that operations research be introduced into the War and Navy departments. Reports of the British work had already been sent from London by American observers, and James B. Conant, then chairman of the National Defense Research Committee, had become aware of operations research during a visit to England in the latter half of 1940. Another stimulant was Blackett's memorandum, "Scientists at the Operational Level," of December 1941, which was widely circulated in the U.S. service departments.

The first organized operations research activity in the United States began in 1942 in the Naval Ordnance Laboratory. This group, which dealt with mine warfare problems, was later transferred to the Navy Department, from which it designed the aircraft mining blockade of the Inland Sea of Japan.

As in Britain, radar stimulated developments in the U.S. Air Force. In October 1942, all Air Force commands were urged to include operations research groups in their staffs. By the end of World War II there were 26 such groups in the Air Force. In 1943 Gen. George Marshall suggested to all theatre commanders that they form teams to study amphibious and ground operations.

At the end of World War II a number of British operations research workers moved to government and industry. Nationalization of several British industries was an important factor. One of the first industrial groups was established at the National Coal Board. Electricity and transport, both nationalized industries, began to use operations research shortly thereafter. Parts of the private sector began to follow suit, particularly in those industries with cooperative research associations; for example, in the British Iron and Steel Research Association.

The early development of industrial operations research was cautious, and for some years most industrial groups were quite small. In the late 1950s, largely stimulated

by developments in the United States, the development of industrial operations research in Britain was greatly accelerated.

ndustrial nvolvement in he United States

Although in the United States military research increased at the end of the war, and groups were expanded, it was not until the early 1950s that American industry began to take operations research seriously. The advent of the computer brought an awareness of a host of broad system problems and the potentiality for solving them, and within the decade about half the large corporations in the United States began to use operations research. Elsewhere the technique also spread through industry.

Societies were organized, beginning with the Operational Research Club of Britain, formed in 1948, which in 1954 became the Operational Research Society. The Operations Research Society in America was formed in 1952. Many other national societies appeared; the first international conference on operations research was held at Oxford University in 1957. In 1959 an International Federation of Operational Research Societies was formed.

The first appearance of operations research as an academic discipline came in 1948 when a course in nonmilitary techniques was introduced at the Massachusetts Institute of Technology in Cambridge. In 1952 a curriculum leading to a master's and doctoral degree was established at the Case Institute of Technology (now Case Western Reserve University) in Cleveland. Since then many major academic institutions in the United States have introduced programs. In the United Kingdom courses were initiated at the University of Birmingham in the early 1950s. The first chair in operations research was created at the newly formed University of Lancaster in 1964. Similar developments have taken place in most countries in which a national operations research society exists.

The first scholarly journal, the *Operational Research Quarterly,* published in the United Kingdom, was initiated in 1950; in 1978 its name was changed to the *Journal of the Operational Research Society.* It was followed in 1952 by the *Journal of the Operations Research Society of America,* which was renamed *Operations Research* in 1955. The International Federation of Operational Research Societies initiated the *International Abstracts in Operations Research* in 1961.

Despite its rapid growth, operations research is still a relatively young scientific activity. Its techniques and methods, and the areas to which they are applied, can be expected to continue to expand rapidly. Most of its history lies in the future.

ESSENTIAL CHARACTERISTICS

Three essential characteristics of operations research are a systems orientation, the use of interdisciplinary teams, and the application of scientific method to the conditions under which the research is conducted.

Systems orientation. The systems approach to problems recognizes that the behaviour of any part of a system has some effect on the behaviour of the system as a whole. Even if the individual components are performing well, however, the system as a whole is not necessarily performing equally well. For example, assembling the best of each type of automobile part, regardless of make, does not necessarily result in a good automobile or even one that will run, because the parts may not fit together. It is the interaction between parts, and not the actions of any single part, that determines how well a system performs.

Interrelationships between parts of a system

Thus, operations research attempts to evaluate the effect of changes in any part of a system on the performance of the system as a whole and to search for causes of a problem that arises in one part of a system in other parts or in the interrelationships between parts. In industry, a production problem may be approached by a change in marketing policy. For example, if a factory fabricates a few profitable products in large quantities and many less profitable items in small quantities, long efficient production runs of high-volume, high-profit items may have to be interrupted for short runs of low-volume, low-profit items. An operations researcher might propose reducing the sales of the less profitable items and increasing those of the profitable items by placing salesmen on an incentive system that especially compensates them for selling particular items.

The interdisciplinary team. Scientific and technological disciplines have proliferated rapidly in the last 100 years. The proliferation, resulting from the enormous increase in scientific knowledge, has provided science with a filing system that permits a systematic classification of knowledge. This classification system is helpful in solving many problems by identifying the proper discipline to appeal to for a solution. Difficulties arise when more complex problems, such as those arising in large organized systems, are encountered. It is then necessary to find a means of bringing together diverse disciplinary points of view. Furthermore, since methods differ among disciplines, the use of interdisciplinary teams makes available a much larger arsenal of research techniques and tools than would otherwise be available. Hence, operations research may be characterized by rather unusual combinations of disciplines on research teams and by the use of varied research procedures.

Methodology. Until the 20th century, laboratory experiments were the principal and almost the only method of conducting scientific research. But large systems such as are studied in operations research cannot be brought into laboratories. Furthermore, even if systems could be brought into the laboratory, what would be learned would not necessarily apply to their behaviour in their natural environment, as shown by early experience with radar. Experiments on systems and subsystems conducted in their natural environment ("operational experiments") are possible as a result of the experimental methods developed by the British statistician R.A. Fisher in 1923–24. For practical or even ethical reasons, however, it is seldom possible to experiment on large organized systems as a whole in their natural environments. This results in an apparent dilemma: to gain understanding of complex systems experimentation seems to be necessary, but it cannot usually be carried out. This difficulty is solved by the use of models, representations of the system under study. Provided the model is good, experiments (called "simulations") can be conducted on it, or other methods can be used to obtain useful results.

PHASES OF OPERATIONS RESEARCH

Problem formulation. To formulate an operations research problem, a suitable measure of performance must be devised, various possible courses of action defined (that is, controlled variables and the constraints upon them), and relevant uncontrolled variables identified. To devise a measure of performance, objectives are identified and defined, and then quantified. If objectives cannot be quantified or expressed in rigorous (usually mathematical) terms, most operations research techniques cannot be applied. For example, a business manager may have the acquisitive objective of introducing a new product and making it profitable within one year. The identified objective is profit in one year, which is defined as receipts less costs, and would probably be quantified in terms of sales. In the real world, conditions may change with time. Thus, though a given objective is identified at the beginning of the period, change and reformulation are frequently necessary.

Determining who makes decisions

Detailed knowledge of how the system under study actually operates and of its environment is essential. Such knowledge is normally acquired through an analysis of the system, a four-step process that involves determining whose needs or desires the organization tries to satisfy; how these are communicated to the organization; how information on needs and desires penetrates the organization; and what action is taken, how it is controlled, and what the time and resource requirements of these actions are. This information can usually be represented graphically in a flowchart, which enables researchers to identify the variables that affect system performance.

Once the objectives, the decision makers, their courses of action, and the uncontrolled variables have been identified and defined, a measure of performance can be developed and selection can be made of a quantitative function of this measure to be used as a criterion for the best solution.

The type of decision criterion that is appropriate to a problem depends on the state of knowledge regarding pos-

sible outcomes. Certainty describes a situation in which each course of action is believed to result in one particular outcome. Risk is a situation in which, for each course of action, alternative outcomes are possible, the probabilities of which are known or can be estimated. Uncertainty describes a situation in which, for each course of action, probabilities cannot be assigned to the possible outcomes.

In risk situations, which are the most common in practice, the objective normally is to maximize expected (long-run average) net gain or gross gain for specified costs, or to minimize costs for specified benefits. A business, for example, seeks to maximize expected profits or minimize expected costs. Other objectives, not necessarily related, may be sought; for example, an economic planner may wish to maintain full employment without inflation; or different groups within an organization may have to compromise their differing objectives, as when an army and a navy, for example, must cooperate in matters of defense.

In approaching uncertain situations one may attempt either to maximize the minimum gain or minimize the maximum loss that results from a choice; this is the "minimax" approach. Alternatively, one may weigh the possible outcomes to reflect one's optimism or pessimism and then apply the minimax principle. A third approach, "minimax regret," attempts to minimize the maximum deviation from the outcome that would have been selected if a state of certainty had existed before the choice had been made.

Each identified variable should be defined in terms of the conditions under which, and research operations by which, questions concerning its value ought to be answered; this includes identifying the scale used in measuring the variable.

Model construction. A model is a simplified representation of the real world and, as such, includes only those variables relevant to the problem at hand. A model of freely falling bodies, for example, does not refer to the colour, texture, or shape of the body involved. Furthermore, a model may not include all relevant variables because a small percentage of these may account for most of the phenomenon to be explained. Many of the simplifications used produce some error in predictions derived from the model, but these can often be kept small compared to the magnitude of the improvement in operations that can be extracted from them. Most operations research models are symbolic models because symbols represent properties of the system. The earliest models were physical representations such as model ships, airplanes, tow tanks, and wind tunnels. Physical models are usually fairly easy to construct, but only for relatively simple objects or systems, and are usually difficult to change.

The next step beyond the physical model is the graph, easier to construct and manipulate but more abstract. Since graphic representation of more than three variables is difficult, symbolic models came into use. There is no limit to the number of variables that can be included in a symbolic model, and such models are easier to construct and manipulate than physical models.

Symbolic models

Symbolic models are completely abstract. When the symbols in a model are defined, the model is given content or meaning. This has important consequences. Symbolic models of systems of very different content often reveal similar structure. Hence, most systems and problems arising in them can be fruitfully classified in terms of relatively few structures. Furthermore, since methods of extracting solutions from models depend only on their structure, some methods can be used to solve a wide variety of problems from a contextual point of view. Finally, a system that has the same structure as another, however different the two may be in content, can be used as a model of the other. Such a model is called an analogue. By use of such models much of what is known about the first system can be applied to the second.

Despite the obvious advantages of symbolic models there are many cases in which physical models are still useful, as in testing physical structures and mechanisms; the same is true for graphic models. Physical and graphic models are frequently used in the preliminary phases of constructing symbolic models of systems.

Operations research models represent the causal relationship between the controlled and uncontrolled variables and system performance; they must therefore be explanatory, not merely descriptive. Only explanatory models can provide the requisite means to manipulate the system to produce desired changes in performance.

Operations research analysis is directed toward establishing cause-and-effect relations. Though experiments with actual operations of all or part of a system are often useful, these are not the only way to analyze cause and effect. There are four patterns of model construction, only two of which involve experimentation: inspection, use of analogues, operational analysis, and operational experiments. They are considered here in order of increasing complexity.

In some cases the system and its problem are relatively simple and can be grasped either by inspection or from discussion with persons familiar with it. In general, only low-level and repetitive operating problems, those in which human behaviour plays a minor role, can be so treated.

When the researcher finds it difficult to represent the structure of a system symbolically, it is sometimes possible to establish a similarity, if not an identity, with another system whose structure is better known and easier to manipulate. It may then be possible to use either the analogous system itself or a symbolic model of it as a model of the problem system. For example, an equation derived from the kinetic theory of gases has been used as a model of the movement of trains between two classification yards. Hydraulic analogues of economies and electronic analogues of automotive traffic have been constructed with which experimentation could be carried out to determine the effects of manipulation of controllable variables. Thus, analogues may be constructed as well as found in existing systems.

Uncovering causal structure

In some cases analysis of actual operations of a system may reveal its causal structure. Data on operations are analyzed to yield an explanatory hypothesis, which is tested by analysis of operating data. Such testing may lead to revision of the hypothesis. The cycle is continued until a satisfactory explanatory model is developed.

For example, an analysis of the cars stopping at urban automotive service stations located at intersections of two streets revealed that almost all came from four of the 16 possible routes through the intersection (four ways of entering times four ways of leaving). Examination of the percentage of cars in each route that stopped for service suggested that this percentage was related to the amount of time lost by stopping. Data were then collected on time lost by cars in each route. This revealed a close inverse relationship between the percentage stopping and time lost. But the relationship was not linear; that is, the increases in one were not proportional to increases in the other. It was then found that perceived lost time exceeded actual lost time, and the relationship between the percentage of cars stopping and perceived lost time was close and linear. The hypothesis was systematically tested and verified and a model constructed that related the number of cars stopping at service stations to the amount of traffic in each route through its intersection and to characteristics of the station that affect the time required to get service.

In situations where it is not possible to isolate the effects of individual variables by analysis of operating data, it may be necessary to resort to operational experiments to determine which variables are relevant and how they affect system performance.

Such is the case, for example, in attempts to quantify the effects of advertising (amount, timing, and media used) upon sales of a consumer product. Advertising by the producer is only one of many controlled and uncontrolled variables affecting sales. Hence, in many cases its effect can only be isolated and measured by controlled experiments in the field.

The same is true in determining how the size, shape, weight, and price of a food product affect its sales. In this case laboratory experiments on samples of consumers can be used in preliminary stages, but field experiments are eventually necessary. Experiments do not yield explanatory theories, however. They can only be used to test explanatory hypotheses formulated before designing

Testing hypotheses

the experiment and to suggest additional hypotheses to be tested.

It is sometimes necessary to modify an otherwise acceptable model because it is not possible or practical to find the numerical values of the variables that appear in it. For example, a model to be used in guiding the selection of research projects may contain such variables as "the probability of success of the project," "expected cost of the project," and its "expected yield." But none of these may be calculable with any reliability.

Models not only assist in solving problems but also are useful in formulating them; that is, models can be used as guides to explore the structure of a problem and to reveal possible courses of action that might otherwise be missed. In many cases the course of action revealed by such application of a model is so obviously superior to previously considered possibilities that justification of its choice is hardly required.

In some cases the model of a problem may be either too complicated or too large to solve. It is frequently possible to divide the model into individually solvable parts and to take the output of one model as an input to another. Since the models are likely to be interdependent, several repetitions of this process may be necessary.

Deriving solutions from models. Procedures for deriving solutions from models are either deductive or inductive. With deduction one moves directly from the model to a solution in either symbolic or numerical form. Such procedures are supplied by mathematics; for example, the calculus. An explicit analytical procedure for finding the solution is called an algorithm.

Algorithms and heuristics

Even if a model cannot be solved, and many are too complex for solution, it can be used to compare alternative solutions. It is sometimes possible to conduct a sequence of comparisons, each suggested by the previous one and each likely to contain a better alternative than was contained in any previous comparison. Such a solution-seeking procedure is called heuristic.

Inductive procedures involve trying and comparing different values of the controlled variables. Such procedures are said to be "iterative" (repetitive) if they proceed through successively improved solutions until either an optimal solution is reached or further calculation cannot be justified. A rational basis for terminating such a process—known as "stopping rules"—involves the determination of the point at which the expected improvement of the solution on the next trial is less than the cost of the trial.

Such well-known algorithms as linear, nonlinear, and dynamic programming are iterative procedures based on mathematical theory. Simulation and experimental optimization are iterative procedures based primarily on statistics.

Testing the model and the solution. A model may be deficient because it includes irrelevant variables, excludes relevant variables, contains inaccurately evaluated variables, is incorrectly structured, or contains incorrectly formulated constraints. Tests for deficiencies of a model are statistical in nature; their use requires knowledge of sampling and estimation theory, experimental designs, and the theory of hypothesis testing (see also STATISTICS).

Sampling

Sampling-estimation theory is concerned with selecting a sample of items from a large group and using their observed properties to characterize the group as a whole. To save time and money, the sample taken is as small as possible. Several theories of sampling design and estimation are available, each yielding estimates with different properties.

The structure of a model consists of a function relating the measure of performance to the controlled and uncontrolled variables; for example, a business may attempt to show the functional relationship between profit levels (the measure of performance) and controlled variables (prices, amount spent on advertising) and uncontrolled variables (economic conditions, competition). In order to test the model, values of the measure of performance computed from the model are compared with actual values under different sets of conditions. If there is a significant difference between these values, or if the variability of these differences is large, the model requires repair. Such tests do not use data that have been used in constructing the model, because to do so would determine how well the model fits performance data from which it has been derived, not how well it predicts performance.

The solution derived from a model is tested to find whether it yields better performance than some alternative, usually the one in current use. The test may be prospective, against future performance, or retrospective, comparing solutions that would have been obtained had the model been used in the past with what actually did happen. If neither prospective nor retrospective testing is feasible, it may be possible to evaluate the solution by "sensitivity analysis," a measurement of the extent to which estimates used in the solution would have to be in error before the proposed solution performs less satisfactorily than the alternative decision procedure.

The cost of implementing a solution should be subtracted from the gain expected from applying it, thus obtaining an estimate of net improvement. Where errors or inefficiencies in applying the solution are possible, these should also be taken into account in estimating the net improvement.

Implementing and controlling the solution. The acceptance of a recommended solution by the responsible manager depends on the extent to which he believes the solution to be superior to alternatives. This in turn depends on his faith in the researchers involved and their methods. Hence, participation by managers in the research process is essential for success.

Using the results of models

Operations researchers are normally expected to oversee implementation of an accepted solution. This provides them with an ultimate test of their work and an opportunity to make adjustments if any deficiencies should appear in application. The operations research team prepares detailed instructions for those who will carry out the solution and trains them in following these instructions. The cooperation of those who carry out the solution and those who will be affected by it should be sought in the course of the research process, not after everything is done. Implementation plans and schedules are pretested and deficiencies corrected. Actual performance of the solution is compared with expectations and, where divergence is significant, the reasons for it are determined and appropriate adjustments made.

The solution may fail to yield expected performance for one or a combination of reasons: the model may be wrongly constructed or used; the data used in making the model may be incorrect; the solution may be incorrectly carried out; the system or its environment may have changed in unexpected ways after the solution was applied. Corrective action is required in each case.

Controlling a solution requires deciding what constitutes a significant deviation in performance from expectations; determining the frequency of control checks, the size and type of sample of observations to be made, and the types of analyses of the resulting data that should be carried out; and taking appropriate corrective action. The second step should be designed to minimize the sum of the costs of carrying out the control procedures and the errors that might be involved.

Since most models involve a variety of assumptions, these are checked systematically. Such checking requires explicit formulation of the assumptions made during construction of the model.

Effective controls not only make possible but often lead to better understanding of the dynamics of the system involved. Through controls the problem-solving system of which operations research is a part learns from its own experience and adapts more effectively to changing conditions.

COMPUTERS AND OPERATIONS RESEARCH

Simulation. Computers have had a dramatic impact on the management of industrial production systems and the fields of operations research and industrial engineering. The speed and data-handling capabilities of computers allow engineers and scientists to build larger, more realistic models of organized systems and to get meaningful solutions to those models through the use of simulation techniques.

Simulation consists of calculating the performance of a system by evaluating a model of it for randomly selected values of variables contained within it. Most simulation in operations research is concerned with "stochastic" variables; that is, variables whose values change randomly within some probability distribution over time. The random sampling employed in simulation requires either a supply of random numbers or a procedure for generating them. It also requires a way of converting these numbers into the distribution of the relevant variable, a way of sampling these values, and a way of evaluating the resulting performance.

Simulation in military games

A simulation in which decision making is performed by one or more real decision makers is called "operational gaming." Such simulations are commonly used in the study of interactions of decision makers as in competitive situations. Military gaming has long been used as a training device, but only relatively recently has it been used for research purposes. There is still considerable difficulty, however, in drawing inferences from operational games to the real world.

Experimental optimization is a means of experimenting on a system so as to find the best solution to a problem within it. Such experiments, conducted either simultaneously or sequentially, may be designed in various ways, no one of which is best in all situations. (R.L.A./W.K.H.)

Decision analysis and support. Since their widespread introduction in business and government organizations in the 1950s, the primary applications of computers have been in the areas of record keeping, bookkeeping, and transaction processing. These applications, commonly called data processing, automate the flow of paperwork, account for business transactions (such as order processing and inventory and shipping activities), and maintain orderly and accurate records. Although data processing is vital to most organizations, most of the work involved in the design of such systems does not require the methods of operations research.

In the 1960s, when computers were applied to the routine decision-making problems of managers, management information systems (MIS) emerged. These systems use the raw (usually historical) data from data-processing systems to prepare management summaries, to chart information on trends and cycles, and to monitor actual performance against plans or budgets.

MIS and DSS

More recently, decision support systems (DSS) have been developed to project and predict the results of decisions before they are made. These projections permit managers and analysts to evaluate the possible consequences of decisions and to try several alternatives on paper before committing valuable resources to actual programs.

The development of management information systems and decision support systems brought operations researchers and industrial engineers to the forefront of business planning. These computer-based systems require knowledge of an organization and its activities in addition to technical skills in computer programming and data handling. The key issues in MIS or DSS include how a system will be modeled, how the model of the system will be handled by the computer, what data will be used, how far into the future trends will be extrapolated, and so on. In much of this work, as well as in more traditional operations research modeling, simulation techniques have proved invaluable.

Personal computers and business

New software tools for decision making. The explosive growth of personal computers in business organizations in the early 1980s spawned a parallel growth in software to assist in decision making. These tools include spreadsheet programs for analyzing complex problems with trails that have different sets of data, data base management programs that permit the orderly maintenance and manipulation of vast amounts of information, and graphics programs that quickly and easily prepare professional-looking displays of data. Business programs (software) like these once cost tens of thousands of dollars; now they are widely available, may be used on relatively inexpensive hardware, are easy to use without learning a programming language, and are powerful enough to handle sophisticated, practical business problems.

The availability of spreadsheet, data base, and graphics programs on personal computers has also greatly aided industrial engineers and operations researchers whose work involves the construction, solution, and testing of models. Easy-to-use software that does not require extensive programming knowledge permits faster, more cost-effective model building and is also helpful in communicating the results of analysis to management. Indeed, many managers now have a computer on their desk and work with spreadsheets and other programs as a routine part of their managerial duties. (W.K.H.)

EXAMPLES OF OPERATIONS RESEARCH MODELS AND APPLICATIONS

As previously mentioned, many operational problems of organized systems have common structures. The most common types of structure have been identified as prototype problems, and extensive work has been done on modeling and solving them.

Though all the problems with similar structures do not have the same model, those that apply to them may have a common mathematical structure and hence may be solvable by one procedure. Some real problems consist of combinations of smaller problems, some or all of which fall into different prototypes. In general, prototype models are the largest that can be solved in one step. Hence, large problems that consist of combinations of prototype problems usually must be broken down into solvable units; the overall model used is an aggregation of prototype and possibly other models.

Solving allocation problems

Resource allocation. Allocation problems involve the distribution of resources among competing alternatives in order to minimize total costs or maximize total return. Such problems have the following components: a set of resources available in given amounts; a set of jobs to be done, each consuming a specified amount of resources; and a set of costs or returns for each job and resource. The problem is to determine how much of each resource to allocate to each job.

If more resources are available than needed, the solution should indicate which resources are not to be used, taking associated costs into account. Similarly, if there are more jobs than can be done with available resources, the solution should indicate which jobs are not to be done, again taking into account the associated costs.

If each job requires exactly one resource (*e.g.*, one person) and each resource can be used on only one job, the resulting problem is one of assignment. If resources are divisible, and if both jobs and resources are expressed in units on the same scale, it is termed a transportation or distribution problem. If jobs and resources are not expressed in the same units, it is a general allocation problem.

An assignment problem may consist of assigning workers to offices or jobs, trucks to delivery routes, drivers to trucks, or classes to rooms. A typical transportation problem involves distribution of empty railroad freight cars where needed or the assignment of orders to factories for production. The general allocation problem may consist of determining which machines should be employed to make a given product or what set of products should be manufactured in a plant during a particular period.

In allocation problems the unit costs or returns may be either independent or interdependent; for example, the return from investing a dollar in selling effort may depend on the amount spent on advertising. If the allocations made in one period affect those in subsequent periods, the problem is said to be dynamic, and time must be considered in its solution. (R.L.A.)

Linear programming. Linear programming (LP) refers to a family of mathematical optimization techniques that have proved effective in solving resource allocation problems, particularly those found in industrial production systems. Linear programming methods are algebraic techniques based on a series of equations or inequalities that limit a problem and are used to optimize a mathematical expression called an objective function. The objective function and the constraints placed upon the problem must be deterministic and able to be expressed in linear form. These restrictions limit the number of problems that can

be handled directly, but since the introduction of linear programming in the late 1940s, much progress has been made to adapt the method to more complex problems.

Since linear programming is probably the most widely used mathematical optimization technique, numerous computer programs are available for solving LP problems. For example, LP techniques are now used routinely for such problems as oil and chemical refinery blending, choosing vendors or suppliers for large, multiplant manufacturing corporations, determining shipping routes and schedules, and managing and maintaining truck fleets. (W.K.H.)

Inventory control. Inventories include raw materials, component parts, work in process, finished goods, packing and packaging materials, and general supplies. The control of inventories, vital to the financial strength of a firm, in general involves deciding at what points in the production system stocks shall be held and what their form and size are to be. As some unit costs increase with inventory size—including storage, obsolescence, deterioration, insurance, investment—and other unit costs decrease with inventory size—including setup or preparation costs, delays because of shortages, and so forth—a good part of inventory management consists of determining optimal purchase or production lot sizes and base stock levels that will balance the opposing cost influences. Another part of the general inventory problem is deciding the levels (reorder points) at which orders for replenishment of inventories are to be initiated.

Principal control systems

Inventory control is concerned with two questions: when to replenish the store and by how much. There are two main control systems. The two-bin system (sometimes called the min-max system) involves the use of two bins, either physically or on paper. The first bin is intended for supplying current demand and the second for satisfying demand during the replenishment period. When the stock in the first bin is depleted, an order for a given quantity is generated. The reorder-cycle system, or cyclical-review system, consists of ordering at fixed regular intervals. Various combinations of these systems can be used in the construction of an inventory-control procedure. A pure two-bin system, for example, can be modified to require cyclical instead of continuous review of stock, with orders being generated only when the stock falls below a specific level. Similarly, a pure reorder-cycle system can be modified to allow orders to be generated if the stock falls below the reorder level between the cyclical reviews. In yet another variation, the reorder quantity in the reorder-cycle system is made to depend on the stock level at the review period or the need to order other products or materials at the same time or both. (S.E.)

The classic inventory problem involves determining how much of a resource to acquire, either by purchasing or producing it, and whether or when to acquire it to minimize the sum of the costs that increase with the size of inventory and those that decrease with increases in inventory. Costs of the first type include the cost of the capital invested in inventory, handling, storage, insurance, taxes, depreciation, deterioration, and obsolescence. Costs that decrease as inventory increases include shortage costs (arising from lost sales), production setup costs, and the purchase price or direct production costs. Setup costs include the cost of placing a purchase order or starting a production run. If large quantities are ordered, inventories increase but the frequency of ordering decreases, hence setup costs decrease. In general, the larger the quantity ordered the lower the unit purchase price because of quantity discounts and the lower production cost per unit resulting from the greater efficiency of long production runs. Other relevant variables include demand for the resource and the time between placing and filling orders.

An inventory example

Figure 3 portrays a highly simplified inventory situation for a single item held in stock. The solid line charts a case in which the stock is replenished in quantities (batches or lots) of 10 units. Assuming that there is no initial stock when the first replenishment batch arrives, the inventory on hand immediately goes from zero to 10 units. Then, as the stock is used or sold, the inventory (or stock on hand) gradually declines. In this simple example the usage rate is a constant one unit per day. When the solid-line diagram is extended over several periods, it points out the necessity of ordering every 10 days, or about 25 times per year, assuming a five-day week and 50 weeks per year.

The dashed line on the diagram depicts a different situation, in which replenishment batches of 12 units are ordered. With the same fixed usage rate of one unit per day, each reorder lasts for 12 days instead of 10. The larger quantity permits the business to get through the year with fewer (20 or 21) orders and results in a lowering of the total costs associated with ordering. The average inventory in the system, however, is higher with the second policy than with the first, and the cost-effectiveness of the policy depends on the relationship between the costs of ordering and the costs of carrying the inventory. In situations like this, formulas developed by operations researchers are used to calculate the correct order quantity for various market conditions.

In this simplified portrayal, inventory varies uniformly between zero units and the order quantity; the average inventory, therefore, can be expressed as the order quantity divided by two, or five units in the first example and six units in the second.

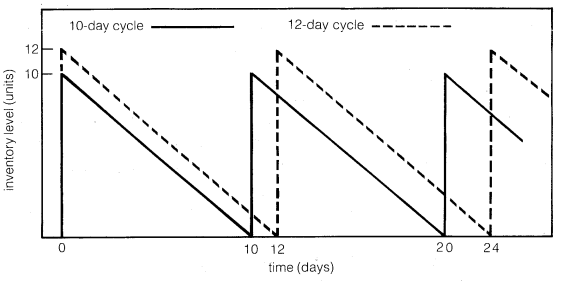


Figure 3: Simple inventory model for a single item held in stock.

Applications in inventory problems

Inventory problems arise in a wide variety of contexts; for example, determining quantities of goods to be purchased or produced, how many people to hire or train, how large a new production or retailing facility should be or how many should be provided, and how much fluid (operating) capital to keep available. Inventory models for single items are well developed and are normally solved with calculus. When the order quantities for many items are interdependent (as, for example, when there is limited storage space or production time) the problem is more difficult. Some of the larger problems can be solved by breaking them into interacting inventory and allocation problems. In very large problems simulation can be used to test various relevant decision rules. (R.L.A./W.K.H.)

Japanese approaches. In the 1970s several Japanese firms, led by the Toyota Motor Corporation, developed radically different approaches to the management of inventories. Coined the "just-in-time" approach, the basic element of the new systems was the dramatic reduction of inventories throughout the total production system. By relying on careful scheduling and the coordination of supplies, the Japanese ensured that parts and supplies were available in the right quantity, with proper quality, at the exact time they were needed in the manufacturing or assembly process.

Two things made just-in-time work—a dogged attention to quality at all levels of the total system obviated the need for parts inventories to cover defectives found in the manufacturing process, and a close coordination of information and plans with suppliers and vendors permitted them to align their schedules and shipments with the last-minute needs of the manufacturer. Elements of the just-in-time approach now have been adopted by numerous companies in the United States and Europe, although many cannot use the system to its fullest extent because their supplier networks are larger and more widely dispersed than in Japan.

The *kanban* system

A second Japanese technique, called *kanban* ("card"), also permits Japanese firms to schedule production and manage inventories more effectively. In the *kanban* system, cards or tickets are attached to batches, racks, or

pallet loads of parts in the manufacturing process. When a batch is depleted in the assembly process, its *kanban* is returned to the manufacturing department and another batch is shipped immediately. Since the total number of parts or batches in the system is held constant, the coordination, scheduling, and control of the inventory is greatly simplified. (W.K.H.)

Replacement and maintenance. Replacement problems involve items that degenerate with use or with the passage of time and those that fail after a certain amount of use or time. Items that deteriorate are likely to be large and costly (*e.g.,* machine tools, trucks, ships, and home appliances). Nondeteriorating items tend to be small and relatively inexpensive (*e.g.,* light bulbs, vacuum tubes, ink cartridges). The longer a deteriorating item is operated the more maintenance it requires to maintain efficiency. Furthermore, the longer such an item is kept the less is its resale value and the more likely it is to be made obsolete by new equipment. If the item is replaced frequently, however, investment costs increase. Thus the problem is to determine when to replace such items and how much maintenance (particularly preventive) to perform so that the sum of the operating, maintenance, and investment costs is minimized.

In the case of nondeteriorating items the problem involves determining whether to replace them as a group or to replace individuals as they fail. Though group replacement is wasteful, labour cost of replacements is greater when done singly; for example, the light bulbs in a large subway system may be replaced in groups to save labour. Replacement problems that involve minimizing the costs of items, failures, and the replacement labour are solvable either by numerical analysis or simulation.

The "items" involved in replacement problems may be people. If so, maintenance can be interpreted as training or improvements in salary, status, or fringe benefits. Failure can be interpreted as departure, and investment as recruiting, hiring, and initial training costs. There are many additional complexities in such cases; for example, the effect of one person's resigning or being promoted on the behaviour of others. Such controllable aspects of the environment as location of work and working hours can have a considerable effect on productivity and failure rates. In problems of this type, the inputs of the behavioral sciences are particularly useful.

Queuing. A queue is a waiting line, and queuing involves dealing with items or people in sequence. Thus, a queuing problem consists either of determining what facilities to provide or scheduling the use of them. The cost of providing service and the waiting time of users are minimized. Examples of such problems include determining the number of checkout counters to provide at a supermarket, runways at an airport, parking spaces at a shopping centre, or tellers in a bank. Many maintenance problems can be treated as queuing problems; items requiring repair are like users of a service. Some inventory problems may also be formulated as queuing problems in which orders are like users and stocks are like service facilities.

Job shop sequencing. In queuing problems, the order in which users waiting for service are served is always specified. Selection of that order so as to minimize some function of the time to perform all the tasks is a sequencing problem. The performance measure may account for total elapsed time, total tardiness in meeting deadlines or due dates, and the cost of in-process inventories.

The most common context for sequencing problems is a batch, or job shop, production facility that processes many different products with many combinations of machines. In this context account may have to be taken of such factors as overlapping service (that is, if a customer consists of a number of items to be taken through several steps of a process, the first items completing the initial step may start on the second step before the last one finishes the first), transportation time between service facilities, correction of service breakdowns, facility breakdowns, and material shortages. (R.L.A.)

A job shop example

A simplified job shop sequencing problem, with two jobs and four machines, is shown in Figure 4. At the top of the figure is the operations sequence of the two jobs. Job A must go first to machine 1, then to 2, then to 3, and finally to 4, and the order of processing on the four machines cannot be changed. The processing time for the job is one hour at each machine, for a total of four hours of machining time. In this example, the job can only be on one machine at a time, as if the job consisted of a single product being processed through four machine tools.

Job B must follow a different sequence. It also starts on machine 1, but then it goes to machine 4, then to 2, and finally back to machine 4. Each machining operation on Job B also requires one hour.

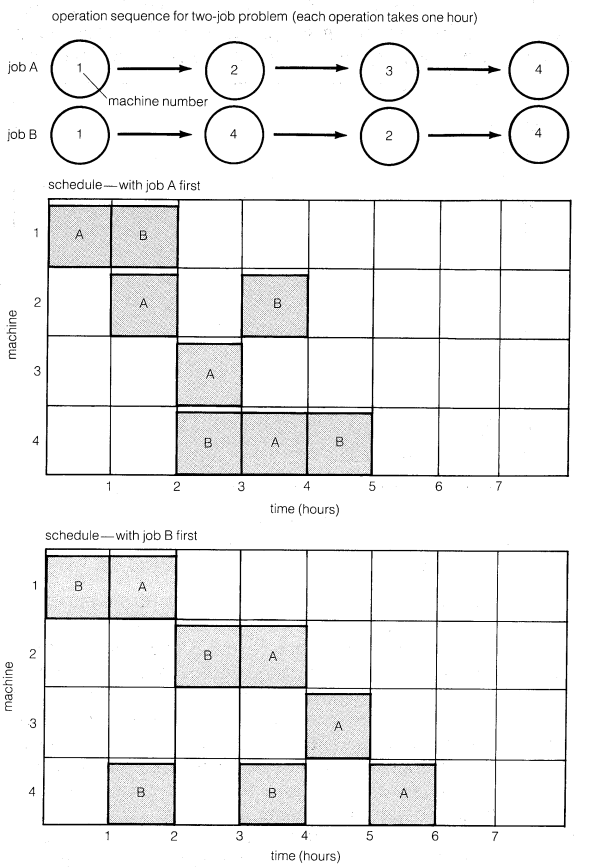


Figure 4: Job shop sequencing problem with two solutions.

Underneath the charts showing the required sequence of operations, two alternative schedules are shown for the two jobs. (In a bar chart, time is shown on the horizontal line, and the bars or blocks represent the time that each operation is scheduled on each of the four machines.) The first schedule assumes that Job A is run first. Once Job A is laid out on the schedule, Job B's operations are placed on the chart as far to the left as possible, without violating the sequence constraints. In this case, the chart shows that both jobs (eight hours of work) can be completed in five hours. This is made possible by running both jobs at the same time (on separate machines) during the second, third, and fourth hours. The second schedule assumes that Job B is run first. This schedule requires a total of six hours, one more than the previous schedule. If the total elapsed time for completion of the two jobs is an important criterion, the first schedule would be superior to the second.

Although this problem is easily solved, solutions to actual job shop sequencing problems require the use of sophisticated models and the calculating power of computers. It is not unusual for job shops to have 5,000 customer orders in process at any given time, with each order requiring 50 or 60 distinct processing or machine operations. The number of combinations of feasible sequences is astronomical in such problems, and they provide many problems in modeling and systems development for operations researchers and industrial engineers. (W.K.H.)

Manufacturing progress function. Because of the enor-

mous complexity of a typical mass production line and the almost infinite number of changes that can be made and alternatives that can be pursued, a body of quantitative theory of mass production manufacturing systems has not yet been developed. The volume of available observational data is, however, growing, and qualitative facts are emerging that may eventually serve as a basis for quantitative theory. An example is the "manufacturing progress function." This was first recognized in the airframe industry. Early manufacturers of aircraft observed that as they produced increasing numbers of a given model of airplane, their manufacturing costs decreased in a predictable fashion, declining steeply at first, then continuing to decline at a lower rate, as shown in Figure 5. When an

Manufacturing progress function

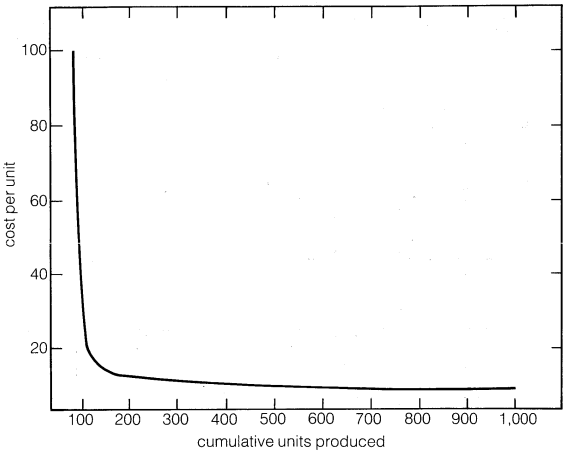


Figure 5: Manufacturing progress function relating cost per unit produced to number of units on a regular scale.

actual cost graph is drawn on double logarithmic paper, as shown in Figure 6, plotting the logarithm of the cost per unit as a function of the logarithm of the total number of units produced results in data points that almost form a straight line. Over the years similar relationships have been found for many products manufactured by mass production techniques. The slope of the straight line varies from product to product. For a given class of products and a given type of production technology, however, the slope appears remarkably constant.

Manufacturing progress functions can be of great value to the manufacturer. It is a consequence of the mathematical relationship portrayed in Figure 6 that the cost of a unit decreases by a constant percentage every time the cumulative volume of production is doubled. The actual percentage depends on the slope of the line. (Figure 6 depicts a 25 percent progress function; that is, the cost per unit decreases by 25 percent when the cumulative volume of production has doubled.) Thus the progress function can serve as a useful tool in estimating future costs. Furthermore, the failure of costs to follow a well-established progress function may be a sign that more attention should be given to the operation in order to bring its cost performance in line with expectation.

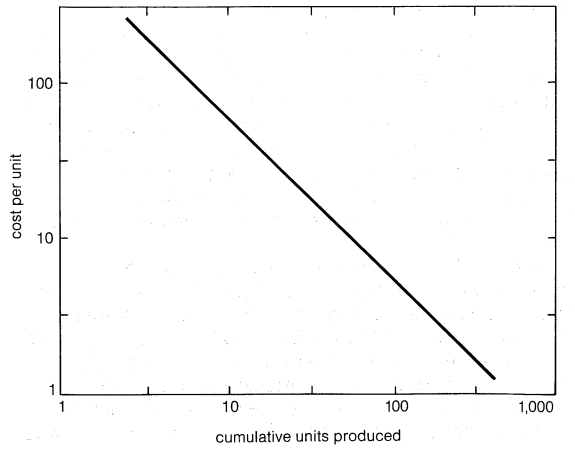


Figure 6: Manufacturing progress function relating cost per unit produced to number of units on a double logarithmic scale.

Though manufacturing progress functions are sometimes called "learning curves," they reflect much more than the improved training of the manufacturing operators. Indeed, although improved operator skill is important in the early start-up of production, the major portion of the long-term cost improvement is contributed by improvements in product design, in machinery, and in the overall engineering planning of the production sequence.

(M.T./W.K.H.)

Network routing. A network may be defined by a set of points, or "nodes," that are connected by lines, or "links." A way of going from one node (the "origin") to another (the "destination") is called a "route" or "path." Links, which may be one-way or two-way, are usually characterized by the time, cost, or distance required to traverse them. The time or cost of traveling in different directions on the same link may differ.

A network routing problem consists of finding an optimum route between two or more nodes in relation to total time, cost, or distance. Various constraints may exist, such as a prohibition on returning to a node already visited or a stipulation of passing through every node only once.

Communication and transportation problems

Network routing problems commonly arise in communication and transportation systems. Delays that occur at the nodes (*e.g.*, railroad classification yards or telephone switchboards) may be a function of the loads placed on them and their capacities. Breakdowns may occur in either links or nodes. Much studied is the "traveling salesman problem," which consists of starting a route from a designated node that goes through each node (*e.g.*, city) only once and returns to the origin in the least time, cost, or distance. This problem arises in selecting an order for processing a set of production jobs when the cost of setting up each job depends on which job has preceded it. In this case the jobs can be thought of as nodes, each of which is connected to all of the others, with setup costs as the analogue of distances between them. The order that yields the least total setup cost is therefore equivalent to a solution to the traveling salesman problem. The complexity of the calculations is such that even with the use of computers it is very costly to handle more than 20 nodes. Less costly approximating procedures are available, however. More typical routing problems involve getting from one place to another in the least time, cost, or distance. Both graphic and analytic procedures are available for finding such routes.

Competitive problems. Competitive problems deal with choice in interactive situations where the outcome of one decision maker's choice depends on the choice, either helpful or harmful, of one or more others. Examples of these are war, marketing, and bidding for contracts. Competitive problems are classifiable as certain, risky, or uncertain, depending on the state of a decision maker's knowledge of his opponent's choices. Under conditions of certainty, it is easy to maximize gain or minimize loss. Competitive problems of the risk type require the use of statistical analysis for their solution; the most difficult aspect of solving such problems usually lies in estimating the probabilities of the competitor's choices; for example, in bidding for a contract on which competitors and their bids are unknown.

Game theory

The theory of games was developed to deal with a large class of competitive situations of the uncertainty type in which each participant knows what choices he and each other participant has. There is a well-defined "end state" that terminates the interaction (*e.g.*, win, lose, or draw), and the payoffs associated with each end state are specified in advance and are known to each participant. In situations in which all the alternatives are open to competition, or some of their outcomes are not known in advance, operational gaming can sometimes be used. The military have long constructed operational games; their use by business is more recent.

Search problems. Search problems involve finding the best way to obtain information needed for a decision.

Though every problem contains a search problem in one sense, situations exist in which search itself is the essential process; for example, in auditing accounts, inspection and quality control procedures, in exploration for minerals, in the design of information systems, and in military problems involving the location of such threats as enemy ships, aircraft, mines, and missiles.

Errors in search problems

Two kinds of error are involved in search: those of observation and those of sampling. Observational errors, in turn, are of two general types: commission, seeing something that is not there; and omission, not seeing something that is there. In general, as the chance of making one of these errors is decreased, the chance of making the other is increased. Furthermore, if fixed resources are available for search, the larger the sample (and hence the smaller the sampling error), the less resources available per observation (and hence the larger the observational error).

The cost of search is composed of setup or design cost, cost of observations, cost of analyzing the data obtained, and cost of error. The objective is to minimize these costs by manipulating the sample size (amount of observation), the sample design (how the things or places to be observed are selected), and the way of analyzing the data (the inferential procedure).

Almost all branches of statistics provide useful techniques for solving search problems. In search problems that involve the location of physical objects, particularly those that move, physics and some fields of mathematics (*e.g.*, geometry and trigonometry) are also applicable.

A "reversed-search" problem arises when the search procedure is not under control but the object of the search is. Most retailers, for example, cannot control the manner in which customers search for goods in their stores, but they can control the location of the goods. This type of problem also arises in the design of libraries and information systems, and in laying land and sea mines. These, too, are search problems, and solution techniques described above are applicable to them.

FRONTIERS OF OPERATIONS RESEARCH

Operations research is a rapidly developing application of the scientific method to organizational problems. Its growth has consisted of both technical development and enlargement of the class of organized systems and the class of problems to which it is applied.

Tactics and strategy compared

Strategic problems. Tactics and strategy are relative concepts. The distinction between them depends on three considerations: (1) the longer the effect of a decision and the less reversible it is, the more strategic it is; (2) the larger the portion of a system that is affected by a decision, the more strategic it is; and (3) the more concerned a decision is with the selection of goals and objectives, as well as the means by which they are to be obtained, the more strategic it is.

Strategy and tactics are separable only in thought, not in action. Every tactical decision involves a strategic choice, no matter how implicit and unconscious it may be. Since the strategic aspects of decisions are usually suppressed, an organization's strategy often emerges as an accidental consequence of its tactical decisions.

Operations research is becoming increasingly concerned with strategic decisions and the development of explicit strategies for organizations so as to improve the quality of their tactical decisions and make even the most immediate and urgent of these contribute to its long-run goals.

The system design problem. Operations research has traditionally been concerned with finding effective solutions to specific operational problems. It has developed better methods, techniques, and tools for doing so. But operations researchers have found that too many of their solutions are not implemented and, of those that are, too few survive the inclination of organizations to return to familiar ways of doing things. Therefore, operations researchers have gradually come to realize that their task should not only include solving specific problems but also designing problem-solving and implementation systems that predict and prevent future problems, identify and solve current ones, and implement and maintain these solutions under changing conditions.

The planning problem. Operations researchers have come to realize that most problems do not arise in isolation but are part of an interacting system. The process of seeking simultaneous interrelated solutions to a set of interdependent problems is planning. More and more operations research effort is being devoted to developing a rational methodology of such planning, particularly strategic planning.

Most organizations resist changes in their operations or management. The organizational need to find better ways of doing things is often not nearly as great as is the need to maximize use of what it already knows or has. This is apparent in many underdeveloped countries that, while complaining about the lack of required resources, use what resources they have with considerably less efficiency than do most developed countries. Operations research, therefore, has been addressing itself more and more to determining how to produce the willingness to change.

Organizations and authority

Types of organization. Operations researchers have become increasingly aware of the need to distinguish between different types of organization because their distinguishing features affect how one must go about solving their problems. Two important classifications exist, the first of which is homogeneous–heterogeneous. Homogeneous organizations are those in which membership involves serving the objectives of the whole (*e.g.*, a corporation or military unit), while heterogeneous organizations are those whose principal objective it is to serve the objectives of its members (*e.g.*, a university or city). The second classification is unimodal–multimodal. Unimodal organizations are hierarchical organizations with a single decision-making authority that can resolve differences between any lower level decision makers. Multimodal organizations have no such authority but have diffused decision making and hence require agreement among the several decision makers in order to reach conclusions.

Since current skills in operations research are largely restricted to homogeneous unimodal organizations, attempts are under way to develop methodologies adequate for improving the other three types of organization.

In order to solve any of the preceding problems more effectively, operations research requires a better understanding of human behaviour, individual and collective, than is currently available. Furthermore, what understanding the behavioral sciences claim to provide is seldom available in a form that lends itself to symbolic representation and hence to operations research methodology. Operations researchers, therefore, are increasingly working with behavioral scientists to develop behavioral theories that are expressible in a more usable form.

As the scope of problems to which operations research addresses itself increases, it becomes more apparent that the number of disciplines and interdisciplines that have an important contribution to make to their solution also increases. An attempt to provide such a higher order integration of scientific activity is being made in the management sciences. (R.L.A.)

Systems engineering

Systems engineering is not so much a branch of engineering as it is a technique for applying knowledge from other branches of engineering and disciplines of science in effective combination to solve a multifaceted engineering problem. It is related to operations research (see above) but differs from it in that it is more a planning and design function, frequently involving technical innovation. Probably the most important aspect of systems engineering is its application to the development of new technological possibilities with the specific objective of putting them to use as rapidly as economic and technical considerations permit. In this sense it may be seen as the midwife of technological development.

The word "systems" is frequently used also in other combinations, especially when elements of technological advance are not so important. Systems analysis is an example. Systems theory, or sometimes systems science, is frequently applied to the analysis of physical dynamic systems. An example would be a complex electrical network

with one or more feedback loops, in which the effects of a process return to cause changes in the source of the process.

In the development of the various engineering disciplines in the 19th and 20th centuries, considerable overlap was inevitable among the different fields; for example, chemical engineering and mechanical engineering were both concerned with heat transfer and fluid flow. Further proliferation of specializations, as in the many branches of electrical and electronic engineering, such as communications theory, cybernetics, and computer theory, led to further overlapping. Systems engineering may be seen as a logical last step in the process. Systems engineers frequently have an electronics or communications background and make extensive use of computers and communications technology. Yet systems engineering is not to be confused with these other fields. Fundamentally a point of view or a method of attack, it should not be identified with any particular substantive area. In its nature and in the nature of the problems it attacks, it is interdisciplinary, a procedure for putting separate techniques and bodies of knowledge together to achieve a prescribed goal in an effective manner.

Systems engineering is a point of view

In general, a systems engineering approach is likely to differ from a conventional design approach by exhibiting increased generality in its basic logical framework and increased concern with the fundamental objectives to be achieved. Thus, at each stage the systems engineer is likely to ask both why and how, rather than merely how.

In addition to systems engineering, it is important to define systems themselves. The systems with which a systems engineer is concerned are first of all man-made. Second, they are large and complex; their component parts interact so extensively that a change in one part is likely to affect many others. Unless there is such interaction, there is little for the systems engineer to do, at least at the systems level; he can turn immediately to the components themselves. Another important characteristic of systems is that their inputs are normally stochastic; that is, the inputs are essentially random functions of time, although they may exhibit statistical regularities. Thus, one cannot expect to foresee exactly what the system will be exposed to in actual operation, and its performance must be evaluated as a statistical average of the responses to a range of possible inputs. A calculation based on a single precisely defined input function will not do.

Systems may also vary depending on the amount of human judgment that enters into their operation. There are, of course, systems such as electrical circuits, automated production equipment, or robots that may operate in a completely determinate fashion. At the other extreme, there are management and control systems, for both business and military purposes, in which machines in a sense do most of the work but with human supervision and decision making at critical points. Clearly these mixed human-machine systems offer the greatest variety both of possibilities and problems for the systems engineer. Aspects of such systems are treated below in the section *Ergonomics and human-factors engineering.*

THE DEVELOPMENT OF SYSTEMS ENGINEERING

Mathematical modeling. The systems approach stems from a number of sources. In a broad sense it can be regarded as simple extension of standard scientific methodology. It is a common procedure in science (and elsewhere) to list all of the factors that might affect a given situation and select from the complete list those that appear critical. Mathematical modeling, perhaps the most basic tool in systems engineering, is a technique encountered in any branch of science that has become sufficiently quantitative. Thus, in this broad sense, the systems approach is simply the inheritor of a tradition that is generations, if not centuries, old.

In looking for more recent and more specific sources for the systems approach, on the other hand, there are two in particular that stand out. First is the general field of communications, particularly commercial telephony, where systems engineering first appeared as an explicit discipline in its own right. Traces of the systems approach are to be found in telephone engineering at least as far back as the beginning years of the century, and systems ideas were fairly common in telephony by the 1920s and 1930s. When Bell Telephone Laboratories, the research arm of the American Telephone & Telegraph Company, was officially incorporated in 1925, its two principal engineering divisions were called respectively Apparatus Development and Systems Development. A complete formal doctrine of the role of systems engineering, however, first emerged in the years after World War II as part of an effort to redefine the policy and structure of the research and development. This doctrine set the engineering effort on a level of logical parity with the research and development efforts and made it of almost comparable actual size, at least with research. The systems engineer had a multitude of functions, with special emphasis on effective utilization of scientific and technical advances in planning new communications systems. This particular set of ideas, of course, reflected the special needs of telephony. Nevertheless, as an example and a point of departure, it had a wide effect. It seems to be one of the reasons why so esoteric a subject as systems engineering advanced as rapidly as it did. (For a detailed discussion of the research and development aspects of systems engineering, see below *Research and development.*)

The importance of systems engineering

Operations research and systems engineering. A second major source for systems engineering is operations research, which originated in a recognizable form in Britain during World War II and initially was concerned with the best employment of military equipment. Typical examples included determining the best employment of a given number of bombers, the best way of arranging convoys against submarine attack, and the best way of using interceptors against a bombing attack. Operations research was effective in such cases and has flourished ever since in both civilian and military contexts.

There exists a clear distinction between operations research and systems engineering. Because operations research is concerned with the best employment of existing equipment, technological uncertainties do not arise. Systems engineering, on the other hand, is normally concerned with the planning of new equipment and such uncertainties may be important. In practice, nevertheless, systems engineering and operations research have a good deal in common. In particular, they share many of the same analytic techniques. This results in large part from the fact that a systems engineer is likely to evaluate the effectiveness of a tentative design by the same methods an operations research specialist would use with actual hardware.

Another reason for overlap is the fact that the distinction between new and existing equipment is not quite clear-cut. Newness in equipment is a relative matter. If the new equipment is sufficiently well based on existing design techniques and seems to involve few enough technical uncertainties, the issue becomes unimportant. The question is one of degree and, to an extent, of judgment.

The growth of systems engineering

Most of the present character of systems engineering derives historically from the early 1950s. There had been some noteworthy events in the years just after World War II, including, for example, the introduction of linear programming in 1947 and the founding of various organizations for continued development of the field in the late 1940s. On the whole, however, this was a period of consolidation of earlier advances. Thus, in the communications field the principal systems were some long-distance transmission systems that had been initiated before the war and had been interrupted by war activities.

In the 1950s the pace of growth accelerated appreciably. The first general textbook on systems engineering appeared in 1957 and was followed by a number of other works that treated both industrial and military applications. These publications proved sufficient to establish systems engineering as an accepted academic discipline, and courses in it are now taught in many universities throughout the developed countries of the world. Professional societies and journals exist in France, India, Japan, West Germany, the United Kingdom, and the United States.

Communications and electronics. The development of systems engineering after 1950 stemmed, in large part,

from the impact of great advances in adjacent fields, notably communications and electronics. An automatic control system is a good example. A control system has the prime characteristic that the components interact extensively and that the system as a whole has certain properties—*e.g.*, stability—that cannot be said to adhere to any individual component. Thus control systems furnished convenient textbook examples for systems engineering.

The development of information theory as a basic starting point for communications engineering, in the years just after World War II, was also influential in shaping the evolution of systems engineering. The various subsystems in many complete systems were found to be held together by what were, in effect, communication channels. Thus ideas of information transfer from one part of the system to another proved useful in understanding the operation of the structure as a whole.

Computers and systems engineering. Systems engineering also profited from the advent of computers and the subsequent development of powerful, high-level programming languages, which affected the field in two principal ways. First, they provided new tools for analyzing complex systems by means of extensive calculations or direct simulation. In the second place, they could be used to digest large amounts of data or as actual constituents of complex systems, especially those concerned largely with information transmission. This opened up the possibility of processing information as well as simply transmitting it in such systems (see also INFORMATION PROCESSING AND INFORMATION SYSTEMS).

The impact of military weapons problems on systems engineering began soon after World War II. A landmark date was 1945, when the development of Nike Ajax, a U.S. air defense missile system, was initiated.

In 1945 available rocket propulsion seemed barely sufficient to give the missile a satisfactory tactical range. It was discovered that achievable range depended on several parameters, such as the weight and size of the warhead, fineness of the missile's aerodynamic design, degree of maneuverability provided by the control system, and shape of the trajectory and average speed along it. Thus an effective systems engineering effort was mounted in which a variety of combinations of the missile's properties were explored, with the objective of achieving the best balance between range and other tactical characteristics.

Control and feedback

Control and feedback questions were also important aspects of the overall systems problem. The whole system was in fact a gigantic feedback loop because the missile was controlled by orders sent it from a ground computer, and the computer input included information on what the tracking radar observed the missile to be doing. Thus there was a closed feedback loop from missile to computer and back to the missile again. There were also such subsidiary feedback loops as that of the autopilot controlling the attitude of the missile, and the dynamic response of the system was further affected by the need to process the radar signals to remove radar "jitter." The analysis of such elaborate dynamical systems involving interlaced feedback paths has become an important special part of the general systems area.

In the 1950s and 1960s systems engineering also grew in other directions, largely as a result of weapons systems projects associated with the Cold War. Thus the Ajax study was concerned with the dynamics of a single isolated missile. On the other hand, the defense systems that grew up in the 1950s involved the coordinated operation of a large number of missiles, guns, interceptors, and radar installations scattered over a considerable area. These were all held together by a large digital computer, which thus became the central element of the system. The SAGE (semiautomatic ground environment) system in the United States is a good example.

During the same years the systems approach also became increasingly identified with management functions. Thus the phrase "systems engineering and technical direction" came into use to describe the role of a systems engineer responsible for both the initial planning of a project and its subsequent management. So-called planning, programming, and budgeting (PPB) techniques were developed to provide similar combinations of systems engineering and financial management.

In nonmilitary fields systems engineering has developed along similar though more modest lines. Early applications were likely to stress feedback control systems in large-scale automated production facilities, such as steel-rolling mills and petroleum refineries. Later applications stressed computer-based management information and control systems somewhat like those that had earlier been developed for air defense. In more recent years the systems approach has occasionally been applied to much larger civilian enterprises, such as the planning of new cities.

SYSTEMS ENGINEERING TECHNIQUES, TOOLS, AND PROCEDURES

If a system is both large and complex in the sense in which these terms have been defined, it may be difficult to find out how it works. A large part of the content of systems engineering consists of techniques for the investigation of such relatively complex situations.

Modeling and optimization. Perhaps the most fundamental technique is the flow diagram, or flowchart, a graphical display composed of boxes representing individual components or subsystems of the complete system, plus arrows from box to box to show how the subsystems interact. Though such a representation is very useful in an initial study, it is, of course, essentially qualitative. A more effective approach in the long run is construction of a so-called mathematical model, which consists of a set of equations, or sometimes simply of tables and curves, describing the interactions within the system in quantitative terms. It is not necessary for the mathematical model to be exact, as long as it serves its purpose. It frequently consists of piecewise linear approximations to basically nonlinear situations (*i.e.*, a series of short straight lines that roughly approximate a curve). After the model has been constructed and checked, a number of mathematical techniques can be employed (including straightforward enumeration and computing) to find out what it says about the actual operation of the system. Often these calculations will have a probabilistic or statistical flavour.

When the components or subsystems interact significantly, it may be possible to achieve essentially the same final level of performance in many different ways. Limited performance by one subsystem may be offset by superior performance somewhere else. These optimization studies, called trade-offs, are important in suggesting how to achieve a given result in the most economical manner. They are equally valuable in suggesting whether or not the proposed result is in fact a reasonable goal to aim for. It may be found, for example, that a modest reduction in performance will permit radical savings in overall cost or, conversely, that the postulated equipment is capable of much better performance than is asked of it, at only nominally greater expense. (It may also turn out that the equipment can supply useful functions not originally contemplated. Computing systems, for example, can usually perform extra chores of record keeping at little increased cost.) For all of these reasons, studies of such variables are an important part of systems engineering, both in the early exploratory phases of a project and in the final design.

Optimization studies

Identifying objectives. The formulation of suitable objectives for the final system is so important a part of the systems engineering process that it deserves special attention. It is, of course, always possible to state the general objectives of a system in vague or perfectionist terms. A sufficiently clear, precise, and comprehensive statement to serve as a basis for engineering studies, however, is another matter. Unless the situation has been well explored in the past, the real choices are not likely to be obvious when the work begins. Thus, the first task of the systems engineer is to develop as clear a formulation of objectives as possible. This usually involves computations and consultation with others interested in the system. Because the final statement must reflect value judgments as well as purely technical considerations, the systems engineer does not try to do this thinking alone but attempts to serve as a working focus and catalyst. Although issues of this sort naturally present themselves with particular force near the

beginning of a systems study, they may recur in subsequent steps. The question of objectives is never really out of the systems engineer's mind.

The principal reason why a satisfactory statement of objectives may present such a problem is simply that most systems have multiple objectives, often in conflict with one another. In the design of transport aircraft, for example, there are a multitude of desirable characteristics, such as range, speed, payload, and safety, to be maximized, as well as undesirable characteristics, such as noise generation and air pollution, to be minimized. Because the same design cannot do the best job in all of these directions, a compromise achieving the most desirable overall performance is required. The most attractive compromise, which may require both study and ingenuity, is not likely to be found at all until some hard thinking has been done about what characteristics are really needed.

Especially difficult problems in defining objectives may arise when an existing technology is transplanted to some new disciplinary area. An example is the application of electronics such as computer techniques to medicine and education. It seldom happens in such cases that the best system is based on a simple one-for-one substitution, such as direct replacement of a classroom teacher by electronic hardware and computer-assisted instruction materials. It is much more likely that the most effective plan will turn out to be a rather complicated mixture of the old and the new. This conclusion, however, is likely to raise basic issues about the actual objectives of the new system, issues made no simpler by the interdisciplinary nature of the situation.

A design example. The design of the commercial transport plane mentioned above is an example of a systems engineering problem. In such a design the aerodynamic lift, the drag of fuselage and wings, the control apparatus, the propulsion system, and such auxiliary hardware as the landing gear all interact substantially. One element cannot be disturbed without affecting the others; all elements and aspects of the total system, and the interactions among them, must be considered. Thus, if designers make the fuselage fatter and the wings smaller in an effort to carry more payload at the same or higher speeds, a new control system might be needed because of the changes produced in the overall mechanical and aerodynamic characteristics of the vehicle. Stronger and heavier landing gear might be needed to withstand higher landing speeds. Almost surely, the new design would call for larger engines and fuel tanks to compensate for greater aerodynamic drag. Thus the designers would have lost ground in some respects and gained in others. The new plane might be more useful for short flights when not much fuel must be carried but less useful for long ones. Obviously, the system objective—the kind of airplane actually wanted—must control the direction of any such study.

The interaction of design elements

The study becomes more interesting if a possible advance in basic technology is considered, such as an improvement in propulsion or aerodynamics, and it is desired to determine how it might best be applied in a new airplane design. The central systems engineering question then would probably encompass the relation between the available new plane characteristics and the needs of the existing air transportation system. Clearly, such an investigation can be made only by going to one of the upper levels in the systems hierarchy.

Finally, to operate the new airplane successfully, a whole series of supporting functions may be required, including routine checkout, maintenance, and spare parts supply, in addition to functions directly involved in the plane's flight. Though, under normal circumstances, these might readily be handled by the existing operating staff, it is part of the user orientation of the systems approach that the systems engineer is expected to anticipate any new requirements and make sure they are properly planned for.

To make adequate comparisons between competing objectives, a logical frame of reference, broad enough to include both, is needed. Thus, the systems engineer may study many situations in the framework of more than one system or a whole hierarchy of systems of steadily increasing generality. In the example of an airplane, the airplane itself is a possible system, as are the group of planes owned by one airline, the total number of airplanes in a particular country, and that nation's transportation facilities. Though the simplest system—the airplane itself—is a satisfactory reference for specific design problems, a more general framework may be needed to approach broader problems. Thus the individual airplane designer may seek to ameliorate air-traffic congestion by improving airplane takeoff and landing characteristics, permitting better utilization of existing airports. The airlines in turn may suggest construction of more and better airports. From the point of view of the transportation system as a whole, the best step might be to invest more money in high-speed rail facilities to carry part of the air-traffic load. In systems engineering the error of studying the problem within too narrow a framework is called the error of suboptimization.

User orientation. The stress on systems objectives has one further consequence worth mentioning; *i.e.,* that systems engineering is likely to be strongly user-oriented. This results naturally enough from the fact that systems objectives usually relate to overall performance, which is what the final user is interested in. The identity of technical interest between the systems engineer and the final user is usually marked; systems engineering is likely to give special consideration to such qualities as reliability, ease of maintenance, and convenience in operation. Moreover, the final step of a systems engineering project is typically an evaluation that attempts to find out how well the system works in the hands of the user.

Tools. The most obvious aspect of systems engineering tools is their great diversity. A leading text in the field states that "there is virtually no scientific discipline which may not be used in the design of some large-scale system" and singles out probability, mathematical statistics, computing, system logic, queuing theory, game theory, linear programming, cybernetics, group dynamics, simulation, information theory, servomechanism theory, and human engineering. To this list might also be added decision theory, nonlinear programming, some elements of econometrics, and communications theory as related to random processes.

In spite of this diversity, many of the tools of systems theory can be grouped under a few major headings. The analytic problems associated with optimization, for example, are a recurrent theme. Probability and statistics are also major areas that carry with them numerous more specialized topics, such as queuing theory and much of communications theory. Finally, computing is a major field for the systems engineer. If all else fails, direct calculation or simulation may produce the desired results.

These fields are all essentially mathematical in nature. The systems engineer may need knowledge and skills of other sorts as well. The mathematical model and associated objective functions that conventionally begin a systems analysis, for example, are only satisfactory to the extent that they adequately represent the real physical situation. The adequacy of a mathematical model in this sense, however, is a matter of physical or engineering rather than mathematical judgment. In some circumstances the systems engineer may also need to know something about experimental procedures in general and, in particular, about ways of maximizing the amount of information from a given testing program. This is particularly likely to happen in urgent high-risk projects, like space exploration or nuclear power generation, in which intermediate testing failures are bound to occur, and the systems engineer, as part of his overall responsibility, must decide what to do next. Even in simpler cases, in which the project should close in a final test and evaluation phase, the systems engineer is responsible for ensuring that this work is adequately carried out. A closely related question is the monitoring of testing procedures for routine quality control purposes. This also is logically part of the systems engineer's duties, which reflect a basic user orientation. When reliability is very important, as in space programs, it may be a major responsibility.

The systems engineer in testing programs

When the systems engineer's job is defined as including significant management responsibility, some acquaintance with modern management techniques is an obvious re-

quirement. The techniques of particular importance are those that bear most directly on cost figuring and scheduling technological developments.

Finally, systems engineering is used in new situations that may involve the application of new discoveries in science or technology to existing technical areas or the application of known science or technology in new contexts. In either case the systems engineer obviously needs considerable substantive knowledge of the fields involved in order to make reliable plans.

It is apparent that no single person is likely to meet all of these specialized qualifications. Thus, systems engineering on any significant scale almost invariably involves a team approach.

APPLICATIONS OF SYSTEMS ENGINEERING

Many useful systems are, in effect, modifications of previous designs. The proportions of the subsystems may be changed, but no substantial function has been added or left out. Chemical-processing plants and information systems, for example, are likely to be of this sort. The basic task of the systems engineer in such a situation is relatively straightforward; it is essentially a matter of reoptimizing the existing design to meet the new conditions.

In other circumstances, however, the basic systems concept represents a more radical break with the past. The new concept may involve the introduction of new functions or the realization of old functions in new ways. On the other hand, it may merely involve a radical change in system parameters (constants).

The development of radically new systems. Radically new systems concepts are like inventions in ordinary engineering. Usually offering a substantial advance in overall performance, more than would be expected from a modest reproportioning of a known system, these clearly deserve special attention. On the other hand, in many cases it is impossible to predict accurately in advance of the development just what performance may be achievable in one or more of the critical elements of the new system. This leaves the systems engineer with a special problem in planning, which is usually addressed by establishing a minimum acceptable level of performance for the critical elements, with the rest of the system so arranged that whatever is realized beyond this level will appear as growth potential in the overall capabilities of the system. Thus definitive optimization studies may be postponed until the system is better understood.

Minimum acceptable level of performance

The Nike Ajax missile system provides an example of the application of a radically new systems concept. The simple realization that the technology was available to provide a missile that could outmaneuver an enemy bomber taking evasive action was perhaps the systems invention in this case. (Guided missiles had been thought of before, but only for use against targets simpler than a rapidly maneuvering airplane.) In a more limited sense, however, the key idea in the overall systems concept was probably the decision to use a command-guidance system, as opposed, for example, to a homing system.

In the command system, both the radars used for tracking the aircraft and the intercepting missile and the computer that calculates how the missile should change course are on the ground. Such a system requires a minimum of control apparatus in the missile. It also allows the missile to follow computer-determined paths that are aerodynamically favourable. This was an especially important consideration at the time in maximizing the range achievable with available propulsion systems. It also allows, through the computer, maximum flexibility in dealing with evasive action by the target. On the other hand, adequate accuracy from the ground-tracking system becomes increasingly difficult in a command system as the range is increased, whereas a homing system is not so limited. Thus the adoption of the command system implied a belief that the ground radars would be accurate enough to provide satisfactory interception even at the limits of the expected field of fire. As the development turned out, tracking accuracies were more than adequate for the purpose, and the surplus provided growth potential toward still longer ranges and higher probabilities of interception. (In other circumstances, of course, a different choice might have been better.)

Long-term systems development. Thus far the description of systems engineering may seem to suggest that systems engineering efforts are essentially episodic. In many situations, however, there are important elements of continuity in what the systems engineer actually does. He is likely, in fact, to work on a series of similar problems as part of a long-term effort. In the case of telephone engineering, systems engineering groups have been set up formally as permanent parts of the overall organizational structure, each group having cognizance over some wide area of telephone technology. Thus continuity in this case may extend over many years.

Permanent systems engineering groups

Telephone engineering makes a particularly good example of this because of the importance of long-range forecasting for the telephone plant. One of the important responsibilities of the systems engineer in this field, for example, involves establishment of performance standards for new items in the plant. In most fields the systems engineer's role in this respect is comparatively modest because performance standards represent value judgments established by others. In telephony the systems aspects are more conspicuous because, at any given instant, the plant is a composite of new and old items, designed at different times in different ways to meet a variety of assumed operating conditions. The pieces must all work together. Any new apparatus or system must be compatible with what already exists. In addition, however, if an orderly evolution of the plant is to continue, the new apparatus must also be suited to the traffic and service needs of the future. Thus the systems engineer's responsibility, to provide a sort of long-term doctrine of performance standards, becomes very important.

Somewhat the same considerations affect the economics of new telephone systems. Here again a carefully developed forecast is essential. On the one hand, there are usually substantial economies of scale in communication apparatus. Thus it pays to take big steps in installing new equipment. On the other hand, capital charges on underutilized equipment are likely to be excessive. Because the successive steps of design, production, and installation of a complicated system may take years, and because meanwhile both technology and service demands may change materially, the situation may increase in complexity, and the systems engineers may find that the choice of the most inviting system involves a calculated risk. They are responsible for judging the degree to which new possibilities should be exploited at any time. The principal result of assigning the telephone systems engineers a continuing organizational position and function is to put them into much more intimate contact with the technical frontier in communications. Their prime responsibility becomes that of monitoring the technical frontier to see what can be put to use in new operating systems.

Major technical advances may, of course, require many years between the original discovery or conception and the time when a practical design becomes feasible. The systems engineers keep in touch with the design and research work as it progresses throughout this interim period. They can exercise a valuable if indirect influence on the investigation simply by noticing weaknesses and errors that need correction for a project to be successful. It frequently happens that an extensive program of systematic measurements is called for before a new systems conception can be implemented, even when the basic conception is well established.

In many cases, technology suggests two or more competitive approaches to the same problem. If a continuing systems engineering organization exists, there is generally no need to make a premature choice between them. Rather, both lines can be followed until it is clear that one is superior or, perhaps, that each has its particular niche in the marketplace. In telephony, for example, this was the case in the long-continued rivalry between microwave transmission systems and those based on transmissions by a coaxial cable.

Competitive approaches to problems

Applications to government and social problems. Thus far, systems engineering has been dealt with in relation to

two principal fields of application. One field is industry, in which the prospects of a further expansion of systems engineering appear bright. Existing applications furnish many good models, and it seems likely that such extensions can take place without raising many unusual problems. The other major field of application has been in military and space systems, and this may have been the principal force in shaping the systems engineering field. The possibility of new applications of systems ideas in nonmilitary areas of government has appeared in the realm of worldwide basic social and economic problems. On the other hand, systems engineering as practiced in other contexts does not automatically transfer easily to this new environment. General interest in the subject dates, however, only from the late 1960s, and the field is incompletely explored.

Systems approach in government

In one experiment in the conversion of military systems engineering techniques, a U.S. state government contracted with four large aerospace companies (each of which had a substantial capability in systems engineering) to study the following four topics: (1) a statewide information-handling system, including a plan for implementation, (2) a program for the prevention and control of crime and delinquency, (3) a waste-management problem, and (4) a systems approach to basic transportation problems.

None of the four studies led to proposals that seemed attractive enough to be implemented by the state, and, in this respect, the experiment was a disappointment, although in view of the wide scope of the problems attacked and the limited effort called for by the study contracts, the result is not surprising. On the other hand, the experiment was useful in advertising the possibilities of systems analysis as applied to civil problems and in illuminating difficulties that may be encountered in making such applications. The experiment stimulated interest in the civil uses of systems methods both inside and outside the United States.

The potential applications of the systems approach to governmental activities are so numerous and so varied, in both the developed and developing worlds, that an exhaustive catalog would be out of the question. Nevertheless, it may be worthwhile to list a few of the most conspicuous possibilities. The most obvious class is made up of massive engineering attacks on very broad socioeconomic problems. These are the situations that seem to have most in common with the applications of systems methods in developing weapons. They include new transportation systems, comprehensive attacks on pollution, and radical reconstruction of urban areas. A concrete issue is the problem of power-plant location, an urgent question in many advanced and developing countries. The systems overtones are obvious. Generating stations are customarily interconnected so that a new plant has an impact on the availability of power over a considerable region, and, of course, the effects of thermal and atmospheric pollution from a given plant may also be widespread.

Other applications of systems analysis in the social sphere tend generally to be smaller and more easily treated. One class consists of the extension of military budgeting and methods of financial control to the nonmilitary world. Another application has been the use of systems analysis to support the technical aspect of foreign-aid programs. Other fields include the possible application of specific items of new technology in such areas as crime detection, fire fighting, and traffic control. Still other studies involve specific aspects of such subjects as housing and other types of building construction. Such studies attempt to be useful rather than broad or necessarily definitive for all time to come.

The applications of systems analysis in civil government obviously still have far to go before their potentialities are exhausted. On the other hand, there are many reasons why these potentialities can be realized only slowly, if at all. Some of them are related to the inherent difficulty of the problems presented—the wide range of both technical and social considerations that may enter certain decisions, for example. Others reflect some of the common characteristics of governmental structure, the necessary bureaucratization of functions, for example, or the frequent problem of overlapping jurisdiction. Still other problems reflect the fact that existing systems analysts are trained preponderantly in the physical sciences and engineering and thus may not be well matched to the socioeconomic issues they are likely to confront, though most systems analysis groups working in socioeconomic questions try to balance their strength by adding appropriate missing skills. The most common problem, however, is probably simply the need to build up an adequate basis for mutual cooperation between systems analysts and government.

As such an evolution proceeds, there may be an increasing tendency for the individual systems analyst to become identified with the substantive area in which he works and to lose his special relation to systems analysis as a distinctive field. Thus, it may be the ultimate fate of systems analysis to disappear as a separate field and instead become an important constituent of the planning function required in many parts of modern society.

(H.W.B./W.K.H.)

Ergonomics and human-factors engineering

The term human-factors engineering is used to designate equally a body of knowledge, a process, and a profession. As a body of knowledge, human-factors engineering is a collection of data and principles about human characteristics, capabilities, and limitations in relation to machines, jobs, and environments. As a process, it refers to the design of machines, machine systems, work methods, and environments to take into account the safety, comfort, and productiveness of human users and operators. As a profession, human-factors engineering includes a range of scientists and engineers from several disciplines that are concerned with individuals and small groups at work.

The terms human-factors engineering and human engineering are used interchangeably on the North American continent. In the Soviet Union the term engineering psychology is preferred; in Europe, Japan, and most of the rest of the world the prevalent term is ergonomics, a word made up of the Greek words, *ergon,* meaning "work," and *nomos,* meaning "law." Despite minor differences in emphasis, the terms human-factors engineering and ergonomics may be considered synonymous. Human factors and human engineering were used in the 1920s and 1930s to refer to problems of human relations in industry, an older connotation that has gradually dropped out of use. Some small specialized groups prefer such labels as bioastronautics, biodynamics, bioengineering, and manned-systems technology; these represent special emphases whose differences are much smaller than the similarities in their aims and goals.

The term ergonomics

The data and principles of human-factors engineering are concerned with human performance, behaviour, and training in man-machine systems; the design and development of man-machine systems; and systems-related biological or medical research. Because of its broad scope, human-factors engineering draws upon parts of such social or physiological sciences as anatomy, anthropometry, applied physiology, environmental medicine, psychology, sociology, and toxicology, as well as parts of engineering, industrial design, and operations research.

THE HUMAN-FACTORS APPROACH TO DESIGN

Two general premises characterize the approach of the human-factors engineer in practical design work. The first is that the engineer must solve the problems of integrating humans into machine systems by rigorous scientific methods and not rely on logic, intuition, or common sense. In the past the typical engineer tended either to ignore the complex and unpredictable nature of human behaviour or to deal with it summarily with educated guesses. Human-factors engineers have tried to show that with appropriate techniques it is possible to identify man-machine mismatches and that it is usually possible to find workable solutions to these mismatches through the use of methods developed in the behavioral sciences.

The second important premise of the human-factors approach is that, typically, design decisions cannot be made without a great deal of trial and error. There are only a few thousand human-factors engineers out of the thousands of thousands of engineers in the world who are design-

ing novel machines, machine systems, and environments much faster than behavioral scientists can accumulate data on how humans will respond to them. More problems, therefore, are created than there are ready answers for them, and the human-factors specialist is almost invariably forced to resort to trying things out with various degrees of rigour to find solutions. Thus, while human-factors engineering aims at substituting scientific method for guesswork, its specific techniques are usually empirical rather than theoretical.

The man-machine model. Human-factors engineers regard humans as an element in systems, and a man-machine model is the usual way of representing that relationship. The simplest model of a man-machine unit consists of an individual operator working with a single machine. In any machine system, the human operator first has to sense what is referred to as a machine display, a signal that tells him something about the condition or the functioning of the machine. A display may be the position of a pointer on a dial, a light flashing on a control panel, the readout of a digital computer, the sound of a warning buzzer, or a spoken command issuing from a loudspeaker.

Having sensed the display, the operator interprets it, perhaps performs some computation, and reaches a decision. In so doing, the worker may use a number of human abilities, including the ability to remember and to compare current perceptions with past experiences, to coordinate those perceptions with strategies formed in the past, and to extrapolate from perceptions and past experiences to solve novel problems. Psychologists commonly refer to these activities as higher mental functions; human-factors engineers generally refer to them as information processing.

Having reached a decision, the human operator normally takes some action. This action is usually exercised on some kind of a control—a pushbutton, lever, crank, pedal, switch, or handle. The action upon one or more of these controls exerts an influence on the machine and on its output, which in turn changes the display, so that the cycle is continuously repeated.

A man-machine system does not exist in isolation; it exists in an environment of some sort. Since the nature of this environment influences the operator's efficiency and performance, the human-factors engineer must be concerned with such environmental factors as temperature, humidity, noise, illumination, acceleration, vibration, and noxious gases and contaminants.

A man-machine example. Driving an automobile is a familiar example of a simple man-machine system. In driving, the operator receives inputs from outside the vehicle (sounds and visual cues from traffic, obstructions, and signals) and from displays inside the vehicle (such as the speedometer, fuel indicator, and temperature gauge). The driver continually evaluates this information, decides on courses of action, and translates those decisions into actions upon the vehicle's controls—principally the accelerator, steering wheel, and brake. Finally, the driver is influenced by such environmental factors as noise, fumes, and temperature.

The simple man-machine model provides a convenient way for organizing some of the major concerns of human engineering: the selection and design of machine displays and controls; the layout and design of workplaces; design for maintainability; and the work environment.

Human factors in large systems. No matter how important it may be to match an individual operator to a machine, some of the most challenging and complex human problems arise in the design of large man-machine systems and in the integration of human operators into these systems. Examples of such large systems are a modern jet airliner, an automated post office, an industrial plant, a nuclear submarine, and a space vehicle launch and recovery system. In the design of such systems, human-factors engineers study, in addition to all the considerations previously mentioned, three factors: personnel, training, and operating procedures.

Personnel, training, and operating procedures

Systems are generally designed for particular kinds of operators. A space vehicle, for example, is designed for a highly select handful of astronauts. Automobiles, on the other hand, are designed to accommodate a wide spectrum of people. In large systems, the specification of personnel requirements is an integral part of systems design.

Personnel are trained; that is, they are given appropriate information and skills required to operate and maintain the system. System design includes the development of training techniques and programs and often extends to the design of training devices and training aids.

Instructions, operating procedures, and rules set forth the duties of each operator in a system and specify how the system is to function. Tailoring operating rules to the requirements of the system and the people in it contributes greatly to safe, orderly, and efficient operations.

APPLICATIONS OF HUMAN-FACTORS ENGINEERING

The basis of human-factors engineering—the consideration of information about human users in the design of tools, machines, jobs, and work environments—has always been present. One of the oldest and most efficient of human implements, the scythe, shows a remarkable degree of human-factors engineering, undoubtedly reflecting modifications made over many centuries: the adroitly curved handle and blade and the peg grasp for the left hand. All of this is in sharp contrast with the conventional snow shovel, a modern implement of generally poor design that has been blamed for many a wintertime back strain.

The need for a more formal approach to these human problems was created when machines became vastly more complex than they had ever been. High-speed jet aircraft, computers, radar, nuclear submarines, communication satellites, space vehicles—all these are products of the past few decades. The fantastic growth in the number and complexity of machines has created entirely new problems about the use of human operators and the way they can be integrated into systems. Moreover, the solution to these new problems cannot be found in the collective wisdom of society. For example, not long ago no one had any way of predicting with any certainty how astronauts could or would function in a weightless environment. Human-factors engineering is, therefore, a child of the times, born of a mechanized civilization.

Range of applications

Applications of human-factors engineering have been made to such simple devices as highway signs, telephone sets, hand tools, stoves, and to a host of modern, sophisticated complexes such as data processing systems, automated factories and warehouses, robots, and space vehicles.

The experience gained in devising these systems has contributed to the realization that even relatively simple devices raise unexpectedly important questions on human use—questions that conventional engineering practice frequently cannot answer.

Push-button telephone. The modern push-button telephone handset provides a good example of a relatively simple device that has required a great deal of human-factors engineering. The layout of the keys in the four rows of three buttons, for example, was selected only after extensive tests on a variety of arrangements: circular, two vertical rows of five buttons, two horizontal rows of five, and a diagonal pattern; the arrangement of the numerals and letters on the keys, in the order of left to right and from top to bottom, was chosen as superior to other arrangements such as that used on many desk calculators, in which the numbers increase from bottom to top. The top-to-bottom design decision was not simply a matter of logic; tests showed that people actually made fewer errors and took less time with that arrangement than they did with the calculator arrangement. Other human-factor considerations in the design of the push-button keyset were the size and style of numerals and letters for maximum legibility, the optimum sizes and spacing of the keys, and the proper force-displacement characteristics of the keys to provide tactile feedback or "feel" when the buttons are depressed.

Similar factors were considered in designing the shape of the handset itself. The locations, separations, and angles between the earpiece and mouthpiece were determined so that the assembly would fit comfortably around the greatest number of different human faces; and the weight of the handset was designed to be neither too light nor too

heavy. In recent years the careful, "user-friendly" design of conventional telephone sets has become more apparent in contrast to some of the new arrivals in the telephone marketplace, which are generally inferior in design and quality.

Space suit. The designing of a much more complicated device, such as a space suit, presents more intricate problems. A space suit is a complete miniature world, a self-contained environment that must supply everything needed for an astronaut's life, as well as comfort. The suit must provide a pressurized interior, without which an astronaut's blood would boil in the vacuum of space. The consequent pressure differential between the inside and the outside of the suit is so great that when inflated the suit becomes a distended, rigid, and unyielding capsule. Special joints were designed to give the astronaut as much free movement as possible. The best engineering has not been able to provide as much flexibility of movement as is desirable; to compensate for that lack, attention has been directed toward the human-factors design of the tools and devices that an astronaut must use.

In addition to overcoming pressurization and movement problems, a space suit must provide oxygen; a system for removing excess products of respiration, carbon dioxide and water vapour; protection against extreme heat, cold, and radiation; protection for the eyes in an environment in which there is no atmosphere to absorb the sun's rays; facilities for speech communication; and facilities for the temporary storage of body wastes. This is such an imposing list of human requirements that an entire technology has been developed to deal with them and, indeed, with the provision of simulated environments and procedures for testing and evaluating space suits.

Typewriter keyboard. Not all human-factors engineering and design is commercially successful. An example is the typewriter keyboard. Several alternative layouts, which are demonstrably superior from a human-factors point of view, have been proposed, beginning as far back as the 1920s. Despite test results which show that alternative layouts are easier to learn, create less operator fatigue, and permit faster typing, the traditional layout persists and now has been carried over into the design of millions of personal computers. In this case, inertia and resistance to change have been more formidable obstacles to efficient ergonomic design than the design itself.

Social problems. The telephone, the space suit, and the typewriter keyboard are but three out of thousands of examples that might have been selected to show how human-factors engineering has been consciously applied to solve technological problems. The same human-factors principles and methods have also been applied to a variety of social problems, such as individualized computer-assisted instruction, nonlethal antiriot equipment for law enforcement agencies, antiterrorist architecture for public buildings, and people movers for airport and urban transportation departments. The modern concern with man's relationship to the total environment implies a much-broadened definition of human-factors engineering and an increasing supply of problems for ergonomic engineers in the future. (Al.C./W.K.H.)

Applications to social problems

Research and development

INTRODUCTION AND DEFINITIONS

Research and development (R and D), a phrase unheard of in the early part of the 20th century, has since become a universal watchword in industrialized nations. The concept of research is as old as science; the concept of the intimate relationship between research and subsequent development, however, was not generally recognized until the 1950s. Research and development is the beginning of most systems of industrial production. The innovations that result in new products and new processes usually have their roots in research and have followed a path from laboratory idea, through pilot or prototype production and manufacturing start-up, to full-scale production and market introduction. The foundation of any innovation is an invention. Indeed, an innovation might be defined as the application of an invention to a significant market need. Inventions come from research—careful, focused, sustained inquiry, frequently trial and error. Research can be either basic or applied, a distinction that was established in the first half of the 20th century.

Basic research is defined as the work of scientists and others who pursue their investigations without conscious goals, other than the desire to unravel the secrets of nature. In modern programs of industrial research and development, basic research (sometimes called pure research) is usually not entirely "pure"; it is commonly directed toward a generalized goal, such as the investigation of a frontier of technology that promises to address the problems of a given industry. An example of this is the research being done on gene splicing or cloning in pharmaceutical company laboratories.

Applied research carries the findings of basic research to a point where they can be exploited to meet a specific need, while the development stage of research and development includes the steps necessary to bring a new or modified product or process into production. In Europe, the United States, and Japan the unified concept of research and development has been an integral part of economic planning, both by government and by private industry.

HISTORY AND IMPORTANCE

The first organized attempt to harness scientific skill to communal needs took place in the 1790s, when the young revolutionary government in France was defending itself against most of the rest of Europe. The results were remarkable. Explosive shells, the semaphore telegraph, the captive observation balloon, and the first method of making gunpowder with consistent properties all were developed during this period.

The lesson was not learned permanently, however, and another half century was to pass before industry started to call on the services of scientists to any serious extent. At first the scientists consisted of only a few gifted individuals. Robert W. Bunsen, in Germany, advised on the design of blast furnaces. William H. Perkin, in England, showed how dyes could be synthesized in the laboratory and then in the factory. William Thomson (Lord Kelvin), in Scotland, supervised the manufacture of telecommunication cables. In the United States, Leo H. Baekeland, a Belgian, produced Bakelite, the first of the plastics. There were inventors, too, such as John B. Dunlop, Samuel Morse, and Alexander Graham Bell, who owed their success more to intuition, skill, and commercial acumen than to scientific understanding.

Early research laboratories

While industry in the United States and most of western Europe was still feeding on the ideas of isolated individuals, in Germany a carefully planned effort was being mounted to exploit the opportunities that scientific advances made possible. Siemens, Krupp, Zeiss, and others were establishing laboratories and, as early as 1900, employed several hundred people on scientific research. In 1870 the Physicalische Technische Reichsanstalt (Imperial Institute of Physics and Technology) was set up to establish common standards of measurement throughout German industry. It was followed by the Kaiser Wilhelm Gesellschaft (later renamed the Max Planck Society for the Advancement of Science), which provided facilities for scientific cooperation between companies.

In the United States, the Cambria Iron Company set up a small laboratory in 1867, as did the Pennsylvania Railroad in 1875. The first case of a laboratory that spent a significant part of its parent company's revenues was that of the Edison Electric Light Company, which employed a staff of 20 in 1878. The U.S. National Bureau of Standards was established in 1901, 31 years after its German counterpart, and it was not until the years immediately preceding World War I that the major American companies started to take research seriously. It was in this period that General Electric, Du Pont, American Telephone & Telegraph, Westinghouse, Eastman Kodak, and Standard Oil set up laboratories for the first time.

Except for Germany, progress in Europe was even slower. When the National Physical Laboratory was founded in England in 1900, there was considerable public comment on the danger to Britain's economic position of German

dominance in industrial research, but there was little action. Even in France, which had an outstanding record in pure science, industrial penetration was negligible.

World War I produced a dramatic change. Attempts at rapid expansion of the arms industry in the belligerent as well as in most of the neutral countries exposed weaknesses in technology as well as in organization and brought an immediate appreciation of the need for more scientific support. The Department of Scientific and Industrial Research in the United Kingdom was founded in 1915, and the National Research Council in the United States in 1916. These bodies were given the task of stimulating and coordinating the scientific support to the war effort, and one of their most important long-term achievements was to convince industrialists, in their own countries and in others, that adequate and properly conducted research and development were essential to success.

The inter-war years

At the end of the war the larger companies in all the industrialized countries embarked on ambitious plans to establish laboratories of their own; and, in spite of the inevitable confusion in the control of activities that were novel to most of the participants, there followed a decade of remarkable technical progress. The automobile, the airplane, the radio receiver, the long-distance telephone, and many other inventions developed from temperamental toys into reliable and efficient mechanisms in this period. The widespread improvement in industrial efficiency produced by this first major injection of scientific effort went far to offset the deteriorating financial and economic situation.

The economic pressures on industry created by the Great Depression reached crisis levels by the early 1930s, and the major companies started to seek savings in their research and development expenditure. It was not until World War II that the level of effort in the United States and Britain returned to that of 1930. Over much of the European continent the Depression had the same effect, and in many countries the course of the war prevented recovery after 1939. In Germany Nazi ideology tended to be hostile to basic scientific research, and effort was concentrated on short-term work.

The picture at the end of World War II provided sharp contrasts. In large parts of Europe industry had been devastated, but the United States was immensely stronger than ever before. At the same time the brilliant achievements of the men who had produced radar, the atomic bomb, and the V-2 rocket had created a public awareness of the potential value of research that ensured it a major place in postwar plans. The only limit was set by the shortage of trained persons and the demands of academic and other forms of work.

Since 1945 the number of trained engineers and scientists in most industrial countries has increased each year. The U.S. effort has stressed aircraft, defense, space, electronics, and computers. Indirectly, U.S. industry in general has benefited from this work, a situation that compensates in part for the fact that in specifically nonmilitary areas the number of persons employed in the United States is lower in relation to population than in a number of other countries.

Outside the air, space, and defense fields the amount of effort in different industries follows much the same pattern in different countries, a fact made necessary by the demands of international competition. (An exception is the Soviet Union, which devotes less R and D resources to nonmilitary programs than most other industrialized nations.) An important point is that countries like Japan, which have no significant aircraft or military space industries, have substantially more manpower available for use in the other sectors. The preeminence of Japan in consumer electronics, cameras, and motorcycles, and its strong position in the world automobile market attest to the success of its efforts in product innovation and development.

TYPES OF LABORATORIES

Company laboratories. Company laboratories fall into three clear categories: research laboratories, development laboratories, and test laboratories.

Research laboratories carry out both basic and applied research work. They usually support a company as a whole, rather than any one division or department. They may be located at a considerable distance from any other part of the company and report to the highest levels of corporate management or even to the board of directors. AT&T Bell Laboratories, the research arm of American Telephone & Telegraph Company (AT&T), is an outstanding example. There the transistor and coaxial cable were developed, pioneer work in satellite communications was carried out, and many computer innovations have been developed.

Development laboratories are specifically committed to the support of particular processes or product lines. They are normally under the direct control of the division responsible for manufacture and marketing and are often located close to the manufacturing area. Frequently used as problem solvers by many sections of each company, development laboratories maintain close contacts with people in manufacturing, advertising, marketing, sales, and other departments with responsibilities for products or processes.

Test laboratories may serve a whole company or group of companies or only a single manufacturing establishment. They are responsible for monitoring the quality of output. This often requires chemical, physical, and metallurgical analyses of incoming materials, as well as checks at every stage of a process. These laboratories may be a part of a manufacturing organization, but many companies give them an independent status.

Government laboratories. The pattern followed by different countries varies widely. The general policy of the U.S. government has been not to set up laboratories of its own, even for military work, but to offer research and development contracts, usually on the basis of competitive bidding, to private companies. The most important reason for this has been a belief that the right place to develop equipment is very close to the place at which it will eventually be manufactured.

Types of government laboratories

There are exceptions to the rule. One is the type of laboratory represented by the National Bureau of Standards, a central authority on problems of measurement and standardization. Another is the type of laboratory supported by the U.S. Department of Agriculture, set up by the government in the belief that research in this field is necessary but that the industry had neither the finances nor the organization to maintain it. The continuing support of successive administrations has resulted in a large and authoritative body carrying out research over a wide field for the benefit of the farming community and thus, indirectly, of the whole nation.

A third type of government laboratory is represented by the U.S. Atomic Energy Commission and its successors, the Energy Research and Development Administration and the Department of Energy's Office of Energy Research. In this case the U.S. government recognized a situation of potential danger and also opportunity of such a nature that it was not practicable for it to be handled by private individuals. It therefore set up a body to deal with the situation, allocating funds directly and maintaining close control of the objectives and timing of research. A similar challenge is faced by the National Aeronautics and Space Administration. Although much of the detailed research and development work is contracted to private industry, overall control, as well as much of the most important work, is handled directly by the central organization.

A different type of policy has been followed in the United Kingdom. A chain of government laboratories supports the requirements of the armed forces and carries out a great deal of the basic and applied research from which new weapons and military techniques emerge. The government laboratories play a major part in negotiating and monitoring the contracts placed with private industry for the eventual development and production of equipment for the armed forces.

In addition to the government laboratories that focus on military R and D, the U.K. government supports civilian establishments such as the National Engineering Laboratory. These have a considerable degree of independence in selecting projects that will bring the greatest benefit to

industry as a whole, and their results are made available to all. They maintain close liaison with the research associations (see below *Research associations*) and with private industry and attempt to concentrate their work in areas that for one reason or another are not covered elsewhere.

In West Germany, as in the United Kingdom, defense research is the responsibility of a chain of government laboratories, but they are much smaller. Most of the work is done for them on contract by the research associations. They place very little research with private industry and call upon it only in the later stages of development.

In Japan there is a chain of laboratories that serves the needs of government departments. They work closely with the research associations that support particular industries. The military laboratories carry out the bulk of defense research and development themselves, and they are also responsible for the placing of contracts with private industry. These are usually confined to the later stages of development and are expected to lead almost directly to production.

The French system is similar, but the directly controlled government laboratories are even smaller and do little more than direct and coordinate work done by the research associations.

Factors common in all countries

In spite of differences in organization, the day-to-day conduct of government-sponsored research and development in all countries has much in common. In every case a comparatively small number of government employees keep in constant touch with the whole of the scientific and technical community and dispense contracts in the way they consider will make the best use of the resources available in the broad national interest. The fact that in some countries it is done in laboratories under direct governmental control, in others in those under private control, and in yet others in those in which responsibility is split is of secondary importance. In every case, government support is important. Even in the United States, with its relatively few government laboratories, government research contracts account for almost half of all R and D expenditures.

Independent laboratories. The concept of a laboratory that maintains itself solely by selling research originated with the Mellon Institute in Pittsburgh before World War I. The difficulties that have to be faced are formidable, for a great deal of research work yields no immediate or obvious reward, and it is extremely difficult to satisfy a customer that he is getting value for his money. Nevertheless, a number of such bodies, including the Battelle Memorial Institute, Columbus, Ohio, and the Stanford Research Institute (now SRI International), Menlo Park, Calif., have become large and successful. These organizations offer the services of workers of high professional standing who cover between them a wide range of disciplines. They undertake studies and investigations on any subject within their competence for fees that are negotiated with each customer; and, although they do not expect to make profits, they are required to be self-supporting.

Another type of organization is represented by Arthur D. Little, Inc., Cambridge, Mass., which is run on strictly commerical lines, seeking to make a commercially viable profit from the resources employed. Only one or two organizations of similar type have been established in western Europe, and they have not grown to a size comparable with those in America.

Both in Europe and in the United States, there are a great number of small laboratories providing specialist analytical, spectrographic, metallurgical, and similar services to industry. Most of their clients are companies that lack adequate facilities of their own and that in the course of time either learn to stand on their own feet or go out of business. But the constant appearance of new companies and the increasing need for technical understanding in established companies results in a slow but steady increase in the number of independent specialist laboratories serving them.

Research associations. A more important part of the industrial research and development effort in western Europe and in Japan is represented by research associations. Most of these organizations are concerned with a single industry. Examples are the British Glass Industry Research Association in Sheffield, the French Petroleum Institute in Paris, the Max Planck Institute for Iron Research in Düsseldorf, and the Textile Research Institute in Yokohama. These laboratories are mainly concerned with the long-term problems of the industries they serve, but they are on occasion called in to help with immediate technical difficulties beyond the powers of local staff. In European countries other than the United Kingdom, they carry out substantial work under contract to the defense departments.

University laboratories. In principle, university laboratories are completely independent and free to investigate anything that interests them. In practice, many of them are anxious to keep in touch with industry and to focus their research effort on problems with practical applications. Similarly, industrial scientists wish to maintain contact with advanced academic research. The result is a constant interchange between universities and industry; industrialists suggest problems for university research and provide funds to support it, and university staffs act as consultants and advisers to industry. In addition, government may play a direct role by funding university research in a wide variety of specialities and research areas.

THE ROLE OF GOVERNMENT

World War I brought home to every government involved the importance of having its armed forces supported by an industry using the most advanced scientific techniques. Since then it has been generally accepted that it is frequently desirable to encourage research and development for reasons of economic growth as well as national security. This has resulted in massive support from public funds for many sorts of laboratories.

Through World War II this support was limited to research and development of direct military significance, but in more recent years the types of equipment used by the armed forces have become so extensive and so complicated that it is no longer practicable to distinguish between the requirements of an efficient armament industry and those of an efficient civilian industry. Advanced communication systems, aircraft engines, computers, and nuclear-power generators have been just as important to one as to the other. This fact has led governments to become the greatest single sponsors of industrial research.

"Spin-off" of research results

During the 1960s it became clear that the "spin-off," or civilian and commercial application of work done under government contracts for defense or space research and development, was giving the industries who participated a crucial advantage over their competitors, particularly over those in countries in which comparable assistance was not available. The dominance of U.S. firms in computer development and in microelectronics was generally attributed to this cause, and the outstanding success of the British aeroengine industry could hardly have been achieved without it. There were obvious examples, such as communication satellites, which derived from work on military rocket propulsion, and more subtle ones, such as the highly reliable electronic components, developed to make communication with and control of space vehicles more reliable, that made it possible to produce television sets with far longer life between failures. The reaction of most industrial countries was to increase government support of private research. In the United Kingdom the Ministry of Technology took responsibility for allocating funds to private industry for research projects with no direct military application. The usual practice has been to contribute 50 percent of the cost of the work, the private company providing the balance.

In the United States and in most western European countries, research contracts placed by government departments originate in the decision of a scientifically or technically oriented executive of the department that certain work should be done. This leads to the preparation of a specification of the work, which is then offered to industry, to private research institutes, and to universities for competitive bidding.

The terms of contract have varied widely. It is common to offer contracts on a cost-plus basis. The contractor

keeps records of the hours worked by the staff and the materials used; these are checked by government auditors and paid for at a negotiated rate, together with a fixed percentage as profit. Criticisms of this system led to fixed-price contracts, but these have the drawback that it is often so difficult to define the end point of a research contract that the contractor can treat a fixed-price agreement as if it were cost-plus. Another problem is that, when the end point can be exactly defined but there are genuine uncertainties in the program, the most attractive bid may come from a contractor who, through ignorance, takes too light a view of the difficulties. Yet another formula that has been tried is to offer contracts on a cost-plus-fixed-profit (rather than cost-plus-percentage) basis.

Fixed-price contracts

In all these cases the main concern of the agency that sponsors the contract is to get the work done as efficiently as possible. With the many uncertainties of research and development, true economy is more likely to lie in high-quality work than in low pricing. Consequently, in every country in which the government is a substantial supporter of private research and development, the departments concerned have set up elaborate systems of monitoring work and of keeping in touch with the performance and capabilities of the companies willing to undertake it. In negotiating contracts, the sponsors attempt to place them where they will be handled most successfully. At the same time, they are concerned to keep together teams that are likely to do good work for them in the future. Within this framework the struggle of the customer to negotiate the best price for a project and that of the contractor to get a good return for the commitment of valuable resources follow normal commercial practice.

Patent rights are often a complex issue when research is carried out by private industry but paid for, at least partially, by government. In some cases the rights are the exclusive property of the government, and in others they belong to the contractor. A common compromise is for the government to retain all rights when anyone uses the patents to supply a government department but for the contractor to retain them when another party is involved. Thus, the government can place production orders with any contractor that it chooses, and the company that carried out the development is obliged to release information to him. If, however, the new contractor wishes to sell in the open market, he is obliged to negotiate a license and pay a royalty to the original development laboratories.

THE MANAGEMENT OF RESEARCH AND DEVELOPMENT ACTIVITIES

Most research and development projects are examples of a project, or one-shot, production system. Here, as opposed to the ongoing activity found in batch or continuous systems, resources are brought together for a period of time, focused on a particular task, such as the development of a new product, and then disbanded and reassigned. The management of such projects requires a special type of organization to administer project resources in an effective manner and maintain clear accountability for the progress of the project. This organization also must avoid the inherent conflict of authority between project managers and managers in the marketing, production, and other departments and coordinate members of R and D teams who are assigned to more than one project and must divide their time among conflicting demands. The management of the whole process is a key to R and D and commercial success.

In industries where continuous innovation and R and D are critical, such as electronics, drugs, robotics, and aerospace, the R and D department usually operates on a corporate level comparable to production, finance, and marketing. A relatively small management group usually sets priorities and budgets and supervises R and D activities. Most research and development personnel are assigned to project activity and report to individual project managers who have considerable autonomy and authority over the people and resources required to complete the project.

The basic purpose of the R and D laboratories of private industry is to provide new products for manufacture and new or improved processes for producing them. One difficulty facing those who plan these projects is the relationship between development costs and predicted sales. In the early stages of development, project expenditures are typically low. They increase to a maximum and decline slowly, disappearing as early production difficulties are overcome and the product settles into a market niche. Similarly, production rises slowly at first, then more rapidly, and finally reaches a plateau. After a time, production starts to fall, sales declining gradually as the product becomes obsolete or abruptly as it is replaced by a new one.

At any particular time, a company may have a number of products at different stages of the cycle. Project managers must ensure that the total development effort required is neither greater nor significantly less than available human and financial resources. Production managers must be satisfied that the eventual demands upon their capacity and resources will be sufficient to keep them fully loaded but not overloaded.

To maintain such a balanced condition, a steady flow of new R and D proposals is required. Each must be studied by technical, commercial, financial, and manufacturing experts. Planning within an R and D organization, then, consists of selecting for development new products and processes that promise to employ the resources available in the most profitable manner. R and D managers have a key part to play in proposing projects as well as in carrying them out.

Key R and D decisions

At each stage of the research and development process, there are numerous technical, financial, and managerial issues that have to be resolved and coordinated with many groups. For example, during the late 1970s and early 1980s several computer and electronics companies in the United States and Europe established major research programs aimed at developing bubble memory devices for large computers. As bubble memories were proved to be technically feasible (*i.e.,* work reliably under normal operating conditions), attention shifted to developing processes to manufacture the memory units at competitive costs. This part of the job proved the most difficult, and by the mid-1980s bubble memories had captured only a minuscule share of the total market for memory devices.

The difficulties in developing the design and production specifications needed to produce low-cost bubble memory units severely tested the mettle of the R and D organizations in several companies in the United States, Japan, and Europe. Each company had to balance the expense of continued R and D investment against the consequences of withdrawing from bubble memory research. Making a decision like this requires a keen sense of the market, a knowledge of the technical issues at hand, and, most importantly, an understanding of the company's priorities and alternatives for R and D funds.

PROJECT MANAGEMENT AND PLANNING TECHNIQUES

Value engineering and cost-benefit analysis. In the areas in which technology advances fastest, new products and new materials are required in a constant flow, but there are many industries in which the rate of change is gentle. Although ships, automobiles, telephones, and television receivers have changed over the last quarter of a century, the changes have not been spectacular. Nevertheless, a manufacturer who used methods even 10 years old could not survive in these businesses. The task of R and D laboratories working in these areas is to keep every facet of the production process under review and to maintain a steady stream of improvements. Although each in itself may be trivial, the total effect is many times as large as the margin between success and failure in a competitive situation.

These efforts to improve existing products and processes have been formalized under the titles of value engineering and cost-benefit analysis.

In value engineering every complete product and every component have their primary function described by an action verb and a noun. For example, an automobile's dynamo, or generator, generates electricity. The engineer considers all other possible methods of generation, calculates a cost for each, and compares the lowest figure with

that for the existing dynamo. If the ratio is reasonably close to unity, the dynamo can be accepted as an efficient component. If not, the engineer examines the alternatives in more detail. The same treatment is applied in turn to each of the parts out of which the chosen component is built, until it is clear that the best possible value is being obtained.

Function versus cost

Cost-benefit analysis approaches the same fundamental problem from a different angle. It takes each part of a product or process and completely defines its function and the basis for measuring its benefits or effectiveness. Then the costs of obtaining each part are reviewed, taking full account of purchased material, labour, investment cost, downtime, and other factors. This focuses attention upon the most expensive items and makes it possible to apply the principal effort in seeking economies at the points of maximum reward. In the effort to improve a product or process, care must be taken to evaluate alternatives on the same "cost" and "benefit" bases so that existing approaches do not enjoy a special advantage just because they are familiar.

These two processes are unending. Every new material, new manufacturing technique, or new way of carrying out an operation gives the engineer a chance to improve his product, and it is from these continuing improvements that the high degree of economy and reliability of modern equipment derives. (T.S.McL./W.K.H.)

PERT and CPM. Project managers frequently face the task of controlling projects that contain unknown and unpredictable factors. When the projects are not complex, bar charts can be used to plan and control project activities. These charts divide the project into discrete activities or tasks and analyze each task individually to indicate weekly manpower requirements. As the work goes forward, progress is charted and estimates are made on the effects of any delays or difficulties encountered during the completion of the project.

In the mid-1950s more sophisticated methods of project planning and control were developed. Two systems based on a network portrayal of the activities that make up the project emerged at about the same time. PERT (Program Evaluation and Review Technique) was first used in the development of submarines capable of firing Polaris missiles. CPM (the Critical Path Method) was used to manage the annual maintenance work in an oil and chemical refinery. Many variations and extensions of the two original techniques are now in use, and they have proved particularly valuable for projects requiring the coordinated work of hundreds of separate contractors. The use of project planning and control techniques based on PERT or CPM are now common in all types of civil engineering and construction work, as well as for large developmental projects such as the manufacture of aircraft, missiles, space vehicles, and large mainframe computer systems.

Figure 7 is a simple example of a network, or "arrow diagram," used in developing an electronic component for a complex system. Each circle on the diagram represents a task or well-defined activity that is part of the project. The number in each circle represents the expected time required to complete the task.

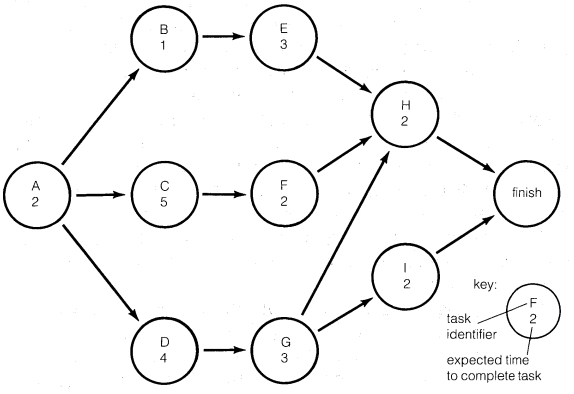


Figure 7: Network diagram for the Critical Path Method problem.

Task A requires two weeks to complete and might, for example, represent the development of general specifications for an electronic unit in question. Tasks B and E might represent two related parts of the design of the unit's power supply, C and F the design of the main functional circuits, and D and G the design of the control circuitry. Arrows indicate the precedence of relationships and depict which tasks must be completed before subsequent tasks can begin. In this example, tasks B, C, and D cannot be started until A has been completed (that is, no one can design specific component items before the general specifications are agreed upon).

Task H requires two weeks to complete but cannot be started until the designs of the power supply and the functional and control circuits have been completed. This task might represent the design of the unit's case or cover, and the case cannot be made final until all of the component designs are completed.

The arrow diagram is an invaluable planning aid for determining how long a project will take to complete. Adding all of the task times together in the example indicates that there are 24 weeks of work to be completed. Note, however, that several tasks can be done simultaneously. For example, once task A has been completed, B, C, and D can be started and worked on concurrently. Thus, the earliest completion date can be determined by looking at all possible "paths" through the network and choosing the longest one, or the one with tasks requiring the most total time. In this example the longest, or "critical," path is A–C–F–H, requiring a total time of 11 weeks.

The arrow diagram yields additional information to the project planner. The earliest possible time that task H can be started is nine weeks after the start of the project (that is, after tasks A, C, and F have been completed). When task A is completed at the end of week 2, tasks B and E do not have to be started immediately in order to complete the project in the minimum possible time; B and E each have three weeks of "slack." The diagram shows that if activity B is started three weeks later than its earliest possible start time (at week 5), it would be completed at the end of week 5; E would then start at the beginning of week 6 and be completed in time for H to begin at its earliest time, the beginning of week 10.

"Slack" in project networks

The notion of slack in a project network is a powerful concept that allows planners to schedule scarce resources efficiently and manage people and equipment so that critical activities are kept on schedule and slack activities are delayed without placing the project in jeopardy.

This simple example is based on CPM logic; it uses single-point task time estimates and assumes that the completion time for the project is the simple sum of the task times along the critical path. PERT logic assumes probabilistic estimates for each task time, with pessimistic, realistic, and optimistic estimates for the completion times of each task.

In actual projects the relationships among the required tasks are often complex, and the arrow diagram for the project might cover the entire wall of an office. Even though it is a time-consuming job to work out arrow diagrams, precedence relationships, task time estimates, and so on for large projects, CPM or PERT is an invaluable aid to planning and control. The proliferation of computer programs that handle critical path and slack time calculations and the development of computer systems capable of handling cost estimates, budget control, resource allocation, and time scheduling promise to make CPM and PERT even more valuable than in the past. (W.K.H.)

BIBLIOGRAPHY

General references. Two handbooks contain a wealth of general information on industrial production systems, methods, problems, and management techniques: GORDON B. CARSON, HAROLD A. BOLZ, and HEWITT H. YOUNG (eds.), *Production Handbook,* 3rd ed. (1972); and H.B. MAYNARD (ed.), *Industrial Engineering Handbook,* 3rd ed. (1971). See also FRANKLIN G. MOORE and THOMAS E. HENDRICK, *Production/Operations Management,* 8th ed. (1980), a classic textbook covering a wide range of topics in nontechnical language; and HARWOOD F. MERRILL (ed.), *Classics in Management,* rev. ed. (1970), an excellent collection of excerpts from the writings of several pioneers in industrial production, including Frederick W.

Taylor, Henri Fayol, and Frank B. and Lillian M. Gilbreth. Two general texts that cover many aspects of the general field treated in this article are ELWOOD S. BUFFA, *Modern Production/Operations Management,* 7th ed. (1983); and RICHARD B. CHASE and NICHOLAS ACQUILANO, *Production and Operations Management: A Life Cycle Approach,* 4th ed. (1985).

Production systems. Various aspects of systems and control are dealt with in B.H. AMSTEAD, PHILLIP F. OSTWALD, and MYRON L. BEGEMAN, *Manufacturing Processes: SI Version,* 7th ed. (1979); MICHAEL PETERS and TERENCE OLIVA, *Operations and Production Management* (1981); and JAMES H. GREENE, *Production and Inventory Control,* rev. ed. (1974), and *Operations Management: Productivity and Profit* (1984).

Early studies of the organization of human effort for production are treated in classics of the 18th and 19th centuries, including ACADEMIE DES SCIENCES. PARIS, *Descriptions des arts et métiers,* 45 vol. (1761–89); ADAM SMITH, *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776, reissued 1981); and CHARLES BABBAGE, *On the Economy of Machinery and Manufactures,* 4th ed. enlarged (1835, reprinted 1971).

Historical views of developments leading to modern mass production methods are J.K. FINCH, *Engineering and Western Civilization* (1951), *The Story of Engineering* (1960); and FRIEDRICH KLEMM, *A History of Western Technology* (1959, reissued 1964; originally published in German, 1954). Technical descriptions of mass production techniques are given by E. PAUL DeGARMO, J. TEMPLE BLACK, and RONALD A. KOHSER, *Materials and Processes in Manufacturing,* 6th ed. (1984). The classical technical works on time and motion studies in manufacturing are FREDERICK WINSLOW TAYLOR, *The Principles of Scientific Management* (1911, reissued 1967); and FRANK B. GILBRETH, *Motion Study: A Method for Increasing the Efficiency of the Workman* (1911, reprinted 1972). RALPH M. BARNES, *Motion and Time Study,* 7th ed. (1980), describes modern industrial engineering methods; ERNEST J. MCCORMICK, *Human Factors in Engineering and Design,* 5th ed. (1982), provides a broad study of the physiological aspects of engineering design. TREVOR I. WILLIAMS, *A Short History of Twentieth-Century Technology c. 1900–c. 1950* (1982), is a good overview; OTTO MAYR and ROBERT C. POST (eds.), *Yankee Enterprise: The Rise of the American System of Manufactures: A Symposium* (1981), is a treatment of mass production revolution; DANIEL NELSON, *Frederick W. Taylor and the Rise of Scientific Management* (1980), is a study of the development of Taylor's ideas; IRA C. MAGAZINER and ROBERT B. REICH, *Minding America's Business: The Decline and Rise of the American Economy* (1982, reissued 1983), is an account of specific problems.

Books written about human and societal problems and adjustments to the industrial milieu include R. BURLINGAME, *Backgrounds of Power: The Human Story of Mass Production* (1949), a popular history and commentary; WILLIAM A. FAUNCE, *Problems of an Industrial Society,* 2nd ed. (1981), on the sociological effects; and HARVEY SWADOS, *On the Line* (1957, reissued 1978), about the problems of assembly line work. Others have focused on the problems of individuals and how they may be approached. Among these are WILLIAM J. DICKSON and F.J. ROETHLISBERGER, *Counseling in an Organization* (1966); ROBERT N. FORD, *Motivation Through the Work Itself* (1969); FREDERICK HERZBERG, *Work and the Nature of Man* (1966, reprinted 1973); and CHARLES R. WALKER and ROBERT H. GUEST, *The Man on the Assembly Line* (1952, reprinted 1979).

Production management. General references with broad coverage include EVERETT E. ADAM, JR., and RONALD J. EVERT, *Production and Operations Management: Concepts, Models, and Behavior,* 2nd ed. (1982); and ROY D. HARRIS and RICHARD F. GONZALEZ, *The Operations Manager: Role, Problems, Techniques* (1981). An analytical approach to production control is SAMUEL EILON, *Elements of Production Planning and Control* (1962). The use of operational research models is covered in EDWARD H. BOWMAN and ROBERT B. FETTER, *Analysis for Production and Operations Management,* 3rd ed. (1967); ELWOOD S. BUFFA and JEFFREY G. MILLER, *Production-Inventory Systems: Planning and Control,* 3rd ed. (1979); and MARTIN K. STARR, *Production Management: Systems and Synthesis,* 2nd ed. (1972). More specialized works are MARTIN K. STARR and DAVID W. MILLER, *Inventory Control: Theory and Practice* (1962); JOHN F. MUTH and GERALD L. THOMPSON (eds.), *Industrial Scheduling* (1963). More general references dealing with strategic issues include WILLIAM J. ABERNATHY, KIM B. CLARK, and ALAN M. KANTROW, *Industrial Renaissance: Producing a Competitive Future for America* (1983); JAMES O'TOOLE, *Making America Work: Productivity and Responsibility* (1981); and a collection of articles from the *Harvard Business Review, Survival Strategies for American Industry,* ed. by ALAN M. KANTROW (1983).

Operations research. *History:* J.G. CROWTHER and R. WHIDDINGTON, *Science at War* (1948); FLORENCE N. TREFETHEN, "A History of Operations Research," in JOSEPH F. MCCLOSKEY and FLORENCE N. TREFETHEN (eds.), *Operations Research for Management,* vol 1, pp. 3–35 (1954); and GREAT BRITAIN. AIR MINISTRY, *The Origins and Development of Operational Research in the Royal Air Force* (1963).

General texts: RUSSELL L. ACKOFF and MAURICE W. SASIENI, *Fundamentals of Operations Research* (1968); STAFFORD BEER, *Decision and Control: The Meaning of Operational Research and Management Cybernetics* (1966); DAVID W. MILLER and MARTIN K. STARR, *Executive Decisions and Operations Research,* 2nd ed. (1969); HARVEY M. WAGNER, *Principles of Operations Research* (1969); DAVID R. ANDERSON, DENNIS J. SWEENEY, and THOMAS A. WILLIAMS, *An Introduction to Management Science: Quantitative Approaches to Decision Making,* 4th ed. (1985); ELWOOD S. BUFFA and JAMES S. DYER, *Management Science/Operations Research: Model Formulation and Solution Methods,* 2nd ed. (1981); GILBERT GORDON and ISRAEL PRESSMAN, *Quantitative Decision Making for Business,* 2nd ed. (1983); and G.D. EPPEN and F.J. GOULD, *Quantitative Concepts for Management: Decision Making Without Algorithms,* 2nd ed. (1985).

Methodology: RUSSELL L. ACKOFF, SHIV K. GUPTA, and J. SAYER MINAS, *Scientific Method: Optimizing Applied Research Decisions* (1962, reprinted 1984); C. WEST CHURCHMAN, *Prediction and Optimal Decision: Philosophical Issues of a Science of Values* (1961, reprinted 1982); PETER C. FISHBURN, *Decision and Value Theory* (1964), and *Utility Theory for Decision Making* (1970, reprinted 1979); and D.J. WHITE, *Decision Theory* (1969). MARILYN TAYLOR THOMPSON, *Management Information, Where to Find It* (1981), is an annotated bibliography of resources.

General techniques: BURTON V. DEAN, MAURICE W. SASIENI, and SHIV K. GUPTA, *Mathematics for Modern Management* (1963, reprinted 1978); JAMES R. EMSHOFF and ROGER L. SISSON, *Design and Use of Computer Simulation Models* (1970); HOWARD RAIFFA and ROBERT SCHLAIFER, *Applied Statistical Decision Theory* (1961, reprinted 1970); THOMAS L. SAATY, *Mathematical Methods of Operations Research* (1959); and CLAYTON J. THOMAS and WALTER L. DEEMER, JR., "The Role of Operational Gaming in Operations Research," *Operations Research,* 5:1–27 (1957).

Specific techniques: (*Programming*): RICHARD E. BELLMAN, *Dynamic Programming* (1957, reprinted 1972); RICHARD E. BELLMAN and STUART E. DREYFUS, *Applied Dynamic Programming* (1962); SAUL I. GASS, *Linear Programming: Methods and Applications,* 5th ed. (1985); RONALD A. HOWARD, *Dynamic Programming and Markov Processes* (1960, reprinted 1964); and G. HADLEY, *Linear Programming* (1962), *Nonlinear and Dynamic Programming* (1964). (*Inventory*): KENNETH J. ARROW, SAMUEL KARLIN, and HERBERT SCARF, *Studies in the Mathematical Theory of Inventory and Production* (1958); G. HADLEY and T.M. WHITIN, *Analysis of Inventory Systems* (1963); and FRED HANSSMANN, *Operations Research in Production and Inventory Control* (1962). (*Replacement and maintenance*): RICHARD E. BARLOW and FRANK PROSCHAN, *Mathematical Theory of Reliability* (1965); D.R. COX, *Renewal Theory* (1962, reissued 1967); and JOHN J. MCCALL, "Maintenance Policies for Stochastically Failing Equipment: A Survey," *Management Science,* 11(5):493–524 (March 1965). (*Queuing*): D.R. COX and WALTER L. SMITH, *Queues* (1961, reprinted 1979); THOMAS L. SAATY, *Elements of Queueing Theory, with Applications* (1961, reprinted 1983); and LAJOS TAKACS, *Introduction to the Theory of Queues* (1962, reprinted 1982). (*Sequencing and coordination*): JAMES E. KELLEY, JR., "Critical-Path Planning and Scheduling: Mathematical Basis," *Operations Research,* 9(3):296–320 (May–June 1961); ROBERT W. MILLER, *Schedule, Cost, and Profit Control with PERT: A Comprehensive Guide for Program Management* (1963); and ROGER L. SISSON, "Sequencing Theory," in RUSSELL L. ACKOFF (ed.), *Progress in Operations Research,* vol. 1, pp. 293–326 (1961). (*Networks*): L.R. FORD, JR., and D.R. FULKERSON, *Flows in Networks* (1962); E.L. LAWLER and D.E. WOOD, "Branch-and-Bound Methods: A Survey," *Operations Research,* 14(4):699–719 (July–August 1966); and JOHN D.C. LITTLE *et al.,* "An Algorithm for the Traveling Salesman Problem," *Operations Research,* 11(6):972–989 (November–December 1963). (*Competitive*): NIGEL HOWARD, "The Theory of Meta-Games," *General Systems,* 11:167–186 (1966), and "The Mathematics of Meta-Games," *General Systems,* 11:187–200 (1966); J.C.C. MCKINSEY, *Introduction to the Theory of Games* (1952); A. RAPOPORT, *N-Person Game Theory: Concepts and Applications* (1970); and JOHN VON NEUMANN and OSKAR MORGENSTERN, *Theory of Games and Economic Behavior,* 3rd. ed. (1944, reissued 1980). (*Search*): RUSSELL L. ACKOFF and MAURICE W. SASIENI, *Fundamentals of Operations Research,* especially ch. 14 (1968); B.O. KOOPMAN, "The Theory of Search," *Operations Research,* 4(3):324–346 (June 1956), 4(5):503–531 (October 1956), and 5(5):613–626 (October 1957); and JAMES B. MACQUEEN, "Optimal Policies for a Class of Search and Evaluation Problems," *Management Science,* 10(4):746–759 (July 1964).

Planning: RUSSELL L. ACKOFF, *A Concept of Corporate Plan-*

ning (1970); H. IGOR ANSOFF, *Corporate Strategy: An Analytic Approach to Business Policy for Growth and Expansion* (1965, reissued 1968), and JAMES C. EMERY, *Organizational Planning and Control Systems: Theory and Technology* (1969).

Implementation: JAN H.B.M. HUYSMANS, *The Implementation of Operations Research* (1970).

Systems engineering. Two classics are PHILIP M. MORSE and GEORGE E. KIMBALL, *Methods of Operations Research,* rev. ed. (1951, reissued 1970), the introductory discussion still useful for orientation but the specific examples mostly from World War II; and *Operations Research for Management,* vol. 1 ed. by JOSEPH F. MCCLOSKEY and FLORENCE N. TREFETHEN (1954), and vol. 2 ed. by JOSEPH F. MCCLOSKEY and J.M. COPPINGER (1956), with a broad range of examples, many of which are still of interest. Later books, with more stress on mathematical techniques, include CHARLES WEST CHURCHMAN, RUSSELL L. ACKOFF, and E. LEONARD ARNOFF, *Introduction to Operations Research* (1957, reprinted 1961), a fair survey of the field but more useful for its description of operations research methodology; and FREDERICK S. HILLIER and GERALD J. LIEBERMAN, *Introduction to Operations Research,* 3rd ed. (1980), with more emphasis on methodology (and probabilistic techniques in particular) and less suited to the general reader. KENICHI OHMAE, *The Mind of the Strategist: The Art of Japanese Business* (1982), is a good survey of strategy techniques.

The earliest significant books directly related to systems engineering are HARRY H. GOODE and ROBERT E. MACHOL, *System Engineering: An Introduction to the Design of Large-Scale Systems* (1957), and ROBERT E. MACHOL, WILSON P. TANNER, and SAMUEL N. ALEXANDER (eds.), *System Engineering Handbook* (1965). These cover the philosophy and methodology of systems engineering; however, both books are directed primarily at large military systems and are less useful for other applications. Another book that covers both philosophical and technical aspects of systems engineering is ARTHUR D. HALL, *A Methodology for Systems Engineering* (1962, reprinted 1968). HAROLD CHESTNUT, *Systems Engineering Tools* (1965), and *Systems Engineering Methods* (1967), deal largely with methodology. JAMES BOTKIN *et al.*, *Global Stakes: The Future of High Technology in America* (1982, reprinted 1984), explores specific methods of development. More general references include STEVEN L. DICKERSON and JOSEPH E. ROBERTSHAW, *Planning and Design, The Systems Approach* (1975); and DEAN KARNOPP and RONALD ROSENBERG, *System Dynamics: A Unified Approach* (1975).

Among shorter works, an article by HENDRICK W. BODE, "The Systems Approach," in NATIONAL RESEARCH COUNCIL. PANEL ON APPLIED SCIENCE AND TECHNOLOGICAL PROGRESS, *Applied Science and Technological Progress: A Report to the Committee on Science and Astronautics, U.S. House of Representatives,* pp. 73–94 (1967), furnishes a broad discussion of the field. G.M. JENKINS, "The Systems Approach," *J. Systems Engng.,* 1:3–49 (1969), offers a valuable elementary introduction to systems engineering. CHARLES HITCH, "Sub-optimization in Operations Problems," *Journal of the Operations Research Society of America,* 1(3):87–99 (May 1953), is a classic.

Ergonomics and human-factors engineering. ALPHONSE CHAPANIS, *Man-Machine Engineering* (1965), a small but authoritative survey of the field; KENYON B. DE GREENE (ed.), *Systems Psychology* (1970), a discussion emphasizing those aspects of human factors that relate particularly to modern systems, both technological and social; ERNEST J. MCCORMICK and MARK S. SANDERS, *Human Factors in Engineering and Design,* 5th ed. (1982), a large textbook covering all aspects of human-factors engineering—the best of its kind; CLIFFORD T. MORGAN *et al.* (eds.), *Human Engineering Guide to Equipment Design* (1963), a highly technical designer's guide; WESLEY E. WOODSON and DONALD W. CONOVER, *Human Engineering Guide for Equipment Designers,* 2nd ed. (1964), a less technical guide. Specific problems are discussed in DAVID STERN, *Managing Human Resources: The Art of Full Employment* (1982); CHRISTOPHER FREEMAN, JOHN CLARK, and LUC SOETE, *Unemployment and Technical Innovation: A Study of Long Waves and Economic Development* (1982); and THEODORE W. SCHULTZ, *Investing in People: The Economics of Population Quality* (1981).

Research and development. *History:* KENDALL BIRR, *Pioneering in Industrial Research: The Story of the General Electric Research Laboratory* (1957); W. RUPERT MACLAURIN and R. JOYCE HARMAN, *Invention and Innovation in the Radio Industry* (1949), a detailed account of the technical problems solved in one area; HARRY W. MELVILLE, *The Department of Scientific and Industrial Research* (1962), a history of a government research department; and ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *International Statistical Year for Research and Development: A Study of Resources Devoted to R and D in OECD Member Countries in 1963/64* (1967–68).

General: DAVID ALLISON (ed.), *The R and D Game: Technical Men, Technical Managers, and Research Productivity* (1969), essays by 15 authors covering the whole field from creativity to administration, with an extensive bibliography; E. DUER REEVES, *Management of Industrial Research* (1967), a discussion of the integration of R and D into corporate structure; T.S. MCLEOD, *Management of Research, Development, and Design in Industry* (1969), basic problems of technical management; CARL HEYEL (ed.), *Handbook of Industrial Research Management,* 2nd ed. (1968), essays by American managers on R and D problems as seen from the boardroom; ELTING E. MORISON, *Men, Machines, and Modern Times* (1966, reprinted 1984), an account of resistance encountered by new inventions; and DEVENDRA SAHAL, *Patterns of Technological Innovation* (1981), a book of case studies on a variety of topics.

Organization: ALEXANDER O. STANLEY and K.K. WHITE, *Organizing the R&D Function* (1965), descriptions of the organizational structures of leading U.S. companies; I.D.L. BALL (ed.), *Industrial Research in Britain,* 6th ed. (1968), details and statistics of U.K. industry; JOHN COCKCROFT (ed.), *The Organization of Research Establishments* (1965), essays by directors of 15 leading U.K. and U.S. laboratories; G. STUART MONTEITH, *R and D Administration* (1969), a detailed description of administrative procedures in the United Kingdom and United States, with an extensive bibliography; and ALLAN COX, *The Cox Report on the American Corporation* (1982).

Government: Report of the Committee on the Management and Control of Research and Development (HMSO 1961); and NATIONAL ACADEMY OF SCIENCES. COMMITTEE ON SCIENCE AND PUBLIC POLICY, *Basic Research and National Goals: A Report to the Committee on Science and Astronautics, U.S. House of Representatives* (1965).

Project management: A popular treatment is JOSEPH J. MODER, CECIL R. PHILLIPS, and EDWARD W. DAVIS, *Project Management with CPM, PERT, and Precedence Diagramming,* 3rd ed. (1983); a more complete work is J. WIEST and F. LEVY, *A Management Guide to PERT/CPM: With GERT/PDM/DCPM and Other Networks,* 2nd ed. (1977). An interesting, practical approach is presented in JOHN MULVANEY, *Analysis Bar Charting: A Simplified Critical Path Analysis Technique* (1969, reissued 1977).

(M.T./S.E./R.L.A./H.W.B./Al.C./T.S.McL./W.K.H.)

Industrial Glass and Ceramics

The common article of commercial manufacture known as glass is normally a transparent, hard, brittle substance formed from certain liquids that have the property of cooling below their freezing point without crystallizing, thus becoming liquids of increasingly high viscosity (stiffness), eventually so stiff that by all ordinary definitions these liquids have the properties of a solid.

Of the many substances that can form glass, the most widely used is silica (SiO_2), usually in combination with lime (calcium oxide, CaO) and soda (sodium oxide, Na_2O) in varying proportions, to make the soda-lime glass used for windows, lamp bulbs, bottles, and inexpensive tableware. Glass can be made from silica alone, but it has a high softening temperature, and is difficult to work. Nevertheless, it is the basic chemical nature of silica that determines the structure of ordinary glass and the way in which it changes from a free-flowing liquid at high temperature to a virtually rigid substance as it cools.

The silica molecule can be visualized as a single atom of silicon, to the four bonds of which oxygen atoms are attached. The bond angles are fixed at 108° to each other, in three dimensions, at the centre. The oxygen atoms thus occupy the four points of a tetrahedron surrounding the silicon. Each oxygen atom plays a role in two such tetrahedra, but the bonds that unite it with its silicon partners do not prevent them and their other three attached oxygen atoms from rotating. There is thus an almost infinite range of possibilities in the relative dispositions of the many tetrahedra that comprise even a minute speck of silica.

The silicon–oxygen bond is very strong; considerable heat energy must be supplied to make such material flow freely, because the fluid condition arises from the continued breaking, remaking, and rebreaking of this bond. Though heating and cooling crystalline silica can result in resumption of the crystalline form, if cooling proceeds rapidly enough, the temperature at which transformation could occur is passed too quickly for the necessary rearrangement of atoms to take place; thereafter, the material is too stiff. The continuous random changes in the disordered structure are brought to a halt, and the structure remains as it was in its last fluid moment; in effect, a liquid whose molecules are immobilized.

Such a random liquid-like structure is a good deal less compact than one in which a state of maximum order prevails. It is full of cavities and holes that are of very considerable importance in practical glassmaking. If a metallic oxide is incorporated into the glass, many of these spaces will be occupied by metal ions, each of which will engage one of the bonds of an oxygen atom of the silica framework, so removing one of the props of the firm three-dimensional structure. This, and the fact that the oxygen-to-metal ion bond is nondirectional, results in a much lower viscosity in the mixture of the two materials at a given temperature. Ordinary glass contains about 15 percent soda to render it workable at moderate temperatures. Lime is also necessary to minimize water solubility.

Industrial ceramics comprise all industrially used solid materials that are neither metallic nor organic. This definition expands the scope of ceramics far beyond the clay-based bodies of the potter and traditional industrial ceramics, as well as beyond the sand-dominated formulas of the glassmaker. In physical structure ceramics may be polycrystalline solids, glasses, combinations of multicrystals with glassy phases, or single crystals. The universal properties of ceramics that account for their enduring utility include: mechanical strength in spite of brittleness; chemical durability (at both normal and elevated temperatures) against deteriorating effects of oxygen, water liquid or vapour, acids, bases, and salts at all concentrations, and organic solvents; hardness, contributing to resistance against erosion; and the ability to be decorated.

Besides primary producers of glasses, cements, and multicrystalline ceramics, there are industries that rework the primary materials into subassemblies, components, or systems that have strategic and economic importance. Often these products, such as computers, electric lamps, or television picture tubes, could not exist without their ceramic constituents, yet the cost of the ceramic materials is small compared to the component and system values.

Ceramics technology draws upon the knowledge and skills of physics, chemistry, rheology, and the design arts and the process capabilities of engineering to convert starting materials into new chemical and mineral compositions, with controlled microstructures (or, more broadly, character), dimensions, properties, and behaviour.

The glass and the ceramics industries share in common most raw materials, many unit operations and processes, and the same underlying sciences, but generally they have developed as two separate industries. This has arisen because of competition for some markets, and because of differences in the technologies. In the following text industrial ceramics derived by compacting particulates and subsequent heat treatment will be differentiated from the products of the glass industry in which the materials are first heated and then formed.

This article is divided into the following sections:

Industrial ceramics 232
- History 232
- Manufacture 232
  - Processing of raw materials
  - Particle bonding and packing
  - Forming processes
  - Heat treatment
- Glazes and enamels 233

Glass products and production 234
- The general properties of glasses 234
  - Chemical properties
  - Physical properties
  - Thermal properties
  - Electrical properties
  - Optical properties
- History of glass production 235
  - From ancient times to the 19th century
  - Development of the modern industry
- Production of conventional forms of glass 237
  - Raw materials
  - Furnaces
  - Production processes
- Production of special glass products 242

Brick and tile production 243
- History 243
- Manufacturing 243
- Colours and textures 244
- Uses of brick 245
- Nonclay brick and tile 245

Abrasives 246
- History 246
- Natural abrasives 246
- Manufactured abrasives 247
  - Silicon carbide
  - Aluminum oxide
  - Synthetic diamond
  - Boron carbide
  - Boron nitride
  - Other manufactured abrasives
- Properties of abrasives 248
- Fabrication into useful forms 248
  - Preparation and sizing
  - Abrasive product manufacture
- Industrial applications 249

Industrial ceramics

HISTORY

Industrial ceramics may be said to have begun with the first barter of the potter's ware. Through ancient times and the Middle Ages, the techniques for making clayware continued to advance, and, with the coming of the industrial and scientific revolutions, the industries were ready to supply the scientist with porcelain for chemical experimentation and manufacturing and also to supply an increased demand for containers, tableware, sanitary ware, and construction materials.

Makers of industrial ceramics were thus prepared for the extraordinary demands brought by the 20th-century growth of the chemical, electrical, electronic, nuclear, automotive, and aerospace industries. For the electrical industry the first products were insulators, such as the familiar porcelain suspensions of transmission lines. For electronics the first new products were insulators for even higher frequencies, with the steatites (a class of magnesium silicate minerals) as the first novel materials. With the science explosion during and after World War II, the applications of ceramics in the electronic field grew to include compact capacitors, piezoelectric transducers, resistor and semiconductor compositions, magnetic materials (hard, soft, memory core, and microwave), and other energy converters. Improved compositions for automotive and aviation spark plugs led to high-alumina (aluminum oxide) ceramics that have been applied to many mechanical, wear, and electronic uses, including even lamps. Nuclear reactors use ceramics for moderators and controls and the ceramic forms of nuclear fuels (uranium and plutonium) in fuel rods. Strong, hermetic seals of alumina and other ceramics to metals are vital in many applications.

MANUFACTURE

Processing of raw materials. Raw materials for industrial ceramics were originally derived from specific locations around the world where geological processes had prepared ore bodies of sufficient chemical purity and appropriate particle sizes and rheological characteristics for each specific product. Examples of this include the English china and ball clays for better grades of porcelain, the Grossalmerode clay that for centuries was considered essential for pots for melting glass, the Vallendar clay of the porcelain enameller, and the Manchurian talcs deemed superior for the early steatites.

Only with the increasing knowledge about the chemical and physical characterization of these materials and their transformations during processing was it possible to replace these geologically prepared, strategic materials, first with beneficiated minerals (minerals that have been concentrated by removing unwanted materials) and then with chemically prepared compounds. Economically, the natural minerals, more or less beneficiated, still are lowest in cost and remain the largest tonnage used in most established ceramic products. Even with the more specialized technical ceramic products, the ceramist at first tended to retain a pinch of his favourite clay in the formula. Completely synthesized ceramics are still a relatively small proportion of total output, though significant in strategic and monetary value.

With the depletion of the geologically prepared resources and with the possible savings in processing costs through faster firing rates and higher yields, synthesized raw materials are becoming acceptable substitutes for natural minerals. Many of the newer ceramics demand raw materials that either never existed in natural minerals or did not exist in useful ore concentrations. One example of the latter kind is silicon carbide. A current classification of materials used in the manufacture of ceramic products includes more than 500 items.

Ceramic raw materials are classified functionally as aggregates, plasticizers, binders, fluxes, and modifiers (chemical, thermal, physical, electrical, colour). Traditional ceramic materials were largely derived from the elements most common in the Earth's crust: oxygen, silicon, aluminum, iron, and so forth, but modern ceramics draw upon nearly every element. Although oxides remain most common, other useful compounds include carbides, nitrides, borides, sulfides, and fluorides.

Mining, crushing, grinding, milling, beneficiation, and mixing of the standard materials for ceramics are similar to the operations in other chemical and engineering industries. Clays going into common bricks, sewer pipe, or stoneware may be processed directly as dug out of the ground, though even here blending, aging, and tempering for uniform distribution of water are practiced. Such impure clays are workable because they contain fillers and fluxes associated with the clay minerals. For whitewares, the clays are washed and purified by settling out some impurities and floating off others. Silicas are purified by washing and separating unwanted minerals by gravity and by magnetic and electrostatic means. Feldspars are increasingly beneficiated by froth flotation, a process in which a frothing agent is added to separate the desired material from impurities. Blending from stockpiles is utilized to maintain constancy of chemical composition.

Use of impure clays

Raw materials for the newer ceramics are generally prepared by chemical reactions involving precipitation, filtering, calcination (firing at high temperature), solid-state reactions, melting, and even vapour formation, as well as other sophisticated methods. Freeze drying has been under intensive development as a means for providing a very fine-particled, free-flowing powder feed. Spray drying has become a widely used technique for removing process water at the raw material preparation plants and in mixed batches at the ceramic-processing factory.

Much of the art of the ceramic technologist has always lain in the compounding of raw materials into batches. The formulas, traditionally jealously guarded secrets, involve selection of the raw materials that confer the desired working characteristics and responses to heat treatment, yielding the sought-for character and properties. The clays must be selected and apportioned among the kaolins and ball clays for workability, fusibility, fired colour, and other requirements. The silicas, likewise, must meet criteria of chemical purity and particle size distribution. The fluxing ingredient is selected from one or more feldspars, nepheline syenites, Cornwall stone, bone ash, fritted (finely divided) glasses, and other candidates. The considerations include balance between sodium and potassium oxides, content of iron oxides and other unwanted impurities, and particle size distribution and behaviour during heat treatment. With the burgeoning of technical ceramics no longer requiring or permitting clays and traditional fluxes, the skill of the batch formulator has become dependent on his awareness of the functions and behaviour of the various organic aids.

Particle bonding and packing. The forming of ceramics involves particle technology, including size reduction (comminution), measurement, and separation; blending and packing of particles; surface chemistry and physics; rheology, or the flow of more or less plastic materials under pressure; and bonding of particles. Other sciences contribute to the understanding and control of heat treatment.

Ceramics came into being because of the useful rheological properties of the clay–water system, with its capability of being shaped, whether as a paste or as a fluid suspension or as a damp, crumbly powder. Modern industrial ceramics often replace clay with organic or inorganic additives to make the raw material easier to form and to provide strength before firing (final heating). Clay-based ceramics undergo both chemical–mineralogical and physical changes during heat treatment to form new phases, including a glassy phase that often constitutes the bond between the grains in the new microstructure. Such heat treatment involves sintering in the presence of a liquid phase. The complex series of occurrences accompanying heat treatments includes gas evolution, oxidation, valence changes, chemical reactions, phase transformations, melting, shrinkage, and still other chemical and physical changes. The trend with modern ceramics technology is to simplify such changes by establishing the chemical and mineral changes before forming. Maximum density is sought in the forming batch, and compaction is achieved by applying the principles of particle packing. Then heat treatment brings about sintering in the solid state with

Changes induced by heat treatment

a minimum of shrinkage; gas evolution; and attendant flaws, such as warping, cracking, or the formation of unwanted pores.

Many significant ceramic products, such as thermal insulation and filters for liquids and gases, require controlled porosity, both in terms of percent by volume and in pore diameters and distributions. In fact, texture comprises the prime distinction between refractories, whitewares, and electronic and glass-ceramics that may otherwise be identical in chemical and mineral compositions.

Forming processes. Clay-based bodies utilize the plasticity of clay–water mixtures to permit easy shaping of the ware, in most cases accompanied by expulsion of air and densification. Forming processes are adapted to the dimensional requirements. The four principal processes are plastic forming, slip casting, pressing, and fusion casting.

Plastic forming. This is the oldest technique and can be visualized from offhand molding, coiling, and throwing on a wheel. In jiggering, the topside of the piece (dinner plate or bowl) is shaped by throwing a wad of body (forming batch of suitable consistency) on a gypsum plaster mold mounted on a rotating spindle; the bottom surface is formed by pressing a contoured pattern against the exposed face of the body. In plastic extrusion a stiff paste is forced through a die of the desired shape and cross section in a continuous column, which can then be cut to length. The body may be prepared separately or may be processed in a continuous apparatus that mixes the raw materials and water (pug mill), passes the mix through a vacuum chamber to remove air and thereby improve plasticity, and then forces it through the die. Structural clay products, such as brick and sewer pipe, are normally formed by this technique. Plastic formed bodies generally contain 15 to 25 percent water by weight.

Slip casting. This process utilizes a fluid suspension of the body with controlled pH (acidity or alkalinity), specific gravity, water content, and additives for deflocculation (prevention of the small particles from combining into larger aggregates) called the slip. The deflocculants (for example, a few tenths of a percent each of sodium carbonate and a sodium silicate) confer the same fluidity at about 25 percent water as would require 100 percent water in their absence. Lower water means lower drying shrinkage and consequently lessened tendency to warp and crack and better dimensional control. The slip is poured into a plaster of paris mold that withdraws enough water from the adjacent suspension to build a layer of solid body. After the desired wall thickness has been attained, excess slip is poured off. Vacuum, pressure, or centrifugal forces are sometimes used in conjunction with slip casting.

Slip casting applications

The slip casting technique is applied to many clay-free, nonplastic particulate materials. By deflocculation of the finer particles, a high-solids-content suspension can be prepared. Either alkaline or acid slips can be prepared at a pH where like electrical charges transferred to the particles cause them to repel each other. Similarly, nonclay bodies can be extruded or die pressed by using organic additives as plasticizers, binders, and lubricants. Like water, these additives are transient constituents, being expelled during the firing operation.

Pressing. Pressing forms a shape or a blank by compacting a free-flowing powder mixture in a die until the particles are bonded strongly enough to permit removal from the die and handling before heat treatment. Water content varies from 5 to 15 percent for clay down to as low as zero for bodies utilizing organic binders. In addition to the complexity of shapes (including holes) that can be formed, dry pressing has the advantage that the low water content results in less drying shrinkage. Pressing is carried out in rigid steel or carbide dies with equipment of increasing sophistication for controlling the rate and uniformity of load application, air expulsion, and ejection from the die. Flexible dies of rubbery polymers are also used for isostatic (uniform pressure) pressing, permitting more uniform density throughout a piece. Plastic bodies are being pressed by procedures that permit simultaneous forming and water removal, thus lessening shrinkage and increasing strength. The steam evolved in warm pressing acts as a parting agent to facilitate removal from the mold. Hot pressing is carried out in molds, usually made of graphite, that can be sufficiently heated by induction or radiation to bring about simultaneous shaping, densification, and in some cases chemical reactions. The resultant body can be of full density (pore free), of the same dimensions as the mold (no drying or firing shrinkage), and completely matured (fired). The chief barrier to wider use of this process has been the short life and high cost of the molds.

Fusion casting. In this process the batch is heated all the way up to melting temperatures and then shaped by casting into a mold, usually made from sand, corresponding to the casting of metals in foundries. So far, it has been largely limited to making large blocks of refractories for glass and metal melting furnaces. Other kinds of forming from melts and from vapours are being used for making ceramics, both in single crystal and in multicrystalline structures.

Heat treatment. Formed ware is generally passed through a drying oven to remove the liquid water carefully without introducing cracking or warping. The basis for the operation is to maintain initially a high humidity, so that the movement of water through the cross section is uniform. Dryers are usually based on heated air, but infrared and dielectric (radio-frequency) heating are being applied.

The goal of subsequent heating (firing) is to obtain cycles of time–temperature–atmosphere that are accurately reproducible. Again the rate of temperature increase must be carefully controlled to avoid spalling (chipping), cracking, warping, nonuniformity within a piece, and nonreproducibility from piece to piece and from day to day. Not only must dimensions be kept within tolerances, but colour and other pertinent physical properties must be maintained within acceptable limits.

Heat treating equipment

Equipment for heat treatment has undergone a considerable advance; the intermittent kilns of the potteries have been superseded by continuous tunnel kilns, and the means for transporting the ware through the cycle is undergoing active innovation.

Regardless of kiln design, there are four basic elements common to all heat treatments: (1) the source of the heat, (2) the containment of the heat, (3) the containment of the ware, and (4) the instrumentation for control of temperature and atmosphere.

Heat sources include combustion of fossil fuels and the passage of electrical current through metallic alloys or ceramics, such as silicon carbide. Induction and arc heating are not common in the ceramics industries, but very-high-frequency induction (radio frequencies in the megahertz range) finds special uses, as have other new sources of heat, such as electron beam, ion beam, and solar and arc imaging.

Containment of heat involves furnace design, including choice of refractories (furnace linings). The availability of refractory materials useful to ever-higher temperatures has been paralleled by advances in insulating refractories, cement-bonded materials that can be cast into the desired shapes, and high-temperature fibres. Lighter weight materials have made faster heating and cooling cycles possible and have contributed to lower fuel or power consumption.

Containment of ware has also moved in the direction of better heat economy and faster cycles with materials of higher thermal conductivity, higher service temperatures, less creep and deformation, and less reaction with the ware.

The use of instrumentation has also advanced rapidly with the innovations in program controllers and analysis of atmospheres. Computer-controlled firing is being evaluated.

GLAZES AND ENAMELS

Glazes and porcelain enamels provide a meeting ground for glass technology and particulate ceramic technology. Glazes are generally used for reasons of appearance but can be designed to improve other properties, such as strength and imperviousness, and to improve smoothness, thus making cleaning easier and improving resistance to electrical tracking (development of paths over the surface along which electrical current can flow).

Glaze formulation

Traditionally, glazes were the province of the ceramic technologist and were frequently synthesized from raw materials at the site on the matured ware during a special gloss firing. Now formulation of glazes has become a specialty of supplier companies that melt glaze, enamel, and bonding formulas and quench them to glassy frits (finely divided material) by rapid cooling in water or passing through steel rollers. Coating slips (suspensions) comprised of milled frits, opacifiers, bonding clays, and various modifiers are applied to the ceramic or metal base by spraying, dipping, or flowing. With ceramic bodies the maturing of the glaze may take place in a single firing with the maturing of the body. With porcelain enamels a carefully cleaned metal surface is coated from a slip (or more rarely by dry dusting), and the glass coating is matured by a rapid heating and cooling requiring only about five minutes at peak temperature.

(A.G.P.)

Glass products and production

THE GENERAL PROPERTIES OF GLASSES

Chemical properties. Although glass is used for many purposes for which other materials are less suitable because of their chemical instability, there is nothing about the glassy state as such that ensures its stability. Water glass, for example, composed of silica and soda alone, is useful only as a temporary glaze because it dissolves readily in water.

Though pure silica is insoluble in water, it has poor resistance to attack by alkaline solutions. Thus, the familiar soda-lime glass contains the agent of its eventual destruction. After soda-lime glass has been in use for a time, moisture in the atmosphere begins to leach sodium ions from the surface; these form sodium hydroxide, an alkaline substance that, in turn, attacks the silica in the glass. Though this etching process causes no significant contamination of most substances likely to be stored in the glass and produces no readily apparent effect, the submicroscopic fissures it creates ultimately have a profound effect on the mechanical strength of the article. Carbon dioxide can also cause deterioration in glass surfaces when moisture is present.

Natural etching of glass

In spite of such deterioration, ordinary glass generally has a long, useful life. Furthermore, the number of substances that will form glass is sufficiently great that special glasses can be formulated that will resist attack by specific substances at normal or elevated temperatures. Boric and aluminum oxides, for example, can be mixed with silica to produce a glass highly resistant to attack by hot alkaline solutions; and glass containing large amounts of alumina and alkaline-earth oxides is resistant to hot sodium vapour that would destroy common soda-lime glass.

In some applications, all of the desired characteristics cannot be combined in a single glass. In such cases, ways can often be found of modifying the surface of the glass itself by some form of chemical treatment that gives the glass surface chemical properties not shared by the substance as a whole. When this is not practicable, a coating of one kind of glass may be applied to the surface of another kind, and the resulting material is, in effect, two-ply glass.

Physical properties. Although everyday experience would seem to suggest that glass is an exceptionally rigid material, its behaviour at room temperature is actually that of an elastic solid. It can be bent or otherwise deformed and, so long as its breaking strain is not exceeded, will return to its original undistorted shape when the strain is removed. Indeed, many of its present-day applications would be impossible unless glass were possessed of some degree of flexibility.

Though everyday experience also suggests that glass is brittle and of low strength, it is actually quite strong and will not fail as a result of simple compressional loading. Glass breakage, even when it results from impact, commences as tensile failure (*i.e.,* the glass is pulled apart, as described below).

When free from surface flaws, glass also possesses immense tensile strength. Generally arising as a result of chemical activity, surface flaws can be so slight as to defy detection, even with optical aid. However, their weakening effect can be measured by comparing the tensile strength of glass that has been in use awhile with the strength of glass in an undamaged state; *i.e.,* with its fire-finished surface unmarred by contact with other substances. Nearest to perfect glass are fine, newly produced glass fibres, which have been found to support loads of up to 500 tons per square inch (about 70,000 kilograms per square centimetre), a tensile strength about five times as great as that of the best practically available steel and about double the theoretical maximum for steel.

The way in which surface defects reduce tensile strength is demonstrated when a diamond cutter is employed to produce a controlled fracture in a sheet of glass. Once the sheet has been lightly scored, pressure applied on the scored side, over a suitable support, will cause separation. The relative rigidity of glass itself provides the lever that concentrates the available force on the very few interatomic bonds at the base of a tiny crack (microcrack). The ultramicroscopic structures fail under an assault by what is, relatively, truly colossal stress, which may amount to a tensile load of thousands of tons per square inch or centimetre. Effective leverage increases as the crack becomes deeper. The liquid structure of the glass presents no crystal boundaries or discontinuities that might arrest the progress of the crack, and it rapidly travels through the glass.

That glass is subject to fatigue fracture can be demonstrated by subjecting it to prolonged static loading. It will eventually fail under stresses far lower than those that a precisely similar sheet of glass can initially withstand. It is believed that such fatigue fractures are assisted by chemical assault. Once a microcrack exists with an orientation that allows it to be opened by the effect of the applied load, silicon ions in the exposed base of the crack can be dissolved out. Though deepening of the crack by such means will initially be very slow, purely mechanical forces will take over in time and the crack will spread very rapidly, eventually proceeding through the medium at a speed higher than that of sound.

Chemical action and cracking

The belief that chemical action does indeed assist in the initial deepening of a crack is supported by the fact that water, as has been seen, readily removes sodium ions from the glass. This property has long been exploited by glassworkers, who commonly wet the cracks made by a cutter before breaking sheets of glass.

Spontaneous fracture of badly annealed glass, in which internal strains are frozen in, can be seen from the preceding paragraphs to be inevitable when the surface has accumulated defects of the kind that are the natural consequence of normal use. Careful annealing to remove such strains is an important part of glassmaking technique. No less important are the methods by which glass surfaces are made more resistant to minor damage or protected against it.

Molten glass acts like a liquid in contrast to ductile metals, which quickly develop weak spots causing them to narrow (neck) and eventually break when drawn by straightforward methods. Glass does not neck when drawn; tubing and rod can be stretched steadily, and almost indefinitely, as long as the viscosity of the material is carefully controlled.

Thermal properties. Glass expands when heated, shrinks when cooled, and is a poor conductor of heat. Though the surface of a slab of glass can lose heat very quickly by radiation into the atmosphere or through contact with any efficient heat-conducting medium, the interior loses heat by a slow process of conduction. In consequence, abrupt changes of temperature in rigid glass result in temperature differences between the interior and surface (thermal shock), creating stress that often leads to breakage. When glass is suddenly cooled, for example, the expanded interior puts the surface in a state of tensile stress. Careful annealing of ordinary glass eliminates the freezing in of strains that might aggravate stress resulting from mild thermal shock. Some kinds of glass expand and contract considerably as a result of temperature changes. Even well-annealed objects made from these are likely to fail as a result of thermally induced tensile stress acting on minor surface damage. Common soda-lime glass, with

Thermal shock

its high coefficient of thermal expansion, is unsuitable for any applications in which thermal shock may be encountered. Borosilicate glass (Pyrex type) is far more resistant to tensile stress; while both the so-called Vycor glass, containing 96 percent silica, and pure silica glass are virtually immune to thermal shock.

The reaction of a glass to changes in temperature is very important in optics, in which use a dimensional change amounting to a single wavelength of light can very seriously affect performance. For critical applications, Pyrex and Vycor types are increasingly used.

Because no real stresses exist in molten glass, sudden extreme chilling of its surface has a contrary effect to that described above. Such chilling immediately creates a very tough outer skin that has a far higher viscosity than the interior. As the glass cools to room temperature and temperature differences diminish between interior and surface, the interior is denied the dimensional shrinkage normal to its temperature condition. In consequence, the interior is left under severe and permanent tensile stress; it exerts a powerful compressional force on the surface. Glass in this condition is very resistant to impact and more resistant than untreated glass to thermal shock.

Electrical properties. Although most glass contains metallic ions (charged atoms or molecules) capable of carrying a current, the high viscosity of glass impedes their movements and electrical activity. Thus, glass is an efficient electrical insulator, though this quality varies with viscosity, which is a function of temperature. Indeed, if glass is first heated sufficiently by normal methods, it can be treated as a conductor and melted electrically.

Though the use of glass as an insulating material in electrical equipment might seem to be ruled out by its conductivity at high temperatures, special varieties have been developed that retain desirable insulating properties at several hundred degrees Fahrenheit. Furthermore, the conductivity of small components can be greatly reduced by prolonged annealing, which ensures maximum compactness of the glass and considerably decreases movement of ions.

Optical properties. Unlike those in most solid substances, electrons in glass molecules are confined to particular energy levels, and hence they cannot absorb and re-emit light energy in the form of photons (units of light energy) by skipping from one energy level to another and back again. Light energy, in consequence, travels through glass instead of being reflected, so that glass is transparent. Furthermore, the molecular units in glass are so small in comparison to light waves of ordinary frequency that their absorption effect is negligible. Light waves in some frequencies, such as the infrared or ultraviolet, can cause glass molecules to vibrate, however, making the glass opaque to light of these frequencies. Furthermore, glass to which certain metallic oxides have been added will absorb wavelengths of certain colours, thus appearing tinted to the eye.

Because of its amorphous liquid structure, glass is transparent in a manner associated with ordinary liquids rather than with transparent crystals. Light is not polarized (constrained to vibrate in a single plane) in passing through glass, nor is it much reflected, as it is in such materials as mica, which has been strained at internal optical boundaries. This optical neutrality can, however, be destroyed locally in glass as a result of internal structural strains such as arise from inadequate annealing.

Refraction of light

Light rays are deviated on passing from one transparent medium to another of different density unless the boundary between the two is at a right angle to the direction in which the rays are travelling. The extent to which light is bent, or refracted, on passing into a medium such as glass depends on the angle at which the light meets the surface (angle of incidence) and on the relative densities of the glass and the other medium involved, usually air. Refraction can be expressed as a constant, the index of refraction, derived mathematically from the ratio of the sines of the angles of incidence and refraction. The constant has a different value for different wavelengths and for different types of glass.

On emerging from glass into air, a light ray will be similarly deviated, but in the opposite sense. That is to say, if the faces of the glass sheet are flat and parallel, the final path will be parallel to the original path of the ray, but if the faces are not parallel (as in a lens or triangular prism), the ray will emerge in an entirely new direction.

Since the index of refraction is different for different wavelengths of light, a prism can be used to separate the constituent wavelengths of "white" light. A lens can be considered as a series of infinitely small prisms the face angles of which are such as to deviate parallel rays of light toward a single point. Though no single lens can do this for all colours, a combination of lenses of different types of glass can do so for several different wavelengths, thus producing a focussed image free from any noticeable false spread of colour.

If the angle at which light rays inside glass arrive on the second surface is large enough, the deviation on emergence will give the light rays a direction parallel and close to the surface on emergence. This critical angle cannot be exceeded without the ray being reflected on the glass–air boundary. Called internal reflection, this phenomenon is widely exploited in making prismatic binoculars, single-lens reflex cameras, and range finders. It also has important new applications in light piping and fibre optics, which are described below.

HISTORY OF GLASS PRODUCTION

From ancient times to the 19th century. The oldest specimens of manufactured glass, beads from Mesopotamia and ancient Egypt, date from about 2500 BC. These beads and very much later specimens are of particular interest because their dates of origin reveal how long a period of time elapsed before glassmakers became aware of the many ways in which hot plastic glass could be manipulated.

Mesopotamian glass

Small glass vessels, nearly 4,000 years old and of presumed Mesopotamian origin, were neither blown nor molded but painstakingly sculpted from solid blocks. Only after a thousand years or more of experience did glass vessels begin to appear in Egypt; these were produced not by exploiting the ductile nature of hot glass but by an extension of the already established method of glazing pottery. Molten glass was poured over a shaped sand core in successive layers until the vessel was sufficiently sturdy. The surface was decorated with molten drops of coloured glass poured over the outer layer and pressed into it before the vessel had cooled. This method was employed for more than a thousand years until the introduction of a simple tool that revolutionized glassworking.

The blowing iron. Probably dating from about 200 BC, the blowing iron is believed to have been used first in Babylon and adopted with enthusiasm by the Romans. Along with the blowing iron (an iron tube about 5 feet, or 1.5 metres, long with a mouthpiece at one end and a knob for holding the soft glass at the other), the craftsman employed a marver, or polished block of iron, and a pontil, a solid iron rod. A blob of soft glass on the knob end of the blowing iron was first rolled into a suitable shape on the marver. Next, it could be blown either into a constraining mold or freely in air. The pontil was employed to roll, squeeze, twirl, or pinch a similar blob of glass into shapes of almost any desired complexity and to weld them to the globe of hot glass on the blowing iron, making, for example, the stem and foot of a goblet. The still soft globe could be cut from the blowing iron with shears; the completed vessel, when cooled, was made to part company with the pontil by means of a sharp rap on the iron (cracking off).

Millefiori glass

Another development in glass manipulation adopted by the Romans led to what became known eventually as millefiori glass. Threads of coloured glass, made by stretching when hot, were laid in carefully arranged bundles so that they could be fused together to make a "cane," which could in turn be cut into slices, all with exactly the same surface pattern. The pattern units were then laid side by side on a suitable sand core and heated until they fused together to form a continuous self-supporting shell.

Both the Romans and the Egyptians showed astonishing skill in the way they used metallic oxides as colorizers. Very small differences in oxide content can drastically

affect the final colour of a glass; yet colours and tints were reproduced time and again with remarkable consistency. Copper was used to make green and ruby-red glass; iron produced black, brown, and green; antimony, yellow; manganese was employed to make purple and amethyst glass. An opaque white glass, made by using tin, was important in glass cameo work, of which the famous Portland vase, made in 1st-century Rome, is an outstanding example. To make this vase, a layer of white glass was superimposed on a darker material and afterward sculpted, pierced, and cut away to leave the white figures in relief against the darker background.

Roman attempts to make flat glass by pouring slabs about one-half inch (12 millimetres) thick were unrewarding. Proper transparency cannot be achieved by such means without grinding and polishing the cast material; the lack of transparency and the difficulty encountered in making any but small panes by this method led to the introduction of stained-glass windows, first used in the Eastern Roman Empire in the early 12th century.

The Middle Ages. Glassmaking skills in Europe declined after AD 200; for about a thousand years, standards remained far below those of the Romans. The range of articles, as well as the quality of the material, was poor; the glass was of inferior colour and marred by streaks and bubbles. The stained-glass windows that began to appear in the new Gothic churches in Europe in the 12th century reached their full splendour in the 13th and 14th centuries. Many of the colours were produced by the method of fusing stains to the surface of the glass. Clear, colourless glass proved extremely difficult to achieve, however.

The real revival of glassmaking skills in Europe came by way of Venice through contact with the Eastern Roman Empire (Byzantium). The Venetians made discoveries and innovations of their own, learning, for example, to eliminate all accidental colorizers from a glass melt by adding countercolorizers. The natural result was a gray glass the overall transparency of which was even less than that of the otherwise slightly tinted glass. So long as the amounts of the original colorizer and of its antidote were small and the thickness of the finished article was slight, however, the loss of transparency was less noticeable than the unwanted colour would have been.

Glass-making in Venice

The Venetians eventually redeveloped all the skills of the Romans. Their products and their secrets were so much sought that regulations were passed forbidding the emigration of workers. The glassworks, said to be a mile long, was moved in its entirety to the island of Murano in 1291 because of the fires it had caused. By the 16th century, however, most of the secrets and all of the manual skills of Venice had spread. Glass of improving quality was being produced over most of Europe in ever-growing amounts.

Although ingredients varied from place to place, most glass produced in the past, as today, was soda-lime glass. Satisfactory for most purposes because it is very stable chemically and of reasonable hardness, soda-lime glass is also easily made. Its moderate softening temperature makes it very workable, and it can be readily resoftened a number of times if necessary to complete an article. Above all, the materials needed for its manufacture were plentiful; sand and limestone exist almost all over Europe, and the soda ash was readily obtainable from the hardwood forests that also provided fuel for the furnaces. Desirable but scarcer materials included potash, produced by burning seaweed and favoured by the Venetians above soda ash, and calcined and crushed river pebbles (selected for their whiteness), which most Italian glassmakers preferred to common sand.

Glass furnace

The 15th-century German authority Georgius Agricola recommended a glass furnace nine to ten yards (eight to nine metres) wide and five times as long, large enough to keep 100 or more craftsmen busy. Agricola's specifications would have made glassworks in the mid-16th century very large compared to other industrial operations at the time. All the ingredients of a glass batch were roasted in order to remove as much moisture as possible before use. Mixing was carried out in a separate furnace; and after being stirred and heated for 24 hours, the ingredients were ready to be shovelled into fireclay pots for melting.

The furnace was roofed over and fitted with flues; its floor, called a siege floor, stood somewhat above ground level. The fuel chamber, which lay underneath this siege floor, had a series of holes along its length. These allowed the fuel to be ignited at many points and provided adequate access of air. The hot gases from the burning fuel entered the melting chamber through a series of vents confined to about three-quarters of the length of the siege floor. This arrangement provided the melting chamber with a relatively cool zone into which pots of molten and refined glass could be pushed to achieve working viscosity as they became needed by the workers.

Window glass. Glassblowing was indispensable to nearly all glassmaking operations until the 20th century. Not only tableware and bottles but flat glass, too, depended on the use of the blowing iron. Until well into the 19th century, the only method of making window glass was, first, to blow a large globe of glass; this was cut from the blowing iron after welding the pontil to a point diametrically opposite. The open side of the glass balloon, 20 inches (50 centimetres) or so in diameter, was then flattened on the marver; then, with the pontil being continuously rotated, the glass was reheated until it became very soft. Immediately after it was removed from the heat, the rate of spin of the pontil was increased; suddenly, the centrifugal stress on the soft glass caused it to flow rapidly outward from the centre and to take the form of a large disk, which was kept spinning until the glass was rigid.

Such glass of course was not truly flat. The disk was very uneven, being thickest near the centre and marked by concentric circular waves; at the very middle was the fractured nub, or crown, marking the point of former attachment to the pontil. Disks more than about five feet (1.5 metres) in diameter were hardly practical. The manufacturing process and the need for great economy in cutting the material, rather than aesthetic taste, account for the sector-shaped fanlights and small window panes in English Georgian houses of the 18th and early 19th centuries. Polished cast-plate glass, first produced in France toward the end of the 17th century, was little used, even by the wealthy, because of its prohibitive cost.

First made in the 17th century, using a large proportion of calcined flints with potash substituted for soda ash, flint glass was initially a failure. It spontaneously developed a maze of fine cracks that made it useless. The addition of lead oxide, however, eliminated this defect and produced a fine lustrous glass, soft enough to be cut and engraved easily and of greater refractive power than the common crown glass. It was also more dispersive of light, but not to the same degree. Telescope makers took advantage of these properties: they placed a flint-glass lens and a crown-glass lens one behind the other to eliminate the worst of the coloured haze surrounding telescopic images of bright objects.

Flint glass

Development of the modern industry. During most of the 18th century, glass, apart from a few small windows, was a rarity in the homes of ordinary citizens. Containers and tableware were commonly made of pewter or earthenware. By the end of the 19th century, however, glass was a common material, pouring from the furnaces in a seemingly endless flood to be etched, scratched, engraved, and plated; rolled, ground, and polished; blown, cast, and molded into every conceivable form and for almost every conceivable purpose.

The first important mechanical manufacturing innovation was the hand-operated split mold, patented in 1821, which was made of two hollow blocks of iron. This device made it possible to blow the whole bottle, instead of the body only, in the mold. Sixty years later, semi-automatic bottle-making machines were coming into use.

The production of flat glass by breaking and spinning a blown globe gave way to the glass cylinder, blown by using compressed air, which could be slit lengthwise, reheated, and allowed to flatten on an iron table under its own weight. Although the natural fire finish was destroyed on one surface, the final product, still far from being truly flat, was flatter than crown glass. Output increased dramatically, so that the glassmaker and the engineer were able to astonish the world by midcentury with the Crystal

Palace in London, made from 300,000 sheets of cylinder-blown glass resting on a lightweight iron framework.

Fundamental to the great increase in output was the introduction of the regenerative furnace, originally developed for the metallurgical industry, in which hot gases were recycled. This furnace provided far more usable heat from a given amount of fuel than had been possible before, and enormously speeded the melting process. Of even greater importance was conversion of the old siege floor into a vast tank in which the materials were melted directly. Fed in at one end, the batch melted and flowed steadily toward the far end, at which it was taken up by the working teams, or later by machines, that converted it into finished glass products.

The greatly increased flow of glass from the new furnaces found steadily improving machines waiting to receive it, and the increasing amounts of mechanical power available began to make plate-glass polishing, though still costly, a practical business proposition that soon started to change the appearance of the main thoroughfares in large European and American cities.

Optical glass

Glass of optically reliable quality was extremely rare at the beginning of the 19th century, except perhaps in France, where a means had been found for producing truly homogeneous flint. Optical glass, that was formulated specifically for the lens and instrument designer, however, did not exist. Its development called for the application of an altogether more scientific and less empirical approach to the problems of the effects of ingredients on the properties of glass. This progress came with the remarkable collaboration of the German scientists Carl Zeiss, Ernst Abbe, and Otto Schott in Jena, with the result that many new elements came into use in glassmaking; the effect on the quality of microscopes was immediate.

The new scientific approach to glassmaking was paralleled by a more scientific attitude toward the properties and possibilities of glass, so that new methods of using and handling glass continue to multiply, as do the applications of glass. If the beginning of the modern industry is dated from the beginning of the 20th century, then the number of important innovations and changes in glassmaking technique during its short history can be seen to outnumber very easily all those of the preceding 4,500 years. So numerous are they that a list would be tediously long and a description impossible. For example, there are at least four major changes in the methods used to produce flat glass; twin grinders that can produce about 400 square yards per hour of polished plate glass; machines that provide 2,000 or more lamp-bulb glasses per minute, and others that produce astronomical lengths of glass filament. There are also many newer glass types: heat-resistant, low-expansion glass; thermally toughened and laminated glasses; photosensitive and photochromic glass; foam glass, glass filters, glass fabrics; glass that is transparent to ultraviolet and to infrared rays; and glass that provides protection against X-rays and gamma radiation. Some of these processes and new glass types are described more fully in the later text.

PRODUCTION OF CONVENTIONAL FORMS OF GLASS

Raw materials. The greater part of glass produced today consists of variations of the silica–lime–soda mixture. Enormous quantities are used for flat glass of all kinds and for containers, including light bulbs. Optical glass, even in technologically advanced countries, accounts for only about 4 percent of the total value of the output; in terms of tonnage, it is very much less.

Modern continuous methods of production involve a high degree of organization. One of the first requirements is a continuous supply of raw materials, consistent in quality, that lend themselves to standardized cleaning treatment. Thus, the glass manufacturer first looks for conveniently situated sources of adequate size that he can own or control. An idea of the standards required of the ingredients at firing time is given by the fact that the iron oxide content of even common domestic ware is unlikely to be more than about one part in 2,000. For higher quality materials, this figure is frequently cut to one part in 10,000 and sometimes very much less.

Soda is normally provided in the form of sodium carbonate (soda ash), and lime as calcium carbonate. Magnesia is added to the batch when dolomitic lime is used. For more specialized glasses, boric acid and borax are employed to provide boron oxide; zinc is directly oxidized to provide zinc oxide; litharge (red lead) is ordinarily used to provide the lead oxide for crystal ware and for dense flint optical glass, in which it may amount to 50 percent by weight of the total batch. Many minor ingredients are used as colorizers, decolorizers, and to assist in refining the glass, or to give it special working properties or special characteristics when finished. Oxides of copper, chromium, cobalt, iron, and nickel are used as colorizers, as are colloidal gold and copper. Antimony and arsenic are useful refining agents: phosphates and fluorides are employed in opal glass. The table lists the principal ingredients of important types of glass.

Chemical Contents of Various Kinds of Glass

	composition (percentage)
Soda-lime silica	silica (SiO_2), 70; soda (Na_2O), 15; lime (CaO), 10; magnesia (MgO), 2.5; alumina (Al_2O_3), 2.5
Borosilicate	silica (SiO_2), 60–80; boric oxide (B_2O_3), 10–15; alumina (Al_2O_3), 1–4
High silica (shrunken glass)	silica (SiO_2), 96; boric oxide (B_2O_3), 3–4
Aluminosilicate	silica (SiO_2), 5–60; alumina (Al_2O_3), 20–40; lime (CaO), 5–50; boric oxide (B_2O_3), 0–10
Light barium crown	silica (SiO_2), 45–50; boric oxide (B_2O_3), 3–5; soda (Na_2O), 1; potassium (K_2O), 7; barium oxide (BaO), 20–30; zinc oxide, (ZnO), 10–15
Dense barium crown	silica (SiO_2), 30–40; boric oxide (B_2O_3), 10–15; zinc oxide (ZnO), 0–10; alumina (Al_2O_3), 0–10
Extra dense flint	silica (SiO_2), 20–40; potassium oxide (K_2O), 0–10; lead oxide (PbO), 50–80
Soft solder glass	silica (SiO_2), 5; boric oxide (B_2O_3), 15; lead oxide (PbO), 64; zinc oxide (ZnO), 16
Antiradiation	silica (SiO_2), 20; lead oxide (PbO), 80
Infrared (semi-conducting)	arsenic (As), 44; tellurium (Te), 24; iodine (I), 32
Photosensitive	silica (SiO_2), 72; soda (Na_2O), 17; lime (CaO), 11; gold (Au), 0.02; selenium (Se), 0.04
Photochromic	silica (SiO_2), 60; soda (Na_2O), 10; boric oxide (B_2O_3), 20; alumina (Al_2O_3), 10; iodine (I), 0.9; silver (Ag), 0.6; chlorine (Cl), 0.3

Use of cullet

One particularly important ingredient of glass has not been mentioned because it does not, or should not, affect the balance of the constituents of the final product. Cullet, or broken waste glass of the same type as that being manufactured, almost always forms from 50 to 75 percent of the batch. Cullet acts as a solvent. Because any glass becomes fluid at a lower temperature than any of its constituents and because glassmaking depends on the intersolution of its ingredients rather than melting, the entire process is enormously speeded by the quick provision of an adequate amount of fluid glass.

Uniformity of sand-grain size is also important in melting. Sand grains of different sizes tend to separate as a result of the vibrations suffered during transit and handling. During the melting process, their behaviour would be unpredictable and the time allowed for homogeneous intersolution of the glass ingredients would sometimes be inadequate. Streaks of glass of differing viscosity would result. This, in turn, would cause undesirable optical effects and possibly provide a mechanically weak final product.

Furnaces. The regenerative furnaces used in large-scale glass production are invariably tank furnaces, normally designed for continuous production throughout the whole useful life of their refractory (heat-proof) linings. These linings eventually corrode to a point at which they become deeply scored, and pieces are likely to break off. A furnace lining may last up to several years; its life depends not only on the material used in making it but also on the type of glass being made, for some glasses are much more corrosive than others.

A tank furnace may be nearly 30 feet (nine metres) wide and five times as long, with a capacity of more than 1,000 metric tons. The frit (glass ingredients), well dried and

thoroughly mixed, is mechanically conveyed to overhead hoppers located at the end of the furnace, through which it is allowed to trickle down across the whole width of the furnace for introduction into the melt by mechanical pushers. The rate of feed is controlled automatically by the withdrawal of glass at the far end of the furnace, so that the level in the tank remains constant.

The furnace may be oil or gas fired; large fuel reserves are necessary in order to ensure continuity of firing. Electric furnaces, employing resistance heating, although used for special glasses produced in relatively small quantities, are too costly to operate for production of common glass. Electric boosters in regenerative furnaces, however, are used; these not only help to increase output but also provide a means of quickly and accurately controlling the heat distribution along the tank.

Heat distribution

Heat distribution is most important if a continuous supply of properly refined glass is to be delivered to the forehearth (point of extraction) at the required viscosity. During the early stages of the melting process, the newly introduced batch is converted into a stiff, doughlike mass in which the ingredients participate in the chemical reactions that will eventually yield glass. Even when odd lumps of broken glass (cullet) cannot be seen, as they can for a foot or two within the tank, the mass looks uncompromisingly gritty and full of bubbles. Steady addition of new material forces the batch further into the tank, where the temperature is maintained at about 2,700°–2,900° F (1,500°–1,600° C). At this temperature common glass is as fluid as water; bubbles formed by trapped air and gases liberated from chemical agents escape. The further end of the tank, toward which the material continuously flows, is much cooler (less than 1,800° F), so that working viscosity is assured at the delivery point.

Production processes. *Flat glass.* All previous methods of obtaining flat glass by flattening blown globes and cylinders and by drawing cylinders are now obsolete. Even the Colburn process for direct continuous production of flat sheet has been superseded. This was a horizontal process, which, though continuous, marred the fire-finished surface of the material. The principal methods used today are vertical drawing of flat sheet and roller casting, flat casting, and floating on molten metal.

Vertical drawing of flat sheet is basically an extension of the principle applied by the hand craftsman who relies on the ductility of very hot glass to draw a bar attached to his pontil into a long, thin rod of uniform cross section. Glass is delivered from the furnace by way of a throat, or bridge, which holds back any overly viscous patches and controls the flow into an adaptation of the forehearth called a drawing trough, or drawing kiln.

Drawing operations commence by dropping a long bait, made of openwork iron, into the hot molten glass passing through the drawing slot of a form, or drawing boat, in the trough. The glass quickly welds to the metal, which is then slowly raised, with glass running from it much as treacle runs from a knife blade. Like treacle on a knife blade, too, it would quickly run toward the middle and form a narrow stream, but for the fact that the edges of the rising sheet are cooled immediately as it rises from the drawing boat by passing them between, but not touching, a pair of water-cooled steel boxes. The edges are then gripped by cooled, electrically powered knurled rollers, which drive the hot, growing sheet upward to the cutting loft. The iron bait is discarded; its only role is to start the process by raising the glass until it can be gripped at the edges, after which the rising sheet will be self-sustaining so long as the gripping rollers continue to turn. The thickness of the sheet can be closely controlled by the speed at which it is drawn from the trough.

The sheet is forced to rise through an annealing lehr (oven) in which close temperature control ensures that it is free from strains when it reaches the cutting loft about 30 feet (nine metres) above the furnace.

In the cutting loft, the glass sheet comes at last under some degree of direct human control. As it rises through the floor it is seized by large suction pads, suspended from overhead gear, and worked hydraulically by a human operator. As the top of the sheet reaches a predetermined height, it automatically triggers a mechanically operated cutter, which sweeps across the glass at an angle that allows for the continued rise of the glass during the cutting operation. A slight movement of the levers attached to the suction pads separates the standard sized sheet, which is then placed on an easel. There, the edges, marred by the gripping rollers when still soft, are cut away. These and the scraps that gradually accumulate as the result of accidental breakage are shovelled into a chute to be returned to the melt as cullet.

Rolled glass

At one time plate glass was produced by pouring it onto large tables with raised edges. Heavy rollers were then employed to squeeze the glass to an even thickness determined by the depth of the table edges. The modern equivalent of this process involves passing hot glass as a continuous ribbon between a pair of rollers so spaced as to control the thickness of the sheet. Though the glass is still hot and plastic when it issues from the forming rollers and passes over other rollers, it has attained a reasonable flatness by the time it reaches the annealing lehr. Though general surface quality suffers through contact with the metal, the final product, while lacking in transparency, is completely adequate for a number of uses. It is often given a texture, usually on one side but occasionally on both, by specially patterned rollers, so that it emerges finally as single- or double-reeded glass, so-called hammered glass, arctic glass, or Georgian glass; all these effects are intended to reduce further, rather than enhance, transparency. The process is also used to make wired glass by resoftening two sheets so that they can be united by rolling pressure over an intervening wire mesh.

The huge quantities of plate glass for the building industry are produced, like the bulk of other modern glass, by regenerative tank furnaces of the continuous type. Modern furnaces can easily deliver continuous ribbons up to eight or nine feet (2.4 or 2.7 metres) wide; these can be processed by twin grinders so that they eventually emerge, ground and polished on both faces, as high-quality plate glass produced at up to about 1,000 feet (300 metres) per hour.

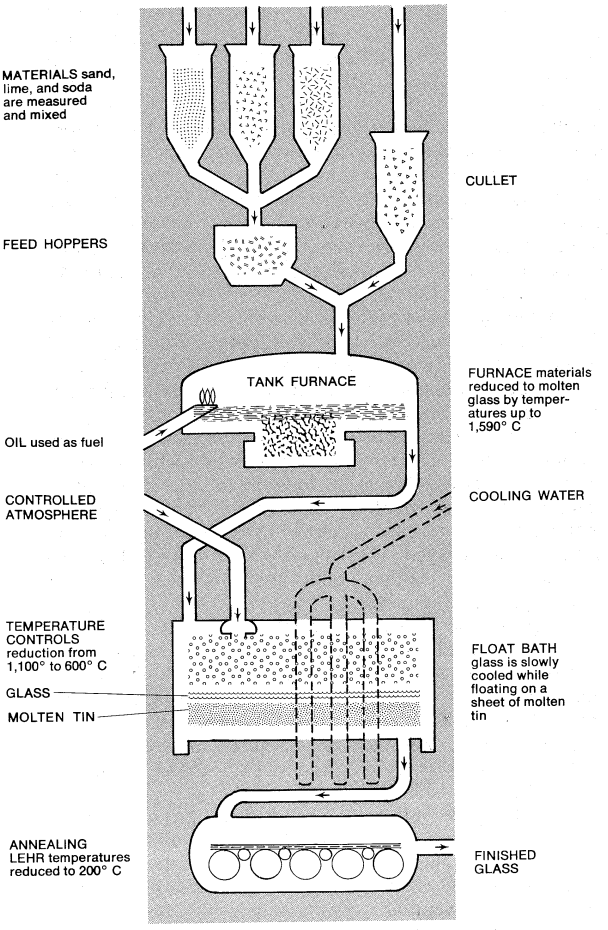
The degree of flatness and parallelism achieved by these machines is of a high standard, which can be improved only by optical methods of grinding and polishing. Indeed, so good is it that selected small pieces of up to several square inches can often be brought to a high optical standard without further grinding by polishing according to traditional handworking methods.

Flame-finished surfaces have already been seen to be impossible to preserve when plastic glass comes into contact with solid materials. The finish can only be restored by reheating, or replaced by grinding with abrasives and then polishing with agents such as rouge or cerium oxide.

Sheet glass of admirable flatness for many common purposes, unmarred by glass-to-metal contact, is produced by the continuous vertical draw processes described above. Its virtues are most apparent, however, in thicknesses of about 1/16 to 1/8 inch (1.6 to 3.2 millimetres). For thicknesses not much greater than these, plate glass was, until the late 1960s, always necessary for purposes in which quality was important. Grinding and polishing costs are disproportionately high relative to weight cost in such material. Yet there is a great demand for glass of good optical quality in these very gauges in which polishing costs are relatively so high. Glass of the required thickness and quality is now being produced in ever-growing quantities without expensive mechanical reworking by means of the float-glass process evolved in Great Britain.

Float glass

In the float-glass process illustrated in the Figure, the frit is melted in an oil-fired continuous tank furnace and, as in making plate glass, emerges as a continuous band. In the new process, it issues through a slot onto the surface of a pool of molten tin contained in a closed chamber, in which the atmosphere is carefully controlled to avoid oxidation of the metal, which would disastrously affect the transparency. Inside the chamber, the glass is heated from above to a temperature that renders it fluid, so that it takes up the form of a plane-parallel film of the required thickness on its bath of molten tin. Once past the high-temperature zone, the glass flows on through the chamber,



Basic steps in the float-glass process.
Drawing by D. Meighan

still supported by the tin, until it is quite hard; then it passes into the annealing lehr. The end result is a glass that retains the fire polish of crown glass and vertically drawn sheet and that also has a degree of flatness and parallelism rivalling glass produced by the more costly grinding and polishing process.

Bottles and containers. The continuous regenerative tank furnaces used for the production of flat glass serve equally well with minor modifications to supply glass for the high-speed production methods of the container industry. Container-production methods consist of molding and of extensions of the ages-old glassblowing process, frequently in combination. By their nature, these are intermittent rather than continuous processes. They could, however, be given a virtual continuity if all the necessary stages in the production of a glass container could be passed through at sufficient speed. There is, however, a limit to the extent at which glass can be hurried through the complex operations necessary to make it into a bottle. The answer to the problem is the multi-station machine, which can be likened to a carousel, at every arm of which a bottle is at some stage of production. The machine revolves not continuously but in a series of stops and starts, as each arm receives its gobbet of hot glass, which has to be shaped and then removed to leave its place vacant at the end of the circuit so that a new measure of hot glass can be processed.

Glass molds are generally fabricated of fine gray cast iron worked to a polished finish. The term iron mold as used by glassmakers refers to a type of mold, made in one piece, for forming shallow, easily removable items. Other molds, even though made of iron, are called press molds and paste molds.

Blow molding

Blow molding demands that the glass be provided in precisely measured amounts of the right consistency when introduced into the blowing mold. Furthermore, unless the container is to be very small, it must be of the right shape if correct distribution of the glass thickness is to be possible. This requirement calls for a preliminary molding operation to make the partially hollowed blank, or parison. At this stage the rim, or collar, of the bottle attains its final form in a removable part of the mold called the neck ring. The glass is introduced into the parison mold either by suction or, downward, by gravity, and it is hollowed either by a puff of compressed air, which also ensures proper overall contact with the mold interior, or by pressure of a suitably shaped die. The parison is then transferred by means of the neck ring to the blow mold in which it is finally shaped.

Press molding is little different in its essentials from the making of a parison by pressing, except that the pressing can be permitted to lose a good deal more heat during processing because it is the end product. Solid molds of low cost can be used instead of expensive opening molds, because they rarely need to be of great dimensional precision. A measured amount of glass is dropped into the mold and pressed to shape by a cooled plunger so that it fills the controlled space between the two units. Ashtrays, dishes, and ovenware are typical items so produced. Quality products are finally polished by high-temperature gas flames on the inside and around their rims while still in the mold.

As in other glass-shaping processes, molding involves careful temperature control. If the glass is too fluid when it is blown, a proper distribution thickness will be impossible; it may even be torn by sticking to the iron. If it is too cool, it will not flow as it should; the surface of the completed item will be of poor quality and any fine detail, such as a trademark or lettering, is apt to be poorly reproduced. Good container processing, especially in large sizes and complex forms, frequently requires that much ingenuity be exercised in distribution of the mold's own mass as part of the means of localized temperature control.

For practical reasons, many items have to be blown into molds despite the fact that their finish must be such as cannot be obtained when glass-to-metal contact is involved. Ordinary table glasses of tumbler shape are an example. The so-called paste mold is an iron mold lined with material, such as soap or beeswax, to which wood flour or sawdust can be made to adhere. Once the flour or sawdust has been converted into an absorbent carbon coating, it can be dampened before the glass is blown into it. The glass is kept rotating while being blown; the steam pressure generated by the heat prevents the glass coming into contact with the mold much as air pressure keeps a hovercraft from contact with the ground. Carboys and similar large containers can be blown into paste molds by skilled workers; at the other extreme are small electric bulbs, which are blown into paste molds at rates of many hundreds per minute.

Until 1821 only the body of a container could be blown into a mold; shape and general finish remained poor for many decades afterward. Even today most mold-blown articles need some finishing before they can be considered complete. Molds and neck rings must be made so they can open and release the product, and however well fitting they may be, a seam mark is made on the glass. Nearly all modern closures require that a perfectly even and smooth finish be given to the rim so that a perfect seal can result. This finish is assured by a short application to the rim of a very hot flame, which causes the surface to flow and obliterate mold marks.

Finishing operations

Articles made on the blowing iron need more extended finishing treatment, as do paste-molded items such as table glasses. Waste material is removed by lightly scoring the surface at the appropriate position. The object is then rotated while being quickly heated by gas jets. This action cracks off the material above the scribed line. A well-directed needle of flame, sometimes preceded by a brief grinding of the edge to square it off, quickly melts the surface around the rim and rids it of the sharp corners that would soon cause it to become chipped if they were allowed to remain.

Light bulbs and tubing. Lamp bulbs can well be regarded as a specialized variety of bottle. Their extreme thinness, however, calls both for special handling methods

and for special techniques to manufacture them economically and in large quantities. A machine can produce light-bulb glasses at rates of up to 33 per second. It is supplied with a continuous ribbon of hot glass, travelling between rollers that preform it with a series of regularly spaced depressions. These glass forms are blown by puffs of compressed air to their final shape and size inside paste molds, which rise around them as the ribbon passes along the conveyor.

Tubing, once produced by a combination of blowing and stretching, is almost all mechanically made today. A hollow mandrel (metal bar), through which air is forced at low pressure, takes the place of the blowing iron. Molten glass is allowed to flow over the downward-pointing mandrel at a rate that maintains an evenly thick coating during the whole process of drawing off. Air is passed through the mandrel so that as the glass is drawn off a tube is formed. A steady pull is maintained by the grip of a pair of moving asbestos-faced belts placed at such a distance from the mandrel that the glass is quite firm and nearly cold by the time it reaches them. Large-diameter tubing and cylinders, which cannot be dealt with in this fashion, are produced by a very similar process. They are drawn off vertically upward over a mandrel consisting of a refractory ring that is positioned on the surface of the molten glass, through which air is passed.

Glass fibres. Once a novelty, glass fibre is now produced in enormous quantities for its thermal and electrical insulating properties and to provide reinforcement of lightweight synthetic resins (plastics) to form boat hulls and car bodies. Filaments as fine as a few thousandths of a millimetre in diameter can be produced. The continuous-filament process is used when the fibres are intended for the making of glass yarn. Molten glass is fed through bushings (removable metal linings) that terminate in very accurately sized spinarets, or forming tips. About 400 such tips may be required to supply the very fine filaments to be united as a single strand of yarn.

The fine streams of glass issuing from the forming dies are drawn off at speeds measured in miles per minute. As they are drawn together, they are thinly coated with an adhesive, which ensures their cohesion as they pass on to the high-speed winders as a united strand. Previously, glass fibres were formed from remelted, uniformly sized glass marbles, but the simplified direct-melt method described is now commonplace and is used to produce many million miles of even quality filaments.

Glass wool

The production of glass wool demands very different treatment. Here, the problem is not to produce an orderly assembly of fibres but a mass of randomly distributed filaments. The system most generally used is the crown process. A steady stream of molten glass from the forehearth is made to fall into a rapidly spinning dish that has hundreds of minute holes in its wall. Fine streams of glass are flung out much as water is forced through the vents of the rotating drum of a spin dryer by centrifugal force. As they issue from the dish, they first meet a blast of air from an air ring that controls their direction downward toward a continuous fine spray of a binding agent. They are then subjected to a further air-jet treatment, the turbulence of which tumbles them into a state of maximum disorder as they fall onto a moving conveyor, on which they form a continuous mat.

Toughened and laminated glass. The rapidity with which cracks travel through glass when it fails under tensile stress was described above. In many applications, such as doors or automobile windshields, failures of this kind would have disastrous results; and neither wired glass nor very thick plate glass can be used. Fortunately, there are two ways—toughening and lamination—of ensuring that glass will not shatter into a mass of lethal fragments on fracture.

It has been known for centuries that a large blob of molten glass dropped into cold water will congeal into a rigid lump that can survive repeated blows with a heavy hammer. It is also known that such blobs (called Prince Rupert's drops after their 17th-century popularizer) immediately collapse into a heap of granules when the tail of the drop is fractured by applying sufficient pressure with a vise or pliers. This condition can be simply explained.

The sudden chilling of the exterior of the drop renders it rigid before the interior has undergone the configurational shrinkage appropriate to its falling temperature. The forces acting to attain greater density place the surface under powerful compressive stress. This compression must be overcome before the surface can be subjected to any tensile stress whatever. It is easiest to do so at the tail end of the drop, at which little material and only relatively minor stress is involved; fracture there immediately releases a considerable amount of stored energy that sends shock waves through the interior, already overstressed in tension, so that it cracks in all directions, causing fragmentation of the entire drop.

Heat toughening of plate glass is carried out by heating the material to just below its softening temperature and suddenly chilling it with jets of air or cold gas. This process produces a condition that allows it to be bent farther, and to be hit about five times harder, than ordinary glass of the same thickness before failure results. An ordinary amount of bending merely goes part way to relieving the compressive stress. When it does break, the sheet disintegrates into small, relatively harmless cubelike fragments that can be seen to have concave edges, resulting from the sudden, long-delayed shrinkage of the interior. All necessary cutting and drilling operations must be carried out before the glass is tempered.

Though toughened glass is widely used in Great Britain for car windshields, it is not acceptable in the United States and some other countries. In certain situations its very toughness can be a disadvantage; it is, for example, very much more rigid than the skulls of the car's occupants. For this reason, laminated glass, made to yield without disastrous results, is often considered to be a more suitable material for windshields.

Laminated glass

Laminated glass suffered a period of eclipse after its introduction because it was originally formed as a sandwich of two sheets of glass bonded to a celluloid interlayer. The celluloid quickly yellowed in some climates, first around the edges, but eventually over a wide zone. Today a very tough interlayer of polyvinyl butyral, which does not suffer from optical aging defects, is used. With flat glass, laminating is a very straightforward process, but modern curved windshields have necessarily complicated matters because the two layers of glass must be perfectly matched. This matching is achieved by placing paired sheets on a suitably shaped frame in an electrically heated furnace in which they are caused to soften and take on their required contours. Next, the thoroughly dried plastic is introduced between the two glass layers in an air-conditioned room the temperature of which does not exceed 60° F (16° C) and which has a maximum relative humidity of 30 percent. Preliminary adhesion is obtained by means of mild electrical heating, and pressure supplied by rubber rollers. The assembly is then placed in an autoclave (closed vessel) in which the air pressure is raised to 50 pounds per square inch (3.5 kilograms per square centimetre) and the temperature to slightly above normal boiling point.

The vinyl layer in ordinary laminated glass, though very thin, is quite tough. Severe impacts will crack the glass but will not damage the interlayer; adhesion is such that splinters cannot fly. The process can also be extended to produce laminates that can resist almost any kind of attack; thicker layers of vinyl or glass (or more layers) can be used to make the material proof against deliberate assault by any means up to small-arms fire. Laminated glass also has the virtue of being a far more effective absorber of sound than unlaminated glass of equal thickness. A pistol shot, even at close range, will not penetrate 3/4-inch- (19-millimetre-) thick laminated glass. Glass of this type is often termed bulletproof.

Glass in electricity and electronics. The uses of glass in electricity and electronics are too numerous to list; some common application problems, however, deserve mention. The lamp globes and tubes described above contain components largely constructed of glass, usually of a type very different from that of the outer shell. The high temperature frequently met in electrical use and the differing thermal expansions of the components involve the use of special glasses at junction points.

In sodium-vapour lamps, an envelope made of ordinary glass would be useless since the intensely active vapour would quickly destroy it. Chemically resistant glass containing large amounts of alkaline-earth oxides and alumina is not only costly to produce but also falls an easy victim to ordinary atmospheric attack. In consequence, sodium-vapour lamps have an envelope of cheap weather-resistant glass with a thin interior layer of the sodium-resistant material.

Glass in insulation

The great dielectric (electrically nonconducting) strength of glass makes it a valuable insulator; a great deal of glass yarn is used in the electrical industries for insulation purposes. High-voltage power lines are usually suspended from toughened glass insulators made up of as many subunits as circumstances demand. Their great mechanical strength withstands the considerable stresses imposed by the heavy cables as they swing in strong winds; failure, when it occurs, takes place not in the glass but in the high-tensile steel units that link the glass components together. Many specially formulated glasses are used in electronics for such common components as insulators, capacitors, and resistors and for many sophisticated devices employed in radar, television, and computer circuitry.

Miniature electronic components are often far too small for their glass parts to be molded by conventional means. In order to overcome this difficulty, thin glass rods are heated until they are sufficiently plastic to be shaped between precision dies. Sintering processes, which heat a substance to coherence without melting, are also used; finely powdered glass mixed with a removable binder is pressed to shape in dies, and then fired. Larger items can be fabricated by mixing the powdered glass with water and shaping it in a plaster mold; it emerges firm but still damp. After being thoroughly dried, it will retain its shape when fired at high temperature.

Heat-resistant glass. When ordinary glass fails as a result of thermal shock, it usually does so on abrupt cooling, because the surface is subjected to tensile stress by the still thermally expanded interior, which cools far more slowly than on the surface. Generally speaking, thick-walled articles fail more often from thermal shock than very thin ones. Microscope slides and cover slips, and even thin table glasses, can survive considerable thermal shock because there is in them so little interior material to apply a stress to the surface. Toughened glass, too, is fairly resistant. In many glass products, such as industrial chemically resistant piping, high-temperature lamps, laboratory glassware, industrial thermometers, and ovenware, toughening after shaping is impracticable. These objects in consequence are fabricated from low-expansion glass, the smaller dimensional changes of which reduce thermally induced surface tensile stresses.

Borosilicate glass expands only about one-third as much as common glass for equal changes of temperature; it also has good chemical resistance and is an excellent electrical insulator. Like ordinary glass, it can be used to produce fibres and fabrics; because of their high melting point, these give excellent protection against fire, as well as being good heat insulators.

The low thermal expansion of Pyrex-type glass gives it a particular value in reflection optics because it reduces the changes of optical figure that arise in concave and convex surfaces as a result of their cross sections being different at different temperatures. The first, and most spectacular, use of low-expansion glass for this purpose was the creation of the 200-inch-diameter mirror for the large reflecting telescope erected on Mt. Palomar in California. Reflecting telescopes of diameters as small as eight inches (20 centimetres) are being made today from Pyrex or other low-expansion material as a matter of course. Although its thermal expansion exceeds that of Pyrex, aluminosilicate glass may be employed in high temperature applications that also require specially high resistance to alkalies.

Shrunken glass

Shrunken glass in its finished form has less thermal expansion than any glass except pure silica. With its very high melting point, it can be used wherever extreme resistance to thermal shock is necessary. Articles are preformed from special borosilicate glass to a size larger than required; the nonsilicate ingredients are then leached out, leaving the material porous. It is then cleansed and heated until the pores vanish and the article assumes its desired dimensions. It is virtually immune to thermal shock and can be used in applications that involve working temperatures of well over 1,800° F (1,000° C).

Tableware. Among all the methods of glass production used at the present time, those employed in the making of fine tableware remain closest to traditional glassmaking techniques. Advantage is taken, however, of the improvement in furnaces and greater knowledge of glass chemistry. The range of colours from metallic oxides in the melt is greater than at any time in the past.

Offhand blowing is still an important process; tools are roughly identical with those of the past. Lead crystal is generally gathered, at temperatures around 1,800° F, from the middle of a fireclay ring, which ensures freedom from any scum or material of anomalous viscosity. The size of the glass gob needed is a matter of judgment, not measurement. The soft glass is then carefully shaped (marvered) on a charred wood block to ensure that it is suitably preformed before blowing, which cannot safely commence until the outer skin is viscous enough to resist expansion sufficiently so that glass flow remains controllable and proper distribution of the material can be assured. Stem and foot are added separately when the globe is firm, but still very hot.

Glassware stems

Stems can be plain or very complex. A hot gob of glass can be squeezed so that a number of evenly separated vertical grooves are formed in it. If these are then folded in they form tubes that remain when the gob is stretched; when the hot stem is twisted, the grooves, too, are twisted, forming a helix of hollow tubes within the stem. In much the same way, the growing stem can be flattened to a square cross section and afterward twisted. Coloured or opaque white-glass canes can be incorporated and twisted to form helices of almost any desired complexity.

Decoration of glass. Engraving processes of fine ware can be carried out on abrasive wheels if the article is substantial; water is used as a coolant and to wash away sludge. The cuts can be of various sections: flat, V-shaped, or curved (concave or convex). The cuts can be buffed to a polish afterward on felt wheels fed with very fine abrasive and water or smoothed with acid.

Fine cuts on thin material and the very complex relief working on sculptured surfaces is carried out by holding the glass against the edges of rotating copper wheels fed with abrasive in an oil medium. Several dozen different sizes of wheels may be used to complete an engraving.

Diamond point stipple engraving is used occasionally; in this technique, minute specks are fractured from the surface. Considerable freedom of treatment is possible by the drypoint-etching technique, consisting of covering the article with an acid-resistant film, scratching through this coating, and then placing the glassware in hydrofluoric acid.

Masking enables broad areas to be etched by acid vapour or by air blasting with fine abrasives. In commercial glassware, considerable use is made of silk-screen printing methods to apply enamels of various colours to the glass; these are afterward rendered permanent by heat. Organic metals, similarly applied, can be used to create a thin metal film by firing the glass.

Diffusion methods

Diffusion methods can also be used, although their primary employment is in the marking of laboratory ware. A copper-silver stain is first applied by screen printing. When the article is fired, the silver and copper ions enter the glass structure to a depth determined by time and temperature control. The printing, being an integral part of the glass substance, is as permanent as the article itself.

Vacuum flasks and jars. The low thermal conductivity of glass does not prevent the contents of jars and bottles from responding as a result of heat transfer by radiation to changes in ambient temperature. Vacuum flasks, often made of low-expansion borosilicate glass, make direct conduction and radiation of heat through the walls of the container impossible. The flask consists of two spaced containers, fused together around their rims. Small pads between the components prevent flexural variations of the space between them during both manufacture and use.

After being united, the air-separated surfaces are silver plated and washed out. After being dried and heated, the flasks are attached to a vacuum pump that exhausts the air from between the walls, and the space is then sealed off.

The silvered surfaces are necessary to prevent losses by radiation across the vacuum space. Heat is simply reflected into the contents and, conversely, to the air surrounding the flask. The only route of heat transfer available is by conduction along the very thin walls of the inner component. This process is so slow that fluids inserted at boiling temperature will still be hot 24 hours later; ice can remain unmelted after two days or more.

PRODUCTION OF SPECIAL GLASS PRODUCTS

Foam glass. A strange but nevertheless useful form of glass that has been called synthetic lava is made in molds from a mixture of ground glass and carbon. When heated, the mixture expands to many times its original volume and sets as a very light, rigid material. Because it is nothing but a mass of thin-walled bubbles, the foam glass can be sawed quite easily; and because the bubbles thus broken and exposed do not communicate with other bubbles, its extreme buoyancy is unaffected. Besides being useful as a buoyant medium, it is employed as insulation, an application in which its rigidity and immunity to insect and atmospheric attack make it superior to natural substances such as cork.

Glass bricks. Entire walls of steel-framed buildings have sometimes been constructed of hollow glass blocks laid with mortar bonding in the same way as ordinary household bricks. Glass bricks are made from molded halves fused together; the partial vacuum created by contraction of the residual trapped air on cooling after sealing makes them highly efficient sound insulators.

Fused spaced window-panes

A similar process is used to unite spaced panes in standard sizes as ready-made double glazing units. The complete and permanent nature of the seal ensures that moisture cannot condense inside the enclosure, makes satisfactory fitting independent of local weather conditions, and greatly simplifies installation problems.

Sulfated glass. The destructive effect of alkaline attack on ordinary glass was described above. Although resistant glasses can be fabricated, their cost or working characteristics can make them impractical in large-scale bottle making. Sulfating is not a glassmaking operation but a form of treatment that consists of changing the constitution of the glass at the surface. The bottles are filled with acid gases, sulfur dioxide and sulfur trioxide. By a process of replacement by diffusion, most of the sodium ions from the surface are removed, with a consequent raising of chemical resistance by factors of up to 100.

Photosensitive glass. By 1943 it had been well established that glass can be rendered photosensitive to ultraviolet light. It was found that the colour of certain types of clear ruby glass can be deepened by exposure to ultraviolet radiation before being subjected to the reheating process that makes them acquire colour. The next step was the making of glass in which no colouring occurs during the reheating process without the preliminary ultraviolet irradiation; any desired pattern can be introduced deep inside the glass itself by interposing a suitably designed ultraviolet–opaque mask between the glass and the irradiating source.

Modern photosensitive glasses contain optical sensitizers, such as univalent cerium ions. Radiation readily removes an electron from these ions; when the glass is heated sufficiently, these can be captured by any gold ions that have been included in the glass, so that gold atoms form and remain in coalescence when the glass is cooled. Unlike the silver halides used in ordinary photographic processes, an image formed by these means is completely grainless because the opaque units are well beyond the threshold of resolution in visible wavelengths.

Photochromic glass. Unlike photosensitive glass, photochromic glass does not depend on heat treatment but on the action of ultraviolet or visible radiation in order to darken it; and unlike ordinary photosensitive materials the process of darkening is completely reversible. The action is due to the inclusion of silver chlorides, bromides, iodides, or their mixture in the glass material; these, in processing, form microscopic crystals thinly dispersed throughout the material. The darkening effect is similar to the action that occurs in ordinary photographic processes but is reversible principally, it is thought, because of the very different sizes of the crystals in the two different materials. A typical single crystal in a photographic emulsion is, by volume, many million times as great as the crystal units in photochromic glass and encourages the neutral silver alone to form stable particles. Additionally, in the impermeable glass material, the neutral chlorine atoms do not diffuse from the zones of reaction as in photo emulsions; the silver and chlorine are not removed by other reactions.

Opal glass. An ultraviolet-sensitive glass containing lithium silicate, which can be converted by heat into an opal glass, has special applications because it then becomes far more readily soluble in hydrofluoric acid. The use of a suitable mask makes it possible to irradiate the glass so that an opal print of almost any desired degree of complexity can be reproduced by heat treatment. This print can then be etched right through the glass with acid. The high resolution obtained makes it possible by this means to "drill" accurately dimensioned, and accurately positioned, holes with distribution densities as high as 320,000 per square inch (50,000 per square centimetre).

Chemically toughened glass. The thermal-toughening process cannot be applied to very thin sections; where these need much more than ordinary flexural strength, chemically toughened glass is used. One method employed involves an ion-exchange process that produces a change in the chemical composition of the surface layer. If a glass that is rich in lithium ions is immersed in molten sodium chloride, many of the lithium ions escape and are replaced on a one-to-one basis by far larger sodium ions. This substitution results in considerable compressive stress being applied to the surface when the glass cools. Flexural strength of up to 98,000 pounds per square inch (6,900 kilograms per square centimetre) can be achieved by this treatment, much more than is possible by thermal toughening. Chemically toughened glass can actually be used to make springs.

Titanized glass. In common with many other forms of glass, bottles do not lend themselves to thermal toughening; maximum strength must be obtained by designing their shapes to withstand the impacts that inevitably result from mechanical handling. They can also be protected to some extent by suitable chemical treatment. The titanizing process, developed in Great Britain, consists of spraying bottles with a titanium compound before annealing, while the bottles are still red hot. Part of the compound volatilizes, leaving a very thin surface layer of titanium oxide, which is soluble in glass. The resulting increase in the working life of returnable bottles, which have to withstand being transported and passing through cleaning and filling plants, is due to the reduction of friction and resistance to the formation of tiny cracks in the treated glass.

Optical fibres. The phenomenon of total internal reflection described at the beginning of this article allows light to be piped around bends through the interior of a glass rod. As long as the ratio of the rod diameter and radius of curvature of the bend is not such as to allow internally reflected rays to arrive at another point of the interior surface at less than the critical angle, only absorption losses will occur.

Glass fibres are essentially very thin rods. It follows that a quarter-inch- (six-millimetre-) diameter bundle of fibres can convey light around a much sharper curve than a single rod of the same diameter. If the individual fibres are sufficiently fine and coherently arranged, the opposite ends of the bundle can be ground and polished. If one end is placed in contact with, or very close to, an object, the image of that object will appear at the distant face. Flexiscopes made from bundles of optical fibres for the investigation of otherwise inaccessible sites are used in medicine and industry. Very small bundles are called optical catheters.

The diameter of individual fibres must be extremely small for good resolution; very fine fibres are hard to handle. The task is simplified by first drawing together a bundle of

Bundling optical fibres

coarser fibres, embedded for protection in a sheath of matrix glass. This group is then drawn together, like millefiori glass (described above) and can be stretched until it is of sufficiently small diameter, normally a few microns. The multiple fibres can be wound on a drum and afterward used to make up coherent bundles of fibrescopes.

Short bundles of fibres can be carefully stacked and fused together as a mass. The resulting slab or sheet has very different properties from those of ordinary glass. When its surfaces are ground and polished, light falling on a point at one surface will be transmitted without deviation to a corresponding point on the distant surface. Used as a faceplate on a cathode-ray tube it acts as an imaging screen that, because it directs the light as a bundle of parallel units, overcomes the difficulties normally arising because of light scatter when photographing the display.

Electrically conducting glass. Aircraft windshields, especially on supersonic aircraft, must withstand environmental conditions very different from those in automobiles. Air friction can raise the surface temperature of the glass to about 310° F (150° C) at low-level high-speed flight; and the ambient temperature may be as low as −200° F (−130° C) at some altitudes.

Laminated toughened glass with interlayers of vinyl up to one-quarter inch (six millimetres) thick is used; in military aircraft, four toughened glass sheets and three interlayers of plastic bonded together give a large degree of protection against bullets and flak splinters. A plating of gold, thin enough to be transparent, makes it possible to pass an electric current through the assembly to warm the glass and prevent ice formation. Very thin wires, less than $1/1{,}000$ inch (0.03 millimetre) thick, can also be embedded for this purpose.

Optical windows. When photographing objects in supersonic wind tunnels, vacuum chambers, or any site into which the investigator or his instruments cannot easily be introduced, the camera is aimed through a special window fitted with glass the surfaces of which are exactly parallel, thus permitting undistorted photography. Even large optical windows are unlikely to deviate from true flatness by more than one-half wavelength of sodium light. For use in applications for which protection against X-radiation and gamma radiation is required, very thick windows of extremely dense flint glass containing up to 80 percent of lead oxide are employed.

Pure silica glass. Ordinary glass is useless for ultraviolet transmission; large aluminized optical mirrors are necessary for ultraviolet spectroscopy in astronomy. Silica itself, however, is transparent to ultraviolet, although its great viscosity even at very high temperatures makes it unsuitable for molding, by ordinary means, of optical lens blanks.

The contemporary method of making 99.5 percent silica glass (fused quartz) satisfies most applications. Even purer silica, transparent to ultraviolet and near infrared radiation as well as to the entire visible spectrum, can be produced by special techniques; the glass composition must be described as 100 percent silica because the impurities amount to no more than one or two parts per 100,000,000. (Fr.J.M.)

Brick and tile production

The brick, first produced in a sun-dried form at least 6,000 years ago, and the forerunner of a wide range of clay building products used today, is a small building unit in the form of a rectangular block, formed from clay or shale or mixtures and burned (fired) in a kiln, or oven, to produce strength, hardness, and heat resistance. The original concept of ancient brickmakers was that the unit should not be larger than what one man could easily handle; today, brick size varies from country to country, and every nation's brickmaking industry produces a range of sizes that may run into the hundreds. The majority of bricks for most construction purposes have dimensions of approximately $2\frac{1}{4} \times 3\frac{3}{4} \times 8$ inches ($5.5 \times 9.5 \times 20$ centimetres).

Structural clay tile, also called terra-cotta, is a larger building unit, containing many hollow spaces (cells), and is used mainly as backup for brick facing or for plastered partitions.

Structural clay-facing tile is often glazed for use as an exposed finish. Wall and floor tile is a thin material of fired clay with a natural or glazed finish. Quarry tile is a dense pressed fired-clay product for floors, patios, and industrial installations in which great resistance to abrasion or acids is required.

Firebrick is used in incinerators, boilers, industrial and home furnaces, and fireplaces. Sewer pipe is fired and glazed for use in sewage-disposal systems, industrial-waste systems, and general drainage. Drain tile is porous, round, and sometimes perforated and is used mainly for agricultural drainage. Roofing tile is made in the form of half-round (Spanish tile) and various flat tiles made to resemble slate or cedar shakes; it is used extensively in the Mediterranean countries.

There are also many products made from cement and aggregates that substitute for, and generally perform the same functions as, the clay products listed above.

HISTORY

Mud brick, dried in the sun, was one of the first building materials. It is conceivable that on the Nile, Euphrates, or Tigris rivers, following floods, the deposited mud or silt cracked and formed cakes that could be shaped into crude building units to build huts for protection from the weather. In the ancient city of Ur, in Mesopotamia (modern Iraq), the first true arch of sun-baked brick was made about 4000 BC. The arch itself has not survived, but a description of it includes the first known reference to mortars other than mud. A bitumen slime was used to bind the bricks together.

Burned brick, no doubt, had already been produced simply by containing a fire with mud bricks. In Ur the potters discovered the principle of the closed kiln, in which heat could be controlled. The ziggurat at Ur is an example of early monumental brickwork perhaps built of sun-dried brick; the steps were replaced after 2,500 years (about 1500 BC) by burned brick.

As civilization spread eastward and westward from the Middle East, so did the manufacture and use of brick. The Great Wall of China (210 BC) was built of both burned and sun-dried bricks. Early examples of brickwork in Rome were the reconstruction of the Pantheon (AD 123) with an unprecedented brick and concrete dome, 142 feet (43 metres) in diameter and height, and the Baths of Hadrian, where pillars of terra-cotta were used to support floors heated by roaring fires.

Enamelling, or glazing, of brick and tile was known to the Babylonians and Assyrians as early as 600 BC, again stemming from the potter's art. The great mosques of Jerusalem (Dome of the Rock), Isfahan, and Teheran are excellent examples of glazed tile used as mosaics. Some of the blues found in these glazes cannot be reproduced by present manufacturing processes.

Western Europe probably exploited brick as a building and architectural unit more than any other area in the world. It was particularly important in combatting the disastrous fires that chronically affected medieval cities. After the Great Fire of 1666, London changed from being a city of wood and became one of brick, solely to gain protection from fire.

Bricks and brick construction were taken to the New World by the earliest European settlers. The Coptic descendants of the ancient Egyptians on the upper Nile called their technique of making mud brick *tōbe*. The Arabs transmitted the name to the Spaniards, who, in turn, brought the art of adobe brickmaking to the southern portion of North America. In the north the Dutch West India Company built the first brick building on Manhattan Island in 1633.

MANUFACTURING

Basically, the process of making brick has not changed since the first fired bricks were produced some thousands of years ago. The steps used then are used today, but with refinements. The various phases of manufacture are as follows: securing the clay, preparation, mixing and forming, drying, firing, and cooling.

Clays used today are more varied than those used by the

first brickmakers. Digging, mining, and various methods of grinding enable the modern manufacturer to utilize many raw materials.

Classes of clays used

Clays used in brickmaking represent a wide range of materials that include varying percentages of silica and alumina. They may be grouped in three classes: (1) surface clays found near or on the surface of the Earth, typically in river bottoms; (2) shales, clays subjected to high geologic pressures and varying in hardness from a slate to a form of partially decomposed rock; and (3) fireclays, found deeper under the surface and requiring mining. Fireclays have a more uniform chemical composition than surface clays or shale.

Securing (winning) the clay. Surface clays are typically recovered by means of power shovels, bulldozers with scraper blades, and dragline operations. Shales are recovered by blasting and power shovels. Fireclays are mined by conventional techniques.

Preparation. Raw clays are often blended to obtain a more uniform consistency. In many cases the material is ground to reduce large rocks or clumps of clay to usable size and is placed in storage sheds. As additional material is stored, samples are blended from a cross section of the storage pile. The material is then transferred to secondary grinders and screens (if necessary) to secure the optimum particle size for mixing with water. In certain processes (*e.g.,* soft-mud), the clay is transferred directly to the mixing area, eliminating all grinding, screening, and blending.

Mixing and forming. All clays must be mixed with water to form the finished product. The amount of water added will depend on the nature of the clays and their plasticity. This water is removed during drying and firing, which causes shrinkage of the units; to compensate for this shrinkage the molds are made larger than the desired finished products. Three basic processes are used in the forming and mixing phase. In the stiff-mud process, the clay is mixed with water to render it plastic, after which it is forced through a die that extrudes a column of clay like the toothpaste squeezed from a tube. The column gives two dimensions of the unit being manufactured; it is cut to give the third dimension. All structural clay tile is made by this process, as is a great percentage of brick. In the older method of forming bricks, the soft-mud process, much more water is used, and the mix is placed in wooden molds to form the size unit desired. To keep the clay from sticking, the molds are lubricated with sand or water; after they are filled, excess clay is struck from the top of the mold. It is from this process that the terms wood-mold, sand-struck, or water-struck brick were derived. Clays with very low plasticity are used in the dry-press process. A minimum of water is added, the material is placed in steel molds, and pressures up to 1,500 pounds per square inch are applied.

Drying. After the bricks are formed, they must be dried to remove as much free water as possible (they could literally explode if subjected to fire without drying). Drying, apart from sun drying, is done in drier kilns with controlled temperature, draft, and humidity.

Firing and cooling. Bricks are fired and cooled in a kiln, an oven-type chamber capable of producing temperatures of 1,600° to more than 2,000° F (870° to 1,100° C), depending on the type of raw material. There are two general types of kilns, periodic and continuous.

The earliest type of kiln, the scove, is merely a pile of dried bricks with tunnels at the bottom allowing heat from fires to pass through and upward in the pile of bricks. The walls and top are plastered with a mixture of sand, clay, and water to retain the heat; at the top the bricks are placed close together and vented for circulation to pull the heat up through the brick. The clamp kiln is an improvement over the scove kiln in that the exterior walls are permanent, with openings at the bottom to permit firing of the tunnels.

A further refinement of the scove kiln, round or rectangular in form, is designated as updraft or downdraft, indicating the direction of heat flow. In these kilns the walls and crown are permanent, and there are firing ports around the exterior.

In so-called periodic kilns, the bricks are placed with sufficient air space to allow the heat from the fires to reach all surfaces. They are placed directly from the drier, and heat is gradually increased until the optimum firing temperature is reached. When they are sufficiently fired, the heat is reduced, and they are allowed to cool gradually before removal from the kiln.

Series of kilns

The periodic kiln was improved in efficiency by placing several kilns in line with connecting passages. The first chamber is fired first and the excess heat passed to the next chamber to start heating. Successively, the various chambers are brought to optimum firing and cooling temperatures, until all bricks have been fired and cooled. This arrangement is known as the moving fire zone. In the more modern fixed fire zone, dried bricks are placed on cars carrying as many as 3,000 or more bricks; the cars start at the cool end of a long tunnel kiln and move slowly forward through gradually increasing temperatures to the firing zone, pass through it, and emerge through decreasing heat zones until cooled.

Automation. Since the development of the tunnel kiln, brickmakers have sought to increase automation in their plants. Handling of the finished product has been automated to the point that bricks emerging from the kiln are now automatically stacked in packages of approximately 500, strapped with metal bands, and stored, shipped, and delivered by mechanical equipment.

In some plants bricks are taken from the cutter machine, placed in the drier or on drier cars by mechanical means, placed on kiln cars by mechanical fingers, removed from the kiln cars mechanically, stacked, strapped, and prepared for shipment without being touched by hand.

COLOURS AND TEXTURES

Colour of clay products may be natural or applied.

Natural colours. These depend on the type of clay used in the production processes. They range from whites through grays, buffs, light to dark reds, and into the purple range. Fireclays are associated with the lighter colours such as the grays and buffs. Ordinary clays and shales are associated with the red ranges. By regulating the oxidizing conditions in the kiln, browns, purples, and blacks can be obtained. The process is known as flashing, and in general the change of colour of the bricks is only on the surface, the body of the unit retaining its natural colour. Some metals, such as manganese, are mixed with the clays to develop special colours.

Applied colours. Colours are applied to many brick products, particularly structural glazed tile, wall and floor tile, and brick. Ceramic glazes are applied to units before or after the firing and cooling stage. If after, the units must be refired. These glazes provide almost all of the basic colours plus some special colours used for "accent" in the design of a wall. The glazes become an integral part of the face of the units since they are burned to the same degree of heat as the units. Finishes and colours other than ceramic glazes are applied to the units either fired or unfired and are surface coatings that conceal the natural colour of the burned unit. In some countries a demand for old brick has led to application of mixtures of cement or lime and sand and many other combinations to give brick an aged appearance.

Types of finishes

Texture. The texture of clay products is directly associated with the manufacturing processes. The soft-mud process produces either a sand- or water-struck finish in a non-uniform texture, which gives the brick (only bricks are made under this process) the appearance of handmade or antique brick. The dry-press process, using steel molds, gives a smooth texture only. This process is seldom used in modern-day brick production but is used in the manufacture of quarry tile as well as floor and wall tile. The stiff-mud process offers the most possibilities for texturing brick. As the prepared clay is extruded through the die, the pressure produces a smooth surface similar to that of concrete when smoothed with a steel trowel. This surface is called the die skin; its removal and further treatment produce other textures.

In wire cutting, a wire placed in front of the column of clay as it comes from the die removes the die skin, creating a semi-rough surface.

In sand finishing, sand is applied to the column of clay by various means to give a very even surface of sand, which is fired into the unit. The desired texture is similar to a wood-mold brick except that the unit is much more uniform in size and in finish. Colour also may be changed by the type of sand used.

Scored finishing is used mostly on tile where the surface of the tile is grooved to give a better bond between the unit and plaster. This is also true of a roughened or combed finish produced by wire brushing or scratching.

Roughened finishing is used when the die skin is removed by various means. In one method, the material cut in removing the die skin may be rolled back into the face of the unit.

Other finishes are applied by rollers on the column to give certain effects such as bark, log, or emblems. Terra-cotta for architectural decoration is both machine extruded and handmade (molded or pressed). It is distinguished from other clay products by the generally larger size of the units. It may be hand carved and used mostly in murals as bas-relief. Both natural and glazed finishes are produced.

USES OF BRICK

By far the largest use of brick products is, as it always has been, in building construction. Many significant changes have taken place in building with brick.

Building construction. It may be roughly accurate to say that about 65 percent of all the brick in the world goes into dwellings, and 35 percent goes into commercial, industrial, and institutional buildings. Construction techniques change yearly and from country to country, but basically most brick and tile are used in walls, with lesser use in roofs and floors.

Walls may be classified in three general categories: load-bearing, non-load-bearing, and veneer. A load-bearing wall supports the loads of a structure, such as floors, equipment, furniture, and people. At one time, buildings were constructed with very thick brick walls carrying all floor and other loads. Design of these walls was not based on engineering data but only on well-intentioned but unscientific building codes. As buildings grew taller, the building code requirements for thickness of a brick wall became economically prohibitive. The last truly high-rise, load-bearing brick structure built under older codes was the Monadnock Building in Chicago (1889–91), 16 stories tall with the brick walls 6 feet (2 metres) thick at the base, tapering to 12 inches (30 centimetres) at the top story. The arrival of structural steel on the building scene put a temporary end to the brick bearing-wall skyscraper, but research in the 20th century has led to a resurgence. Thinner walls can be designed for high-rise buildings and built safely at a reasonable cost. Apartment buildings in Switzerland, Germany, Denmark, England, and other countries have risen 15 or more stories supported by brick bearing walls no more than 12 inches (30 centimetres) thick. The use of reinforced brickwork (a combination of brick, reinforcing steel, mortar, and cement grout) permits even thinner walls; a 17-story building designed to resist earthquake shocks uses only 11-inch walls.

Types of bearing walls

Bearing walls may be classified into five general groups: (1) brick, including brick tied together with cross brick (headers) or with metal ties; (2) composite walls of brick and tile tied together with headers or metal ties; (3) cavity walls, in which the inner and outer wythes (tiers) of units are tied together with metal ties but separated by an air space usually two or more inches in width; (4) reinforced walls, similar to cavity walls except that steel is placed in the cavity and the cavity filled with a soupy mortar (grout); (5) single unit walls, using a unit of necessary thickness to meet design requirements.

Non-load-bearing walls carry only their own weight and may be any one of the types discussed under load-bearing walls. This type of wall is used to close in a steel or concrete frame building. It is usually carried by supports, normally steel shelf angles at each floor, and is called a panel wall. When the wall is supported at the base only, it is called a curtain wall.

Veneer walls are similar to non-load-bearing walls in that they carry no weight except their own. The brick or tile is fastened to a backing, but it does not exert a common action with the backing. Perhaps the most common use is brick veneer on wood frame dwellings. Other examples are architectural terra-cotta and thin ceramic veneer on monumental buildings.

Tile roofs are popular in the Mediterranean area and in the Low Countries of western Europe. In Italy, craftsmen have developed an art of using relatively thin tile to form self-supporting arches. Tile roofs in many other areas, particularly on residences, have been used extensively in the past, but economic considerations limit their use now; in addition to the cost of the tile is the cost of roof framing to support the heavier weight of the tile.

Miscellaneous uses in building construction include retaining walls, brick floors, patios, and walks. Most of these uses are decorative as well as utilitarian. The retaining wall of reinforced brick provides an economical means of restraining earth movement and at the same time maintains a continuity of architectural effect, particularly if the adjoining structure is built of brick.

Brick floors, patios, and walks utilize the physical properties of brick, such as resistance to abrasion and to the elements. Paving brick, per se, is practically nonexistent, except for replacement where roads and streets were brick paved long ago. Industrial floor brick, however, supplies many industries whose manufacturing and handling processes require floors that resist acids and provide a high degree of resistance to abrasion. Brick floors and patios, beside providing a long-lasting, low-maintenance material, offer the designer a medium for developing architectural effects in both colours and patterns.

Clay drainage products. Sewer pipe plays an important part in the world's ecology. An almost impervious material because of its firing, denseness, and glazing, it can carry highly corrosive waste materials that few other products can handle economically.

Drain tile performs a service that ensures a higher yield in farm production of food throughout the world. Many farming areas are plagued with too much water at the wrong times. Drain tile reduces the water level during these times, thereby allowing the root growth of plants to penetrate deeper into the soil, which, in turn, permits them later to resist hot dry weather.

Art in clay products. The walls of ancient Babylon were decorated work, and brick has served decorative and artistic purposes ever since. Terra-cotta, molded, formed, and glazed, has been used for years in murals. Many architects use patterns of brickwork to relieve the monotony of a large expanse of plain wall, while others deliberately use the starkness of large plain walls to emphasize form.

NONCLAY BRICK AND TILE

Following the introduction of portland cement in the 19th century, a growing number of products have appeared that resemble clay products in size and intended use. In some countries the production of them, if reduced to brick equivalents, exceeds that of clay products. A brief review of these products, the material used, and the manufacturing processes may serve to suggest the interrelation between these and clay products.

Concrete block is a large unit, usually 8 inches high, 16 inches long, and of various thicknesses, made from a mixture of cement and an aggregate, which may be cinders, limestone, or expanded clay or shale (burned in a rotary kiln). The mixture of cement, aggregate, and a minimum of water is placed in steel molds and vibrated to compact the mixture. The formed units are removed from the molds and cured either in air, steam, or under autoclaving processes (steam under pressure).

Concrete brick is a mixture of cement and aggregate, usually sand, formed in molds and cured. Certain mineral colours are added to produce a concrete brick resembling clay. Concrete pipe is made of cement and aggregate and cured as above. Used as a substitute for clay sewer pipe, it does not have as much resistance to the corrosive action of certain acids. Concrete drain tile and concrete roofing tile are produced similarly.

Sand–lime brick is a product that uses lime instead of cement. It is usually a white brick made of lime and selected

sands, cast in molds and cured. Production is limited, with greater use in the United States and Germany. (J.A.Le.)

Abrasives

Abrasives are sharp, hard materials used to wear away the surface of softer, less resistant materials. Included within the term are both natural and manufactured substances ranging from the relatively soft particles used in household cleansers and jeweller's polish to the hardest known material, the diamond. Abrasives are indispensable to the manufacture of nearly every product made today.

Abrasives are used in the form of grinding wheels, coated abrasives (sandpaper), honing stones, polishes, cut-off wheels, tumbling and vibratory mass-finishing media, sandblasting, pulpstones, ball mills, and still other tools and products. Only through the use of abrasives is industry able to produce the highly precise components and ultrasmooth surfaces required in the manufacture of automobiles, airplanes, and space vehicles, mechanical and electrical appliances, and machine tools.

HISTORY

The use of abrasives goes back to earliest man's rubbing of one hard stone against another to shape a weapon or a tool. The Bible mentions a stone called *shamir* that was very probably emery, a natural abrasive still in use today. Ancient Egyptian drawings show abrasives being used to polish jewelry and vases. A statue of a Scythian slave, called "The Grinder," in the Uffizi art gallery in Florence, shows an irregularly shaped natural sharpening stone used to whet a knife.

Sand and pieces of flexible hide were early man's sandpaper. Later, craftsmen tried to fix abrasive grains to flexible backings with crude adhesives. A 13th-century Chinese document describes the use of natural gums to fix bits of seashell to parchment. About two centuries later, the Swiss began coating crushed glass on a paper backing.

Early sand and glass abrasives lacked sharpness, and by the 19th century early abrasive products like the natural sandstone that man had formed into the "grinding wheel" no longer met the needs of developing industry. In 1873 Swen Pulson, working in the Norton and Hancock Pottery Company, Worcester, Massachusetts, won a jug of beer by betting that he could make a grinding wheel by combining emery with potter's clay and firing them in a kiln. Pulson succeeded on his third try; this incident signalled the end of unsatisfactory glue-and-silicate bonded products and the birth of the vitrified grinding wheel.

Just before the beginning of the 20th century, when nature's abrasives emery, corundum, and garnet were falling short of industry's demands, a process was developed to make aluminum oxide and silicon carbide with the electric furnace. In 1955 the General Electric Company succeeded in manufacturing diamond. Like other man-made abrasives, synthesized diamond proved superior in many applications to the natural product, which had been used in grinding wheels since 1930.

First synthetic diamonds

Once used only when precise dimensional accuracy and smooth surfaces were necessary, abrasives have become a widely applied industrial tool. Higher grinding-wheel speeds, more powerful grinding machines, and improved abrasives have steadily augmented their role.

NATURAL ABRASIVES

Most natural abrasives have been replaced by manufactured or synthetic abrasives because nearly all industrial applications demand consistent properties. With the exception of natural diamond, most of nature's abrasives are too variable in their properties. Natural abrasives, however, are still important.

Diamond. Diamond is the hardest abrasive and, in fact, the hardest material known. India was the main source of diamonds until the 1860s when they were discovered in South Africa, which became the largest supplier of mined diamonds worldwide. Lesser quantities are found in Brazil, Australia, Guyana, and Venezuela.

Stones not suitable for use as gems (called bort or boart) are designated for industrial purposes. The portion used as abrasive is crushed into various sizes for grinding wheels, lapping (fine polishing) powders, coated abrasive belts, and polishing disks. Industrial diamonds vary in colour from off-white to black.

The shape of a diamond particle is important to its performance. Natural diamonds are sorted for shape. Blocky shapes are suited for more severe grinding operations such as cutting concrete building blocks. More splintery shapes are reserved for grinding the cemented tungsten-carbide cutting tools used in industry.

Corundum. Corundum is a naturally occurring form of aluminum oxide, found mainly in South Africa. Other deposits have been found in India, Canada, and the southern part of the United States. Corundum varies in its properties from location to location and even from within the same location, making it unsuitable for use in grinding wheels. Its principal utility is in polishing and glass grinding.

Emery. Also a natural form of aluminum oxide, emery is generally more impure than corundum. It consists of smaller aluminum oxide crystals imbedded in an iron oxide matrix. Principal sources are Turkey and Greece. Its rather round grain shape makes it more suitable for polishing than for grinding. Along with corundum, emery was an important early abrasive and is still used in emery cloth sandpaper.

Garnet. Far superior to flint, and still valuable for coated abrasive products in the woodworking, leather, and shoe industries, garnet possesses considerable toughness. When its cutting crystal edges break down, garnet fractures sharply to present new cutting edges.

Of the seven known forms of garnet, the best for abrasive use comes from the Adirondack Mountains in New York and is mined by a process similar to panning for gold.

Flint. Flint, or flint quartz, is the most commonly used abrasive in the manufacture of coated abrasive paper (sandpaper). Though flint is found all over the world, not all varieties are suitable as abrasives. The best type varies in colour from a dull grayish white to a faint pink and comes from the eastern United States, where it is mined from open pits. Lumps, varying in size from four to 10 inches (10 to 25 centimetres) in diameter, are crushed into sharp crystals. Flint lacks durability and hardness and loses its abrasive cutting action rather quickly.

Quartz. Quartz is the main ingredient of sandstone, which was for years hewn from natural deposits and formed into the first grinding wheels. Until recently these were used in the glass, cutlery, and woodworking edge-tool industries. The best stones came from Ohio and West Virginia. Although still in use where production rate is not a factor, as on a farm, they have been replaced by faster, more efficient manufactured grinding wheels and equipment.

In other forms, quartz is found in dense stones in the United States and in Belgium. These are used as honing and sharpening stones for putting the final, keen edge on cutlery and various cutting tools and are used for honing machined parts and products to accurate size and smooth finish.

Quartz in the form of sand is used as an inexpensive, disposable abrasive for sandblasting buildings and metal structures, such as bridges, when it is impractical to recover the abrasive. In a crushed or grain form, it is used in buffing and polishing wheels.

Pumice. Volcanic in origin, pumice is the frothy part of lava. The best quality, in lump form, is mined in Italy. Its mild action is utilized in metal polishing, glass cutting, furniture finishing, scouring powders, and soaps.

Tripoli. Tripoli (or diatomite), an opaline silica consisting of the fossilized remains of microscopic plants, is also used in scouring powders and soaps, in grease sticks for buffing, and as a polish ingredient. The principal United States sources are Missouri and Oklahoma. Other sources are Iceland and Kenya.

Other natural abrasives. Talc is the softest and mildest abrasive. It was used in the past to polish grains of rice. In addition to sandstone, other varieties of silica, such as rottenstone, which comes from the weathering of impure limestones, and diatomite, are used in polishes. Calcined

(roasted) diatomite is an effective polishing agent for use on nickel plate. Others include feldspar, chert, and chalk.

MANUFACTURED ABRASIVES

By the 1970s, manufactured abrasives represented the larger share of world production and use. The two most important manufactured abrasives, silicon carbide and aluminum oxide, were developed in the late 19th century by use of the electric furnace. These proved to have the consistent quality required by industry. Their principal value is for use in grinding wheels, coated abrasive products, and polishing and buffing wheels. Because large amounts of electric power are required in their manufacture, electric furnace abrasives are made where power is cheap and plentiful.

Silicon carbide. Silicon carbide was first produced by Edward G. Acheson, an electrical engineer in Monongahela City, Pennsylvania. In 1891 he combined a mixture of clay and powdered coke in a small electric furnace. The resulting hard, shiny crystals were able to scratch glass and were first used for polishing precious stones. Initially, the price of silicon carbide was nearly equal to that of the gemstones it polished.

In the manufacture of silicon carbide, a mixture of pure silica sand and carbon in the form of finely ground coke is built up around a carbon electrical conductor within a brick furnace. Electrical current is passed through the carbon conductor, generating heat and bringing about a chemical reaction in which the carbon in the coke and silicon in the sand combine to form silicon carbide. A little sawdust is added to the mixture. The sawdust burns, making the mixture porous and permitting the escape of great quantities of inflammable gas.

At the completion of the furnace run, the product consists of a core of silicon carbide crystals loosely knitted together, surrounded by partially or entirely unconverted raw material. The quality of the abrasive produced depends largely upon the choice of raw materials and the accuracy of the control and operation of the furnace.

Several companies produce silicon carbide under such trade names as Carborundum, Crystolon, and Carbolon.

Silicon carbide manufacture

The method used to manufacture silicon carbide today is basically the same as that developed by Acheson. The major changes are in the size of the furnace and control of the ingredients, techniques, and the understanding of the process. A furnace run takes about 36 hours, during which temperatures vary from 4,000° to 4,500° F (2,200° to 2,480° C) in the core and drop to about 2,500° F (1,370° C) at the outer edge. The furnaces are the shape of a large trough with electrodes at either end. The energy consumption exceeds 100,000 kilowatt-hours per run. Furnaces are about 40 feet long, 10 feet high, and 10 feet wide and can produce about 25,000 pounds of silicon carbide per run.

Aluminum oxide. Manufacture of aluminum oxide abrasive closely followed the development of silicon carbide. Although there are records of previous experiments to produce emery and corundum abrasives, the first successful attempts were made in the laboratories of the Ampere Electro-Chemical Company at Ampere, New Jersey, in 1897. Rock bauxite, the main ingredient of which is aluminum oxide, was used as the raw material.

The results were sufficient to arouse interest. Several hundred tons of the new abrasive were distributed throughout industry for testing, and the outcome of this testing was the construction of a plant at Niagara Falls in which the manufacture of crystalline aluminum oxide on a commercial scale was begun in 1901.

Bauxite is a claylike substance. Found near the surface, it is one of Earth's most plentiful minerals. It is also the source of aluminum. Principal mines are in Suriname, Guyana, and the United States. Bauxite consists of aluminum oxide, Al_2O_3, combined with water and associated with varying quantities of other impurities. By heating the ore to drive off the combined water before the material goes into the electric furnace (an operation called calcining), the percentage of aluminum oxide is greatly increased over its content in the natural state.

Arc-type electric furnaces are used for manufacturing aluminum oxide. Two carbon electrodes are suspended in the furnace. The furnace is charged with the crude ore. Current passes from one electrode to the mass, passes through it, then arcs back to the other electrode. This produces an intense heat of more than 2,000° F, which melts the bauxite and reduces the major portion of the impurities to their base metals. Small percentages of coke and iron borings combine with the impurities, thus helping to settle them to the bottom of the mass. During the furnace run of approximately 16 hours, fresh bauxite charges are continually added.

At the end of the furnace run, the large steel jacket or outer shell of the furnace is lifted, and the hot chunk of aluminum oxide is allowed to cool. Cooling time varies from one day to one week, depending on the type of crystal structure desired.

Both the chemical composition and the crystalline structure of the abrasive are important in the control and variation of the abrasive properties. A wide variation is possible; this characteristic has been utilized in tailoring aluminous abrasives to specific industrial applications. The types vary in colour from one that is pure, snow-white, of large crystals, and friable—used for the grinding of heat-sensitive metals—to the nearly black, finely crystalline, less pure kinds used for heavy-duty operations in foundries and steel mills.

Recent developments include alloy abrasives, made by adding, for example, zirconium oxide to aluminum oxide and the spinel (a hard crystalline mineral) and aluminum oxide combinations. Produced by a number of companies under a variety of trade names, aluminum oxide is the most important abrasive used.

Synthetic diamond. Synthetic diamonds, initially more expensive than natural diamonds, have proved to be more efficient in the grinding wheels used for manufacturing and sharpening cutting tools made of cemented tungsten carbide. Because of the more irregular surface of the synthesized diamond, the bond between the grinding wheel and the diamond particles lasts longer.

Diamond is now manufactured in a variety of shapes and sizes for different types of grinding operations. By changes in the process, the diamond can be manufactured in shapes suitable for cutting ceramics and glass.

Metallic coating on the surface of the manufactured diamond also improves bonding. Nickel and copper have proved to be the most successful coatings. Both are heat conductors and serve to dissipate the heat of grinding and, thus, act as a bonding interlayer, which improves bonding strength.

Boron carbide. A product of the electric furnace, boron carbide is harder than silicon carbide but is not suitable for use in grinding wheels or coated abrasive products. It does have some utility as a lapping (fine polishing) abrasive, substituting for diamond in the polishing of hard materials. It is made from the combination of coke and boric acid at very high temperature. In solid form, boron carbide becomes extremely wear resistant.

Boron nitride. Cubic boron nitride was first synthesized by the General Electric Company. Its hardness is reported to be twice that of silicon carbide and two and one-half times that of aluminum oxide. High-pressure techniques are used in making the material, which has strong crystals and sharp points with clearly defined edges. Despite its hardness, the initial forms of boron nitride did not have the mechanical strength necessary for grinding.

Cubic boron nitride with a bonded nickel coating is a useful abrasive. Although more expensive than industrial diamond abrasive, its grinding efficiency on certain tool-steel applications has made it economically feasible to use. Continued experiments in its manufacture may result in lower costs for this newest of manufactured abrasives.

Other manufactured abrasives. Other manufactured materials employed as abrasives are neither as hard nor as abrasive on work surfaces as those listed above. Glass beads manufactured specifically for use in blasting machines, utilizing air pressure to propel them against a workpiece, remove rust and scale. Basically a cleaning abrasive, they do not materially alter the dimensions of the workpiece.

Metal shot made of bronze, iron, steel, and steel alloy

is used for shot blasting and cleaning in foundries, and by forging companies to clean castings and parts and to alter surface finishes. Shot is also used for relieving internal stresses.

Steel wool, made by combing the wire that forms steel fibres (with equipment similar to that used in carding wool), is used to clean and finish wood and many other materials. The size of the steel fibres is changed by varying the size of the teeth in the card-clothing, the more abrasive product being the one with coarser wires.

PROPERTIES OF ABRASIVES

One of the most important properties necessary in an abrasive material is hardness. The abrasive must be harder than the material it is to grind, polish, or remove. Hardness of the various abrasive materials is measured on either the Mohs scale or the Knoop hardness scale. The Mohs scale, first described in 1812, measures resistance to indentation as judged by which material will scratch another. This scale, which assigns numbers to natural minerals, has been widely accepted and is used by mineralogists. The Knoop hardness test employs an elongated diamond-shaped indenting device and measures the indentation it makes in a given test material. With the Knoop test, the hardness of extremely brittle materials including glass and even diamonds can be measured without harming either the indenter or test piece.

Measuring hardness

Toughness or body strength characteristics are significant to abrasive function. Ideally, a single abrasive particle resharpens itself by the breakdown of its dull cutting or working edge, which exposes another cutting edge within the same particle. In manufactured abrasives it is possible to achieve some degree of control over this property by varying grain shape during the crushing or sizing operation, by making changes in the purity of the abrasive, by alloying abrasives, and by controlling the crystal structure within abrasive grains. Thus abrasives can be developed to meet the operating conditions found in a variety of applications.

Interaction between the abrasive and the material being ground prevents the use of one abrasive as a universal medium. When silicon carbide is used on steel, or aluminum oxide on glass, there is some reaction, which has yet to be clearly defined. Regardless, the effect is rapid dulling and inefficient abrasive action. Attrition resistance is the name given to this third, very significant property.

FABRICATION INTO USEFUL FORMS

Preparation and sizing. All abrasives, with the exception of the naturally appearing fine powders such as talc and iron oxide, must be crushed to the particle size required for use. Sizes in use vary from 4 grit, which measures about 1/4 inch (six millimetres) in diameter, to as fine as 900 grit, which measures about six microns (0.00024 inch) or about one-tenth the thickness of a human hair. In some cases, even finer powders are required when used for the polishing of scratch-free surfaces on high-quality optical lenses and mirrors for high-power telescopes.

For the coarser sizes the crushed grain is measured by a series of test screens as established in most countries by government standards. For sizes 240 grit and smaller, the size of the grains is usually measured by a settling rate or sedimentation test.

Crushing methods have a significant effect on the cutting action and the strength of abrasive grains. Heavy crushing pressures, for example, tend to create splintery, sharp, and weak grains. These penetrate easily, remove material at a fast rate, and their cutting edges break down easily for resharpening. This shape is required in many coated abrasive products.

When a grinding wheel or coated abrasive belt is to be used to grind away large amounts of material under heavy pressures, a strong pulverized abrasive grain stands up longer. The sharp edge is not as necessary because heavy pressures are used that effect the penetration of the abrasive into the material being ground.

Abrasive product manufacture. *Grinding wheels.* The most important abrasive product manufactured is the grinding wheel. Made of abrasive grain and a binder, or "bond," it is a self-sharpening cutting tool. As the grains on the periphery become dull, they are shed from the surface of the grinding wheel, and fresh, sharp cutting edges are exposed. The ability to resharpen is controlled by the nature of the bond used and the ratio of abrasive to bond, measured by volume.

In the manufacture of grinding wheels, abrasives of the proper size and bonding materials are weighed and mixed together in a power mixing machine. When thoroughly mixed, measured amounts of the abrasive and bond mixture are evenly distributed in steel molds. The mold is placed in a powerful hydraulic press, and the mixture is compressed to the desired wheel size, allowing some oversize in dimensions for finishing operations. Pressures vary with wheel size and may be upward of several hundred tons. Some shaving and shaping of the wheels is done before they are baked or fired.

The majority of grinding wheels made have a vitreous, ceramic bond, made of clays and feldspars. The so-called vitrified wheel is fired in high-temperature kilns at temperatures of 2,300° F (1,260° C). Electric-, oil-, and gas-fired kilns are used. The length of the "burn" varies with wheel size and can be as long as two weeks.

The remaining 35–40 percent of the grinding wheels manufactured have organic type bonds using either resins, rubber, or shellac as the bonding material. These wheels are baked at temperatures of between 300° and 400° F (150° and 200° C).

The lower temperature permits the inclusion of steel rings, molded-in threaded bushings, or fibreglass reinforcements, which become baked into the wheel and serve to make it more resistant to breakage from side pressure. The bushings help to hold the wheel in place on certain grinding machines.

Organic wheels can be made much thinner than vitrified wheels, and they are used in place of metal saw blades for cutting a great variety of materials.

Nearly all grinding wheels must be finished after they have been baked or fired. In a process called truing, the wheels are cut to final size, and the outside glazed layers resulting from the kiln are removed, making the sides of the wheel parallel and the size of the arbor hole accurate; at the same time the working surface of the wheel is sharpened. Wheels are trued with conical steel cutters, by rubbing in beds of steel shot, and by actually grinding them with grinding wheels.

Truing a grinding wheel

Grading of wheels assures that they have the correct resistance to wear. Grade or hardness, determined by the amount of bond, permits the grinding wheel to keep itself sharp and free cutting in a variety of conditions. Grinding wheels used for heavy grinding operations are "harder" and are made with greater amounts of bond, retaining the abrasive particles longer under severe conditions such as those found in steel mills and foundries. In the tool room where industrial cutting tools are sharpened, softer wheels, with less bond per unit of abrasive, are required so that as soon as the abrasive grains start to dull, and the possibility of building up heat develops in the sensitive cutting tool being ground, the wheel will resharpen and shed the dulled grains.

Finally, grinding wheels are checked for balance to assure that they will run without vibration.

Grinding wheels six inches in diameter and larger are usually speed tested. The wheel is rotated at a speed at least 50 percent greater than the maximum allowable operating speed. This is a nondestructive measure of the wheel's bursting strength.

The same basic processes are used in the manufacture of abrasive bricks, sticks, and formed shapes. These are used as rubbing blocks, sharpening stones, honing stones, and shaped abrasive media for tumbling or mass finishing.

Sandpapers. Sandpapers (coated abrasive) are the next most significant abrasive product. They consist, basically, of a single layer of abrasive particles held to a flexible backing material.

Manufacture begins with huge rolls of backing material, either paper, cloth, or a combination of the two. The backing is fed to the making machine where the first layer of adhesive is applied. Next, the layer of abrasive is ap-

plied, either by gravity or electrostatically. The electrostatic method orients the slivery type of abrasive used, with the sharp ends facing out. The process can also control the spacing of the grains. This is an advantage in that wide spacing helps to alleviate loading problems when grinding soft, stringy, or gummy materials.

After being coated with the abrasive layer, the product is draped in long festoons in which it dries partially. Then it is run through another sizing operation and a second layer of adhesive is applied. The product is draped, allowed to dry thoroughly, and wound into large-diameter rolls.

The adhesives used to bond the abrasive to the backing are water soluble, waterproof, or a combination of the two. Water-soluble types are used for dry grinding operations and in household sanding, and occasionally on wood-sanding commercial operations. Resin or resin-over-glue types currently in use have the flexibility associated with the soluble types. The all-resin is best for severe operations; its properties are such that the heat of grinding actually increases the adhesive's holding strength.

Disks for right-angle grinders are die cut. Sheets are cut in the standard lengths and widths used in production and household applications. Strips are slit, cut to length, and joined by their ends to make up the coated abrasive belts that have become an essential part of industry, replacing several kinds of abrasive wheels used previously.

The cutting action of coated abrasive products is determined by the abrasive used, the grit size, the density or spacing of the grit, the strength of the adhesive, and the flexibility of the backing material. Flexibility can be achieved by flexing or breaking the bond.

Other products. Other products use the abrasive in the form of grains or powders. In addition to the sizing operation, many types are specifically treated, by calcining, acid, or heating, to make them more suitable for use as lapping abrasive or perhaps as sandblasting grain. For use in lapping and polishing, the abrasive is usually mixed with a vehicle such as mineral or seal oil. Polishing sticks consist of waxes or greases impregnated with various-sized abrasive grains, depending on the particular requirements of the work.

Two materials used really for cleaning rather than grinding are still abrasive in nature. Glass beads, pressure blasted onto a surface, remove rust, scale, and carbon. These have replaced much hand cleaning with steel wool. Steel wool still has some applications.

INDUSTRIAL APPLICATIONS

Grinding, the most important abrasive application, is in some way involved in the manufacture of almost every product. This use may be direct, as when the product requires pieces that must be made within close dimensional tolerance limits, or a very smooth surface, or when used on materials too hard to be machined by conventional cutting tools; or indirect, as when, for example, grinding wheels are used to sharpen cutting tools. The materials that are used to make cutting tools must of course be hard in order for the cutting tool to cut and retain its sharp edge. Abrasive grinding wheels are the only means for sharpening the dull edges of such tools.

Grinding wheels in use in industry today rotate at peripheral speeds of up to 180 miles (290 kilometres) per hour. The abrasive wheel may throw a 40-foot stream of bright yellow sparks and remove upward of 1,000 pounds (450 kilograms) of metal per hour while grinding the imperfections from the surface of a bar of stainless steel. Or the grinding wheel may be as small as 0.022 inch (0.55 millimetre) in diameter, may rotate at 150,000 revolutions per minute, and may grind miniature precision ball bearings to accuracies measured in millionths of an inch.

In the automotive industry, only abrasives can produce the tight fit required between piston rings and cylinders to prevent the escape of compressed gasoline vapours. Valves and valve seats are ground. Bearing surfaces in the engine, transmission, and wheels need specific finish, size, and roundness to assure frictionless rotation. These can be achieved only with abrasive tools.

Abrasive machining, the use of abrasives rather than high-speed steel or tungsten carbide cutting tools, makes use of the self-sharpening grinding wheel and eliminates tool sharpening costs. The ability to grind hardened materials without the previously necessary prehardening machining saves intermediate part-handling operations.

Speed and improved grinding systems, machines, and grinding tools are the main reason for the increased importance of grinding. One-hundred-horsepower motors, automatic loading equipment, high-speed grinding wheels removing unbelievable amounts of hard-to-grind materials, ultrafine tolerances, and machines costing $100,000 and more are part of the new abrasives systems that are capable of fantastic rates of production. Yet the abrasive products portion of the total cost to grind the part may be as low as 5 percent, even though the grinding wheel, with diamond as the abrasive, might cost $14,000. Or the wheel might cost 28 cents and be used to shape a die used in the manufacture of tableware.

Relatively new products, such as plastic bonded tumbling media for mass finishing a multitude of parts, have eliminated time-consuming, hand-deburring operations, a plague to the aircraft industry, in which the high cost of labour makes handwork prohibitively expensive. Parts with rough edges are tumbled in a rotating barrel of loose abrasive or preshaped abrasive pieces. As the mass slides, burrs are ground away, surfaces are finished, and edges are smoothed. When the same parts and media are mixed in a vibrating tub, the process becomes even more productive.

Abrasive wheels have replaced steel saws in many places. Thin, abrasive cutoff wheels are capable of sawing through nearly every material known, at rates faster than those of metal saws, while generating less heat and producing a better cut surface. Some space-age metals, because of their hardnesses, can be cut only with abrasive wheels. Granite, marble, slate, and various building blocks are cut to size with diamond abrasive wheels. Grooves for expansion joints and for the reduction of wet-weather skidding accidents are cut in concrete runways and highways by blades with a metal centre, onto which are brazed metal segments with the diamond abrasive mixed throughout.

Tool sharpening

The sharpening of all types of tools continues to be a major grinding operation. Drills, saws, reamers, milling cutters, broaches, and the great spectrum of knives are kept sharp by abrasives. Coarser grit products are used for their initial shaping. Finer grit abrasives produce keener cutting edges. Ultrasharp tools must be hand honed on natural sharpening or honing stones. Even grinding wheels themselves may require some sharpening. Specially designed steel disks or diamond tools are used to remove dull abrasive cutting edges and create a sharp cutting surface.

In foundries and steel mills, grinding wheels and coated abrasive belts remove the unwanted portions of castings, forgings, and billets. Abrasive grit is pressure blasted against the metal to clean it in preparation for painting. Metal shot is used on softer metallic castings.

The roster of unusual applications for abrasives includes frozen fish cut into fish sticks; animal-gut sutures and guitar strings ground to constant diameters; human skin blemishes and birthmarks removed or bones shaped by plastic surgery; spots and discolorations removed from suede clothing; toothpicks ground round; and stairs made skidproof by abrasive grain. (D.J.B.)

BIBLIOGRAPHY

Industrial ceramics: ERNST ROSENTHAL, *Pottery and Ceramics* (1949), is a good introduction to the manufacture, properties, and uses of clay-based wares; M. CHANDLER, *Ceramics in the Modern World* (1968), extends the treatment, while F. and S.S. SINGER, *Industrial Ceramics* (1963), goes into more technical detail. For the individual product categories, selected references include F.H. NORTON, *Fine Ceramics* (1970), for a well-illustrated coverage of whitewares, and the same author's *Refractories,* 4th ed. (1968), remains the outstanding text for high-temperature materials. V.V. VARGIN, *Technology of Enamels* (1968; orig. pub. in Russian, 1965), is the most current treatment of porcelain enamelling, though A.I. ANDREWS, *Porcelain Enamels,* 2nd ed. (1961), is still authoritative. B.E. WAYE, *Introduction to Technical Ceramics* (1967), and E.C. HENRY, *Electronic Ceramics* (1969), are brief, informative treatments on newer ceramic products. To supplement a basic text such as W.E. WORRALL, *Raw Materials* (1964), current data on raw ma-

terials and processing appear each year in special issues of the trade journals *Ceramic Industry* for January (also see the annual *Ceramic Data Book*) and *Ceramic Age* for July. W.D. KINGERY's misleadingly named *Introduction to Ceramics* (1960) marked the transition from teaching of ceramics by separate industries to consideration of the principles underlying the full scope of ceramics as a broad class of materials. Symposia such as R.M. FULRATH and J.A. PASK (eds.), *Ceramic Microstructures* (1968), have brought together the producers of ceramics and those who are using them in specific, often extreme environments.

Glass and glass products: T.K. DERRY and T.I. WILLIAMS, *A Short History of Technology from the Earliest Times to A.D. 1900* (1960), provides a useful summary of the development of glass technology. C.J. PHILLIPS, *Glass: The Miracle Maker* (1941), is intended to be of interest and use to such users of glass as architects and engineers. J. HOME DICKSON (ed.), *Glass: A Handbook for Students and Technicians* (1951), and F.J.T. MALONEY, *Glass in the Modern World* (1967), are useful introductions to the subject for non-specialist readers with some scientific or technical knowledge, and the latter is probably the most up-to-date work of this type currently available in English. F. GIFFIN, *Glass* (1963), is suitable for non-scientific readers. Among more highly technical studies, G.O. JONES, *Glass* (1956), and G.W. MOREY, *The Properties of Glass* (1956), both contain a great deal of useful data; L. HOLLAND, *Properties of Glass Surfaces* (1964), is an exhaustive theoretical study of a narrower and more specialized field. B.E. MOODY, *Packaging in Glass* (1963), deals with the production of bottles and jars of most kinds as well as with their design and with the associated problems of filling and mechanical handling.

Brick and tile production: H.C. PLUMMER, *Brick and Tile Engineering,* 2nd ed. (1962), is a comprehensive technical publication on brick manufacturing, sizes, design, and uses, including a bibliography of intensive investigation into the behaviour of clay products in structural systems; B. FOERSTER (ed.), *Man and Masonry* (1960), is a pictorial presentation of masonry structures, new and old, as related to the environment of man; *Technical Notes,* a monthly publication of the Structural Clay Products Institute, McLean, Virginia, covers all phases of brick and tile production and use (previous issues are updated as research reveals new data); E.A. SPEISER and H.M. HERBERT, "Ancient Mesopotamia," *Nat. Geogr. Mag.,* 99:41–105 (1951), is an excellent coverage of the ancient use of brick with references to brickmaking; S.E. JONES, "When in Rome," *ibid.,* 137:741–789 (1970), is an illustrated history of masonry, including construction during Hadrian's reign and the brick dome on the Pantheon.

Abrasives: KENNETH B. LEWIS, *The Grinding Wheel* (1959), published by the Grinding Wheel Institute, contains general information on the various types of grinding operations, presented from a nontechnical point of view. Another nontechnical source is P.S. HOUGHTON, *Grinding Wheels and Machines* (1963). The Norton Company, Worcester, Massachusetts, publishes, and updates annually, booklets of general and semitechnical information on abrasives; titles include *Lectures on Grinding, Lectures on Coated Abrasives,* and *Abrasives and Grinding Wheels.* Another source of nontechnical information is the booklet series published by the Abrasive Grain Association, Cleveland, Ohio; titles include *History of Abrasive Grain, Physical and Chemical Properties of Abrasive Grain, Use of Abrasive Grain in the Manufacture of Coated Abrasives, Use of Abrasive Grain in Polishing,* and *Use of Abrasive Grain in Lapping and Metallographic Polishing. Proceedings* of the INDUSTRIAL DIAMOND ASSOCIATION and of the AMERICAN SOCIETY FOR ABRASIVE METHODS are published annually with contents consisting of reprinted versions of papers presented at their conferences. Topics vary from very general to highly technical. An important periodical, published eight times each year, is *Grits and Grinds.* Various papers on abrasives are also available from technical societies such as the Society of Manufacturing Engineers, the American Foundrymen's Society, and the American Society for Abrasive Methods.

(A.G.P./Fr.J.M./J.A.Le./D.J.B.)

Chemical Process Industries

Although the chemical industry may be described simply as the industry that uses chemistry and manufactures chemicals, this definition is not altogether satisfactory because it leaves open the question of what is a chemical. Definitions adopted for statistical economic purposes vary from country to country. Also the Standard International Trade Classification, published by the United Nations, includes explosives and pyrotechnic products as part of its chemicals section. But the classification does not include the man-made fibres, although the preparation of the raw materials for such fibres is as chemical as any branch of manufacture could be.

The complicated characteristics of the chemical industry. The scope of the chemical industry is in part shaped by custom rather than by logic. The petroleum industry is usually thought of as separate from the chemical industry, for in the early days of the industry in the 19th century crude oil was merely subjected to a simple distillation treatment. Modern petroleum industrial processes, however, bring about chemical changes, and some of the products of a modern refinery complex are chemicals by any definition. The term petrochemical is used to describe these chemical operations, but because they are often carried out at the same plant as the primary distillation, the distinction between petroleum industry and chemical industry is difficult to maintain.

Metals in a sense are chemicals because they are produced by chemical means, the ores sometimes requiring chemical methods of dressing before refining; the refining process also involves chemical reactions. Such metals as steel, lead, copper, and zinc are produced in reasonably pure form and are later fabricated into useful shapes. Yet the steel industry, for example, is not considered a part of the chemical industry. In modern metallurgy, such metals as titanium, tantalum, and tungsten are produced by processes involving great chemical skill, yet they are still classified as primary metals.

The boundaries of the chemical industry, then, are somewhat confused. Its main raw materials are the fossil fuels (coal, natural gas, and petroleum), air, water, salt, limestone, sulfur or an equivalent, and some specialized raw materials for special products, such as phosphates and the mineral fluorspar. The industry converts these raw materials into primary, secondary, and tertiary products, a distinction based on the remoteness of the product from the consumer, the primary being remotest. The products are most often end products only as regards the chemical industry itself; a chief characteristic of the chemical industry is that its products nearly always require further processing before reaching the ultimate consumer.

Thus, paradoxically, the chemical industry is its own best customer. An average chemical product is passed from factory to factory several times before it emerges from the chemical industry into the market.

The industry is also its own chief competitor because the same product can often be made in a number of different ways. Phenol, for example, can be made from benzene (which itself can be obtained either from coal or from petroleum) by at least four different processes, and it can also be made from other raw materials. The methods used for the manufacture of caustic soda and sodium carbonate have changed drastically several times.

There are many routes to the same product, and many uses for the same product. The largest use for ethylene glycol, for example, is as an automobile antifreeze, but it is also used as a hydraulic brake fluid. Further processing leads to many derivatives that are used as additives in the textile, pharmaceutical, and cosmetic industries; as emulsifiers in the application of insecticides and fungicides; and as demulsifiers for petroleum. The fundamental chemicals, such as chlorine or sulfuric acid, are used in so many ways as to defy a comprehensive listing.

Because of the competitiveness within the chemical industry and among the chemicals, the chemical industry spends large amounts on research, particularly in the highly industrialized countries. The percentage of revenue spent on research varies from one branch to another; companies specializing in large-volume products that have been widely used for many years spend less, whereas in the newer fields, competition can be met only by intensive research efforts.

Economic aspects. In most fields the United States is the largest producer of chemicals. West Germany, the United Kingdom, France, Italy, and some other European countries are also large producers, and so is the Soviet Union. Japan in the 1960s came into prominence as a very large producer in certain areas. Investment in the chemical industry as a percentage of total investment in a given country may range from 5 to 15 percent for the less developed countries; for the industrial countries it averages about 6 to 8 percent. For some of the underdeveloped countries this percentage can fluctuate widely; for example, the installation of one sizable fertilizer factory could change the percentage markedly in a specific country.

Earlier in the 20th century, there was a marked distinction between economies that were based on coal as a fossil fuel and those based on petroleum. Coal was almost the unique source of the aromatic hydrocarbons. Two forces, however, have worked together to change this situation. First, aromatics can now also be obtained from petroleum, and indeed all hydrocarbon raw materials are now almost interchangeable; second, modern transportation technology makes possible very large-scale shipments by sea, not only of petroleum, crude or in various stages of refinement, but also of natural gas, refrigerated and condensed to a liquid.

Statistics from the chemical industry as a whole can be very misleading because of the practice of lumping together such products as inexpensive sulfuric acid and expensive dyes or fibres; included in some compilations are cosmetics and toiletries, the value of which per pound may be artificially high. Chemical industry statistics from different countries may have different bases of calculation; indeed the basis may change from time to time in the same country. An additional source of confusion is that in some cases the production is quoted, not in tons of the product itself, but in tons of the content of the important component.

For purposes of simplicity, various divisions of the chemical industry, such as heavy inorganic and organic chemicals and various families of end products, will be described in turn, and separately, although it should be borne in mind that they interact constantly. In some cases, a certain raw material will be followed a considerable distance, but an occasional glance at the end products will help to understand to what places, in terms of the final consumer, the whole process is directed. The first division to be discussed is the heavy inorganic chemicals, starting at the historical beginning of the chemical industry with the Leblanc process. The terms heavy chemical industry and light chemical industry, however, are not precisely exclusive, because numerous operations fall somewhere between the two classes. The two classes do, however, at their extremes correlate with other differences. For example, the appearance of two kinds of plants is characteristically different. The large-scale chemical plant is characterized by large pieces of equipment of odd shapes and sizes standing independent and immobile. Long rows of distilling columns are prominent, but because the material being processed is normally confined in pipes or vessels, no very discernible activity takes place. Few personnel are in evidence.

The light chemical industry is entirely different; it involves many different pieces of equipment of moderate size, often of stainless steel or lined with glass or enamel. This equipment is housed in buildings like those for, say, assembling light machinery. Numerous personnel are present. Both types of plant require large amounts of capital. (A.St.)

This article is divided into the following sections:

Heavy chemical industries 253
- Inorganic chemicals
  - Sodium carbonate and other alkalies
  - Sulfuric acid
  - Carbon disulfide
- Fertilizers 254
  - Potassium
  - Phosphorus
  - Nitrogen
- Halogens and their compounds 256
  - Chlorine
  - Fluorine
  - Bromine
  - Iodine
- Organic chemicals 258
  - Aromatic hydrocarbons
  - Aliphatic hydrocarbons
  - Elastomers
  - Film materials
  - Carbon black
  - Alcohols and their derivatives

Soaps and detergents 260
- History 261
  - Use
  - Early soap production
  - Early synthetic detergents
- Soap manufacturing processes and products 262
  - Raw materials and additives
  - Soap production processes
  - Anionic detergents
  - Nonionic detergents
  - Cationic detergents
  - Ampholytic detergents
  - Finishing synthetic detergents

Dyes 265
- History 265
  - Natural dyes
  - Synthetic dyes
- Raw materials for manufacture of dyes 266
- Manufacturing processes for intermediates 266
- Manufacturing processes for dyes and pigments 267

Pharmaceutical industry 268
- Production of pharmaceuticals 268
  - Raw materials and their purity
  - Preparation from natural sources
  - Preparation by chemical synthesis
- Preparation of dosage forms 273
  - Solutions
  - Semisolid dosage forms
  - Solid dosage forms
  - Sterile medicaments
  - Dispersions
  - Radioactive dosage forms

Explosives 275
- Types of chemical explosives 276
- Black powder 276
  - Manufacture of black powder
  - Ignition of black powder
  - Safety fuse
- Nitroglycerin 277
- Dynamite 277
  - Ammonium nitrate
  - Low-freezing dynamite
- Other explosives 278
- Modern high explosives
  - Ammonium nitrate–fuel oil mixtures
  - Water gels
- Nitrocellulosic explosives 279
- Blasting caps 280
  - Electrical firing
  - Delay systems
- Detonating cord 281
- Military explosives 281
  - TNT
  - Picric acid and ammonium picrate
  - Shaped charges
- Other industrial applications 282
  - Explosive rivets
  - Explosive bonding

Rubber 282
- Natural rubber 282
  - History
  - The rubber tree
  - Modern production
  - Rubber types and grades
  - Chemical and physical properties
  - Distinctive uses of natural rubber
- Synthetic rubber 285
  - History
  - Types of synthetic rubber
  - Reclaimed rubber
- Compounding 288
- Processes 289
- Major applications 290

Plastics and resins 290
- History of the plastics industry 291
- Formation and use of plastic materials 291
  - Polymerization
  - Thermosetting resins
  - Thermoplastic resins
  - Other types of resins
- Plastics manufacture and fabrication 297
  - Natural raw materials
  - Chemicals and additives
  - Basic processing techniques
  - Fabrication and finishing
- Industrial plastics 302
  - Structural foams
  - Sheets and films
  - Laminates, reinforced plastics, and composites
  - Miscellaneous industrial plastics

Man-made fibres 304
- History and development 304
- Fundamentals of production 305
  - Spinning
  - Stretching and orientation
  - Filament modification
  - Physical form of man-made fibres
- Natural polymer fibres 307
  - Rayons
  - Cellulose ester fibres
  - Protein fibres
  - Natural rubber fibres
- Synthetic fibres 310
  - Polyamide fibres
  - Polyester fibres
  - Polyvinyl derivatives
  - Polyolefin fibres
  - Polyurethane fibres
- Inorganic synthetic fibres 314
  - Glass fibres
  - Aluminum silicate fibres
  - Metallic fibres
  - Miscellaneous inorganic synthetic fibres
- Economic aspects 317

Paints and varnishes 317
- Historical development 317
  - Origins
  - Emergence of paint science and technology
- Paint manufacture 318
- Pigments and other colorants 318
  - Colour and hiding
  - Pigments
- Varnishes and other paint binders 320
  - Natural binders
  - Modern vehicles
- Films 321
  - Film formation
  - Characteristics
- Economic aspects 322

Papermaking 322
- The papermaking process 322
  - Historical development
  - Fibre sources
  - Processes for preparing pulp
  - Manufacture of paper and paperboard
  - The world paper industry
- Paper properties and uses
  - Substance and quantity measurement
  - Strength and durability
  - Optical properties
  - Paper grades

Heavy chemical industries

In the chemical process industries there is no precise definition of what production volume qualifies as heavy, but there is no question about the identity of the perennial leaders of the group of heavy chemicals. In the late 20th century at least 50 chemical products have been made in quantities exceeding 1,000,000,000 pounds per year in the United States alone. Sulfuric, phosphoric, and nitric acids; ammonia; gaseous nitrogen, oxygen, and chlorine; lime; ethylene; sodium hydroxide; and sodium carbonate were consistently at the top of the list. All of these compounds and many others are made in enormous tonnages, though none of them is consumed by individual persons as are other chemical products, such as aspirin, toilet soap, or automobile tires. Instead, each of them is used by its own producer or some other industrial establishment in making a succession of other products.

A common feature of heavy chemicals is their versatility: they can all be used in dozens or hundreds of ways. Another is the orientation of their producers toward sources of raw materials rather than toward the ultimate consumer's preferences. That is, the producer of sulfuric acid is quite sure of a market as long as the acid is of satisfactory quality and competitive price; the availability of the fundamental raw material (sulfur) and its economical conversion into sulfuric acid are of more concern than short-term fluctuations in the demand for any particular retail product. (Ed.)

INORGANIC CHEMICALS

Sodium carbonate and other alkalies. In 1775 the French Academy of Sciences offered an award for a practical method for converting common salt, sodium chloride, into sodium carbonate, a chemical needed in substantial amounts for the manufacture of both soap and glass. Nicolas Leblanc, a surgeon with a bent for practical chemistry, invented such a process. His patron, the duc d'Orléans, set up a factory for the process in 1791, but work was interrupted by the French Revolution. The process was not finally put into industrial operation until 1823 in England, after which it continued to be used to prepare sodium carbonate for almost 100 years.

The Leblanc process. The first step in the Leblanc process was to treat sodium chloride with sulfuric acid. This treatment produced sodium sulfate and hydrogen chloride. The sodium sulfate was then heated with limestone and coal to produce black ash, which contained the desired sodium carbonate, mixed with calcium sulfide and some unreacted coal. Solution of the sodium carbonate in water removed it from the black ash, and the solution was then crystallized. From this operation derives the expression soda ash that is still used for sodium carbonate.

It was soon found that when hydrogen chloride was allowed to escape into the atmosphere, it caused severe damage to vegetation over a wide area. To eliminate the pollution problem, methods to convert the dissolved hydrogen chloride to elemental chlorine were developed. The chlorine, absorbed in lime, was used to make bleaching powder, for which there was a growing demand.

Chlorine for bleaching powder

Because calcium sulfide contained in the black ash had a highly unpleasant odour, methods were developed to remove it by recovering the sulfur, thereby providing at least part of the raw material for the sulfuric acid required in the first part of the process. Thus the Leblanc process demonstrated, at the very beginning, the typical ability of the chemical industry to develop new processes and new products, and often in so doing to turn a liability into an asset.

The ammonia soda (Solvay) process. The Leblanc process was eventually replaced by the ammonia-soda process (called the Solvay process), which was first practiced successfully in Belgium in the 1860s. In this process, sodium chloride as a strong brine is treated with ammonia and carbon dioxide to give sodium bicarbonate and ammonium chloride. The desired sodium carbonate is easily obtained from the bicarbonate by heating. Then, when the ammonium chloride is treated with lime, it gives calcium chloride and ammonia. Thus, the chlorine that was in the original sodium chloride appears as calcium chloride, which is largely discarded (among the few uses for this compound is to melt snow and ice from roads and sidewalks). The ammonia thus regenerated is fed back into the first part of the process. Efficient recovery of nearly all the ammonia is essential to the economic operation of the process, the loss of ammonia in a well-run operation being no more than 0.1 percent of the weight of the product.

A use for calcium chloride

Electrolytic process. Later in the 19th century the development of electrical power generation made possible the electrochemical industry. This not clearly identifiable branch of the chemical industry includes a number of applications in which electrolysis, the breaking down of a compound in solution into its elements by means of an electric current, is used to bring about a chemical change. Electrolysis of sodium chloride can lead to chlorine and either sodium hydroxide (if the NaCl was in solution) or metallic sodium (if the NaCl was fused). Sodium hydroxide, an alkali like sodium carbonate, in some cases competes with it for the same applications, and in any case the two are interconvertible by rather simple processes. Sodium chloride can be made into an alkali by either of the two processes, the difference between them being that the ammonia soda process gives the chlorine in the form of calcium chloride, a compound of small economic value, while the electrolytic processes produce elemental chlorine, which has nearly innumerable uses in the chemical industry, including the manufacture of plastic polyvinyl chloride, the plastic material produced in the largest volume. For this reason the ammonia soda process, having displaced the Leblanc process, has found itself being displaced, the older ammonia soda plants continuing to operate very efficiently but no new ammonia soda plants being built.

Other important processes. The need for sodium carbonate in the manufacture of soap and glass that led to the Leblanc process also led to the creation of the alkali industry and the chlor-alkali industry, another of the historic landmarks of the chemical industry (see *Chlorine* below).

Sulfuric acid. Sulfuric acid is by far the largest single product of the chemical industry. The chamber process for its preparation on the scale required by the Leblanc process might be regarded as the most important long-term contribution of the latter.

Chamber process. When sulfur is burned in air, sulfur dioxide is formed, and this, when combined with water, gives sulfurous acid. To form sulfuric acid, the dioxide is combined with oxygen to form the trioxide, which is then combined with water. A technique to form the trioxide, called the chamber process, developed in the early days of the operation of the Leblanc process. In this technique, the reaction between sulfur dioxide and oxygen takes place in the presence of water and of oxides of nitrogen. Because the reaction is rather slow, sufficient residence time must be provided for the mixed gases to react. This gaseous mixture is highly corrosive, and the reaction must be carried out in containers made of lead.

Contact process. Lead is a material awkward to use in construction, and the process cannot deliver acid more concentrated than about 78 percent without special treatment. Therefore, the chamber process has been largely replaced by the contact process, in which the reaction takes place in a hot reactor, over a platinum or vanadium compound catalyst, a substance that increases the speed of the reaction without becoming chemically involved.

Uses. Of the very large world production of sulfuric acid, almost half goes to the manufacture of superphosphate and related fertilizers. Other uses of the acid are so multifarious as almost to defy enumeration, notable ones being the manufacture of high-octane gasoline, of titanium dioxide (a white pigment, also a filler for some plastics, and for paper), explosives, rayon, the processing of uranium, and the pickling of steel.

Sources of sulfur. Because sulfuric acid is indispensable to so many industries, its primary raw material is of the greatest importance. The needed sulfur is obtainable from a number of sources. Originally, sulfur came chiefly from certain volcanic deposits in Sicily. By the beginning of the 20th century this source was insufficient, but the supply

was augmented by sulfur that occurs underground in the southern United States. This sulfur is not mined but is recovered by the so-called Frasch process, in which the sulfur is melted underground by hot water and the mixture brought to the surface in liquid form.

Other sources of sulfur include the ore iron pyrite, an iron–sulfur compound that can be burned to produce sulfur dioxide, and some natural gases, called sour gas, that contain appreciable quantities of hydrogen sulfide. Certain metal sulfides, such as those of zinc and copper, are contained in the ores of those metals. When these ores are roasted, sulfur dioxide is given off. Sulfur is usually shipped in its elemental form rather than in the form of sulfuric acid.

Under some circumstances, the sulfuric-acid stage of manufacture can be avoided. Ammonium sulfate, a fertilizer, is normally made by causing ammonia to react with sulfuric acid. In many parts of the world, abundant supplies of calcium sulfate in any of several mineral forms can be used to make the ammonium sulfate by combining it with ammonia and water. This process brings the sulfur in the calcium sulfate deposits into use. Because deposits of calcium sulfate throughout the world are very extensive, development of such a process would make the available resources of sulfur almost limitless.

Abundance of calcium sulfate supplies

The sulfur present in low percentages in fossil fuels is a notorious source of air pollution in most industrial countries. Removal of sulfur from crude oil adds to the sulfur supply and reduces pollution. It is less easy to remove the sulfur directly from coal.

Carbon disulfide. Carbon disulfide is made by the reaction of carbon and sulfur. Carbon comes from natural gas, and the sulfur may be supplied in the elemental form, as hydrogen sulfide, or as sulfur dioxide. The chief uses of carbon disulfide are for the manufacture of rayon and for regenerated cellulose film. These two products are made in such large quantity that carbon disulfide is a heavy chemical, by any standard.

FERTILIZERS

Fertilizers represent one of the largest market commodities for the chemical industry. A very large industry in all industrialized countries, it is a very important one for introduction as early as possible into developing countries.

The crucial elements that have to be added to the soil in considerable quantities in the form of fertilizer are nitrogen, phosphorus, and potassium, in each case in the form of a suitable compound. These are the major fertilizer elements, or macronutrients. Calcium, magnesium, and sulfur are regarded as secondary nutrients; and it is sometimes necessary to add them. Numerous other elements are required only in trace quantities; certain soils may be deficient in boron, copper, zinc, or molybdenum, making it necessary to add very small quantities of these. As a great industry, however, fertilizers are based on the three elements mentioned above.

Nitrogen is present in vast quantities in the air, making up about 78 percent of the atmosphere. It enters the chemical industry as ammonia, produced through fixation of atmospheric nitrogen, described below. For phosphorus and potassium, it is necessary to find mineral sources and to convert them into a form suitable for use. These three elements are not used in fertilizer only, however; they have other uses and interact with other facets of the chemical industry, making a highly complicated picture.

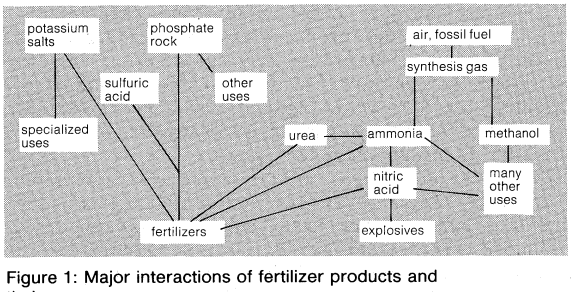


Figure 1: Major interactions of fertilizer products and their uses.

A schematized overview of some of these interactions is presented in Figure 1.

Potassium. The simplest part of this diagram is the portion representing potassium. The element potassium is seventh in order of abundance in the Earth's crust, about the same order as sodium, which it resembles very closely in its properties. Although sodium is readily available in the sodium chloride in the ocean, most of the potassium is contained in small proportions in a large number of mineral formations, from which it cannot be economically extracted. When the use of potassium salts as fertilizers began in the second half of the 19th century, it was believed that Germany had a monopoly with the deposits at Stassfurt, but many other workable deposits of potassium salts were later found in other parts of the world. World reserves are adequate for thousands of years, with very large deposits in the Soviet Union, Canada (Saskatchewan), and Germany (East and West).

Resemblance of potassium to sodium

Potassium chloride is the principal commercial form of potash, and some potassium nitrate is also produced. About 90 percent of the production of these goes to fertilizers. For other purposes, the similar sodium salts are cheaper, but for a few special uses potassium has the advantage. Some ceramic uses require potassium, and potassium bicarbonate is more effective than sodium bicarbonate in extinguishing fires.

Phosphorus. Phosphorus presents a somewhat more complicated picture. It has many uses other than in fertilizers. By far the largest source is phosphate rock, although some use is made of phosphatic iron ore, from which the phosphorus is obtained as a by-product from the slag. As with potassium, there are extensive reserves. The largest deposits are in North Africa (Morocco, Algeria, Tunisia), the United States (largely Florida), and the Soviet Union, but there are also sizable deposits in numerous other countries.

Phosphate rock is found in deposits of sedimentary origin, laid down originally in beds on the ocean floor. The rock consists largely of the insoluble tricalcium phosphate, together with some other materials, including some fluorine. To be used as a fertilizer, phosphate must be converted to a form that is soluble in water, even if only slightly so.

Phosphoric acid (H_3PO_4) has three hydrogen atoms, all of which are replaceable by a metal. Tricalcium phosphate, in which all three of the hydrogen atoms are replaced by calcium, must be converted to the soluble form, monocalcium phosphate, in which only one hydrogen atom is replaced by calcium. The conversion is done by sulfuric acid, which converts the phosphate rock to superphosphate, widely used as fertilizer. This operation requires very large tonnages of sulfuric acid.

The fertilizer industry is not only a matter of manufacturing the right chemical but also of distribution, getting the right material to the right place at the right time. Fertilizers are made centrally, but must be distributed over a large agricultural area. A fertilizer factory is, typically, a large installation, characterized by enormous storage silos; the product is manufactured all the year round, but it requires considerable space to store it until the few weeks during which it is distributed on farmlands.

Manufacturing considerations

The weight of the superphosphate is greater than that of the original phosphate rock by the amount of the sulfuric acid added; the superphosphate also carries the dead weight of the calcium sulfate that is formed in the manufacturing process. This dead weight can be reduced by replacing sulfuric acid with phosphoric acid (itself obtained by the action of sulfuric acid on phosphate rock, followed by separating the products; or else by an electric furnace process). This process results in triple superphosphate, in which all the calcium originally in the phosphate rock appears as calcium monophosphate. The useful content of the fertilizer, expressed as the percent of phosphoric oxide, is increased from 20 percent in ordinary superphosphate to about 45 percent in the triple variety, resulting in a better than twofold reduction in the amount of material that must be distributed to provide a given amount of the useful oxide.

Instead of using either sulfuric or phosphoric acid to treat the phosphate rock, nitric acid can be employed. One of

the resulting products, calcium nitrate, is itself a fertilizer, so what is obtained is one of the many varieties of mixed fertilizers. Instead of neutralizing phosphoric acid with calcium, which contributes nothing but dead weight, ammonia can be used, giving ammonium phosphate, in which both constituents contribute fertilizer elements. Such improvements in fertilizers are constantly being made.

Many other compounds of phosphorus are used. One group is comprised of phosphoric acid and various phosphates derived from it. The acid itself is used in soft drinks for its pleasant taste when sweetened and its nutritive value. Other food applications include the use of disodium phosphate in processed cheese; and phosphates in baking powder, flameproofing, and the treatment of boiler water in steam plants. An important use of some of the phosphates is in detergents, discussed below.

Elemental phosphorus exists in many allotropic forms. White phosphorus is used in rodent poison, and by the military for smoke generation. Red phosphorus, comparatively harmless, is used in matches. Ferrophosphorus, a combination of phosphorus with iron, is used as an ingredient in high-strength low-alloy steel. In addition, the many organic compounds of phosphorus have varied uses, including those as additives for gasoline and lubricating oil, as plasticizers for plastics that otherwise would be inconveniently rigid, and, in some cases, as powerful insecticides, related to nerve poisons.

Nitrogen. The production of nitrogen not only is a major branch of the fertilizer industry but it opens up a most important segment of the chemical industry as a whole.

Farm manure long supplied enough nitrogenous fertilizer for agriculture, but late in the 19th century it was realized that agriculture was outgrowing this source. A certain amount of ammonium sulfate was available as a by-product of the carbonization of coal, and the large deposits of sodium nitrate discovered in Chile helped out for a time. The long-range problem of supply, however, was not solved until just before World War I, when the research of Fritz Haber in Germany brought into commercial operation the method of ammonia synthesis that is used, in principle, today. The immediate motivation for this great development was Germany's need for an indigenous source of nitrogen for military explosives. The close interrelation between the use of nitrogen for fertilizers and for explosives persists to this day.

Haber process for ammonia

Because nitrogen comprises 78 percent of the air, there is a little more than 11 pounds of nitrogen over every square inch of the earth's surface. Nitrogen, however, is a rather inert element; it is difficult to get it to combine with any other element. Haber succeeded in getting nitrogen to combine with hydrogen by the use of high pressure, moderately high temperatures, and a catalyst.

The hydrogen for ammonia (NH_3) is usually obtained by decomposing water (H_2O). This process requires energy, in some cases supplied by electricity, but more often from fossil fuels. In some cases the hydrogen is obtained directly from the fossil fuel, without decomposing water.

Haber used coke as a fuel. Carbon can burn either to carbon dioxide or, if the supply of air is kept short, to carbon monoxide, by a process known as the producer gas reaction. The gaseous product is a mixture of carbon monoxide with the nitrogen that was originally in the air.

The red-hot coke can also be heated with steam to yield carbon monoxide and hydrogen, a mixture known as water gas. It is also possible to carry out a water-gas shift reaction by passing the water gas with more steam over a catalyst, yielding more hydrogen, and carbon dioxide. The carbon dioxide is removed by dissolving it in water at a pressure of about ten atmospheres; it can also be utilized directly, as noted below. Starting then from water gas, and converting a certain proportion of the carbon monoxide to carbon dioxide and hydrogen, it is possible to arrive at a mixture of carbon monoxide and hydrogen in any proportion.

Synthesis gas. In Figure 1, the words synthesis gas have been shown as the source of two products, ammonia and methanol. It is not quite the same synthesis gas in the two cases, but they are closely related. The mixture of carbon monoxide and hydrogen described above is the synthesis gas that is the source of methanol. But ammonia requires nitrogen, which is obtained from the producer gas by causing it to undergo the water-gas shift reaction, yielding hydrogen. Ammonia requires much more hydrogen, which is obtained from water gas subjected to the water-gas shift. And so, by appropriate mixing, ammonia synthesis gas of exactly the right composition can be obtained.

Water-gas shift reaction

The above description is a simplified account of how synthesis gas, either for ammonia or for methanol, is obtained from fossil fuel as a source of energy, but it gives an idea of the versatility of the operations. There are many possible variations in detail, depending largely on the particular fuel that is used. The nitrogen industry, which has grown steadily since shortly after World War I, was originally based largely on coke, either from coal or lignite (brown coal). There has been a gradual change to petroleum products as the fossil fuel. As is true with many other branches of the chemical industry, the latest trend is to move to natural gas.

The carbon dioxide removed during the preparation of the synthesis gas can be caused to react with ammonia, often at the same plant, to form urea, $CO(NH_2)_2$. This is an excellent fertilizer, highly concentrated in nitrogen (46.6 percent), and also useful as an additive in animal feed to provide the nitrogen for formation of meat protein. Urea is also used for an important series of resins and plastics by reaction with formaldehyde, derived from methanol.

Ammonia can be applied as a fertilizer in numerous forms, ranging from the application of liquid ammonia beneath the surface of the soil, or solutions of ammonia in water (also containing other fertilizer ingredients), or as ammonium nitrate, or other products from nitric acid, which itself is derived from ammonia. Ammonia also has other uses within the chemical industry. The small amount of ammonia consumed in the course of making sodium carbonate by the ammonia soda process formerly amounted to a considerable volume. Ammonia is used in one process for making rayon, as a refrigerant in large commercial refrigeration establishments, and as a convenient portable source of hydrogen. Hydrogen can be compressed into cylinders, but ammonia, which forms a liquid on compression, packs far more hydrogen into the same volume; it is decomposed by heat into hydrogen and nitrogen; the nitrogen is used to provide an inert atmosphere for many metallurgical operations.

Ammonia as a source of hydrogen

Nitric acid. By far the most important use of ammonia within the chemical industry is to produce nitric acid (HNO_3). Nitrogen and oxygen can be made to combine directly with one another only with considerable difficulty. A process based on such a direct combination, but employing large quantities of electrical power, was in use in the 1920s and 1930s in Norway, where hydroelectric power is readily available. It has not proved economical in modern conditions.

Ammonia burns in air, or in oxygen, causing the hydrogen atoms to burn off, forming water and leaving free nitrogen. With the aid of a catalyst, platinum with a small percentage of the related metal rhodium, ammonia is oxidized to oxides of nitrogen that can be made to react with water to form nitric acid.

Nitric acid treated with ammonia gives ammonium nitrate, a most important fertilizer. Ammonium nitrate, moreover, is also an important constituent of many explosives. Three fundamental explosive materials are obtained by nitrating (treating with nitric acid, often in a mixture with sulfuric acid): cellulose, obtained from wood, gives cellulose nitrate (formerly called nitrocellulose); glycerol gives glyceryl trinitrate (formerly called nitroglycerin); and toluene gives trinitrotoluene, or TNT. Another explosive ingredient is ammonium picrate, derived from picric acid, the relationship of which appears more clearly in its systematic name, 2,4,6-trinitrophenol.

A minor but still important segment of the explosives industry is the production of detonating agents, or such priming compositions as lead azide [$Pb(N_3)_2$], silver azide (AgN_3), and mercury fulminate [$Hg(ONC)_2$]. These are not nitrates or nitro compounds, although some other detonators are, but they all contain nitrogen, and nitric acid is involved in their manufacture.

Related to the explosives are the rocket propellants. A rocket-propelled missile or spacecraft launch vehicle must carry both reactive components (fuel and oxidizer) with it, either in different molecules or in the same molecule. Essentially, rocket propellants consist of an oxidant and a reductant. The oxidant is not necessarily a derivative of nitric acid, but may also be liquid oxygen, ozone (O_3), liquid fluorine, or chlorine trifluoride.

Other uses for nitric acid not related to explosives or propellants include the production of cellulose nitrate for use in coatings. Without a pigment it forms a clear varnish much used in furniture finishing. Pigmented, it forms brilliant shiny coatings referred to as lacquers. At one time a fibre similar to rayon was made from cellulose nitrate.

Nitrating benzene (Figure 2) yields nitrobenzene, which can be reduced to aminobenzene, better known as aniline. Aniline can also be made by reacting ammonia with chlorobenzene, obtained from benzene. Benzene and ammonia are required in either case. Similar treatment applied to naphthalene ($C_{10}H_8$) results in naphthylamine. Both aniline and naphthylamine are the parents of a large number of dyes, but today synthetic dyes are usually petrochemical in origin (see the major section *Dyes* below). Aniline, naphthylamine, and the other dye intermediates lead also to pharmaceuticals, photographic chemicals, and chemicals used in rubber processing.

Synthetic dyes from petrochemicals

The above account gives an idea of the importance, not only for fertilizers but for many other products, of the process known as fixation of atmospheric nitrogen; that is, taking nitrogen from the air and converting it into some form in which it is usable. The tremendous increase in the production of fertilizers has led to the erection of huge ammonia plants. The plants must have available a source of fossil fuel, but petroleum and natural gas are easily transported, so that there is a tendency to locate the plants near the ultimate destination of the product.

A typical plant comprises all the equipment for the preparation of the synthesis gas on the requisite scale, along with equipment for purifying the gas. In the synthesis of ammonia (but not of methanol) any compound of oxygen is a poison for (reduces the effectiveness of) the catalyst, and so traces of carbon dioxide and carbon monoxide must be carefully removed. Compressing the gas to the desired pressure requires extensive engineering equipment. The higher the pressure the greater the yield but the higher the actual cost of compression. The higher the temperature the lower the yield, but the temperature cannot be lowered indefinitely to obtain better yields because lower temperatures slow down the reaction. The temperature used is of the order of 500° C (930° F). The choice of temperature and of pressure is a carefully worked out compromise to give optimal results. The yield equals the amount of nitrogen and hydrogen that combine to form ammonia in any one pass through the converters. Only a fraction is converted each time, but after each pass the ammonia is removed and the remaining gas is recycled. Atmospheric nitrogen contains about 1 percent argon, a totally inert gas, which must be removed from time to time so that it does not build up in the system indefinitely. There is also usually a nearby nitric acid factory and equipment for producing ammonium nitrate in the exact grain size for convenient application as fertilizer.

Optimization of temperature and pressure

The growing plants of agricultural crops do not receive all of their nitrogen from synthetic fertilizer. A certain proportion is supplied by natural means. Some plants, notably beans, have a symbiotic relationship with nitrifying bacteria that are able to "fix" the nitrogen in the air and to combine it into a form available for plant life. This natural synthesis takes place without the necessity for pressures of several hundred atmospheres or of high temperatures. In many laboratory syntheses of natural products, great success has been obtained by quiet reactions, without extreme conditions, by a process of following nature gently, rather than by brute force. Ammonia may some day be synthesized by some process that more nearly resembles the natural way; research is being carried out along these lines, and it may be that a far easier approach to fixation of atmospheric nitrogen will be part of the chemical industry of the future. (A.St./Ed.)

HALOGENS AND THEIR COMPOUNDS

Chlorine. *Uses.* The first large-scale use of chlorine was in the manufacture of bleaching powder for use in making paper and cotton textiles. Bleaching powder was later replaced by liquid chlorine, which also came into widespread use as a germicide for public water supplies. Presently the principal use of chlorine is in making chemical compounds. Important inorganic chemicals made by direct action of chlorine on other substances include sulfur chloride, thionyl chloride, phosgene, aluminum chloride, iron(III) chloride, titanium(IV) chloride, tin(IV) chloride, and potassium chlorate.

Organic chemicals made directly from chlorine include derivatives of methane (methyl chloride, methylene chloride, chloroform, and carbon tetrachloride); chlorobenzene and *ortho-* and *para*-dichlorobenzenes; ethyl chloride; and ethylene chloride.

Commercial preparation. Of several processes that have been used for the manufacture of chlorine, the oldest employed the reaction of hydrochloric acid with manganese dioxide. The procedure was inefficient, and its commercial application was short-lived.

A process introduced around 1868 by the English chemist Henry Deacon was based on the reaction of atmospheric oxygen with hydrochloric acid, which was available as a by-product of the Leblanc process for making soda ash; when the Leblanc process became obsolete, the Deacon process fell into disuse.

The chlor-alkali industry—in which chlorine and caustic soda (sodium hydroxide) are produced simultaneously by electrolytic decomposition of salt (sodium chloride)—has become the principal source of chlorine during the 20th century. As noted earlier, in the two important versions of the electrolytic process, brine is the electrolyte (in which the passage of electric current occurs by the movement of charged particles called ions), and graphite rods are the anodes (positive terminals). The difference between the two processes derives from the distinct behaviour of iron and of mercury when those metals are used as cathodes (negative terminals).

Chlor-alkali industry

In brine, the two substances susceptible to chemical reduction are positively charged sodium ions and neutral water molecules. At a reversible cathode, reduction of sodium ions requires a higher voltage than does the reduction of water molecules, and application of a voltage high enough to reduce sodium ions would effect reduction of a considerable amount of water but of a very small number of sodium ions. The reaction occurring at the surface of an iron cathode is represented by the following equation:

$$\underset{\text{water}}{2H_2O} + \underset{\text{electron}}{2e^-} \rightarrow \underset{\text{hydroxide ion}}{2OH^-} + \underset{\text{hydrogen gas}}{H_2}$$

At a mercury cathode, on the other hand, appreciable reduction of water requires a much higher voltage than that needed at an iron cathode. This so-called overpotential is so great, in fact, that the electrode voltage can be raised to that needed for the reduction of sodium ions without affecting the water molecules.

Passage of a direct electric current through brine is attended by chemical changes at the surfaces where the electrodes come in contact with the electrolyte. At the graphite anode, chloride ions present in the dissolved salt are converted by oxidation to elemental chlorine, which is led away through a vent. At the iron cathode, reduction of water takes place, according to the equation shown above. The hydrogen gas is removed, while the hydroxide ions remain in the solution. The net result is that chloride ions and water are consumed and chlorine gas, hydrogen gas, and hydroxide ions are produced. Complete conversion of chloride to hydroxide is not practical, but as brine is continuously introduced at the top of the cell, a solution containing nearly equal amounts of salt and caustic soda is withdrawn at the bottom. Purification of the effluent liquor yields solid sodium hydroxide containing only a small amount of salt.

Successful production of chlorine and caustic soda in these cells requires that the two products be separated, because upon mixing they would react with one another.

The chlorine is kept away from the caustic by interposing a diaphragm between the electrodes: such cells are commonly called diaphragm cells.

Mercury cell

In the other main variant of the chlor-alkali process, the so-called mercury cell is employed. The cathode in such a cell is a shallow layer of mercury flowing across the bottom of the vessel; graphite anodes extend down into the brine electrolyte. A powerful direct current is caused to pass between the graphite rods and the mercury surface. At the anodes, chloride ions are converted to chlorine gas, as in the diaphragm cell; the reaction occurring at the mercury cathode, however, differs from that at an iron cathode. Positively charged sodium ions in the brine migrate to the mercury surface, where the voltage is high enough to reduce them to sodium metal without reducing the water because of the above-noted overpotential of mercury. The metallic sodium formed at the cathode dissolves in the mercury, and the solution (called an amalgam) flows out of the cell into another vessel, where it is brought into contact with water, which reacts with the sodium to form sodium hydroxide and hydrogen.

The overall result of operating a mercury cell is the same as that of operating a diaphragm cell: sodium chloride and water are changed into sodium hydroxide, chlorine, and hydrogen. Use of the mercury cell, however, makes it possible to generate the sodium hydroxide in the absence of salt, so that evaporation of the caustic liquor produces solid sodium hydroxide completely free of sodium chloride. The higher purity of the product makes it more desirable for certain applications, notably in the manufacture of rayon.

(J.V.K.)

Fluorine. *Aluminum refining.* The fluorine industry is intimately related to the production of aluminum. Alumina (aluminum oxide, Al_2O_3) can be reduced to metallic aluminum by electrolysis when fused with a flux consisting of sodium fluoroaluminate (Na_3AlF_6), usually called cryolite. After starting the process, the cryolite is not used up in massive quantities, but a small supply is needed to make up for inevitable losses. Cryolite is a rare mineral, however, found in commercial quantities only in Greenland. The supply is limited, and it has other uses in glass, in enamels, and as a filler for resin-bonded grinding wheels.

The supply problem was solved by the development of synthetic cryolite. For this synthetic, however, a source of fluorine was needed. Fluorine is actually somewhat more abundant in the Earth's crust than chlorine, but most of it is distributed in various rocks in very small quantities. In a form available to the industrial chemist, it is much scarcer than chlorine. Until the 1960s almost the only source was fluorspar (CaF_2), a mineral long known and used as a flux in various metallurgical operations. It is still so used, in quantities larger than before, because the processes that are coming into greatest use for making steel, the basic oxygen process and the electric furnace, use two to three times as much flux as the earlier open-hearth furnaces did. The mineral fluorspar is widely distributed, but the supplies of good quality ore are not large; it has been found necessary to utilize lower grade ores, making the processing more expensive. A very large reserve that can be tapped for fluorine is the 3 percent or so that is present in some phosphate rock. In the past this fluorine content was seldom recovered; the future will undoubtedly see a major reversal.

Sources of fluorine

Refrigerants. These inorganic uses, as a flux and in the manufacture of aluminum, formerly comprised almost the whole of the fluorine industry. The organic fluorine industry, a separate branch, began in the late 1920s with the discovery by Thomas Midgley, Jr., of the United States, of the fluorine-containing refrigerants. A new refrigerant was needed for the domestic refrigerators that were just beginning to be produced on a large scale. Ammonia was unsuitable because even a very minute leak would give an unpleasant smell, and breakdown would release poisonous quantities of the gas. Although many fluorine compounds were known to be poisonous, Midgley found some that were remarkably nontoxic. They also had the physical properties required for a refrigerant, and were totally odourless.

The most used of these is Freon 12 (CCl_2F_2), dichlorodifluoromethane; also used is Freon 22 ($CHClF_2$), chlorodifluoromethane. Several analogous compounds containing carbon, fluorine, chlorine, and sometimes hydrogen are available.

Isotope separation. The next advance in the fluorine industry was connected with the development of the atomic bomb during World War II. It was necessary to separate the small proportion of the fissionable isotope uranium-235 from other, nonfissionable uranium isotopes. This separation could be done by diffusion, working with uranium hexafluoride, a gas. Fluorine at that time was made only occasionally on a small laboratory scale, and it had a reputation for intense chemical reactivity and for being difficult to handle. The solution to the problem of large-scale preparation of elemental fluorine, which required the development and introduction of novel, fluorine-resistant materials of construction, made this important element generally available. Fluorine manufacture is now routine. Other uses have been developed: as a component in some rocket propellants, for the preparation of the extremely reactive interhalogen compounds such as chlorine trifluoride (ClF_3), used for cutting steel, and for the preparation of sulfur hexafluoride, an extremely stable gas that has been employed as an insulator in electrical applications.

Other uses. Nonstick frying pans have been coated with a fluorocarbon resin, the best known of which is polytetrafluoroethylene. There are several other fluorocarbon and fluorinated hydrocarbon resins; some have highly specialized applications in the aerospace industry.

Nonstick coating for utensils

Fluorinated compounds are also used in textile treatments; some are soil-release agents that make fabric easy to wash. The salt sodium fluoroacetate is an extremely powerful rodenticide; it has been reported to give good control of rats, but it must be used with great care. Sodium bifluoride is used as a laundry sour; it also removes iron stains without weakening the fabric.

A minor but important use of fluorine in some countries is in the fluoridation of drinking water in the interest of dental health.

Bromine. The properties of bromine are significantly different from those of fluorine and chlorine, and it is far less abundant. Discovered in the early 19th century, in the form of its salts (bromides) in the bitterns remaining after evaporating seawater and extracting the sodium chloride, it was obtained later from Stassfurt, Germany, as a by-product in the production of potassium salts and from other salt deposits and salt lakes. Its main use was originally for bromides in medicine, still a minor use. Bromine first became of industrial importance with the development of the modern photographic process, in which the light-sensitive material is an emulsion of minute particles of silver bromide (together with silver chloride, or iodide, or both) in gelatin.

Tetraethyllead was another of Thomas Midgley's discoveries in the 1920s. Long the only effective agent in preventing "knock" in gasoline engines, tetraethyllead is now supplemented by tetramethyl lead, a similar compound. Although the knock problem was solved, a method was needed to get all traces of lead out of the engine cylinder. This removal was achieved by the addition of small quantities of a scavenger, ethylene dibromide, often in a mixture with ethylene dichloride.

Bromine obtained from brines

For a time the expanding world automobile industry threatened a scarcity of bromine, obtained from brines from the Great Lakes region and Searles Lake in the United States, and from the Dead Sea, which contains about 0.5 percent bromine. To meet the demand it was necessary to turn to seawater, which contains about 70 parts per million bromine.

To produce bromine from seawater, very large volumes of water must be processed. A preferable site for the operation is a neck of land projecting into the ocean so that water can be taken from one side and discharged to the other, avoiding the problem of processing the same water. The water is made acid with a little sulfuric acid and then treated with chlorine, which releases bromine from the bromides.

A current of air removes the bromine as a very dilute

mixture of bromine with air. The bromine is absorbed in sodium carbonate, after which treatment with sulfuric acid releases the bromine again in a much more concentrated form.

By far the greater part of the bromine produced is converted to ethylene dibromide by treatment with ethylene. Most of the ethylene dibromide is used in gasoline as a scavenger for lead; but it is also used as a fumigant, as a solvent for certain gums, and for further syntheses. The next most important bromine compound is methyl bromide, which is used as a fumigant, sometimes as a fire extinguisher, and for further syntheses.

Iodine. Iodine enters the chemical industry on a smaller scale. The largest producer is Japan, where iodine is obtained from seaweed. Seawater contains only about 0.05 part per million iodine, but some species of seaweed are able to concentrate this iodine manyfold, so that commercial extraction of the iodine is possible.

Medical uses of iodine

The most important industrial use of iodine compounds is the small amount of silver iodide used with silver bromide in photography. Iodine is important also in medicine (although this is not a large-scale use) in the treatment of certain thyroid conditions, and it is added to common table salt to prevent such conditions. It is also used directly as a disinfectant. Iodine is a component of a few useful dyes. The laboratory chemist frequently makes use of iodine or iodine compounds in synthesis and also in analysis. Crystalline silver iodide is useful in cloud seeding.

(A.St.)

ORGANIC CHEMICALS

The heavy chemical industry, in its classical form, was based on inorganic chemistry, concerned with all the elements except carbon and their compounds, but including, as has been seen, the carbonates. Similarly the light chemical industry uses organic chemistry, concerned with certain compounds of carbon such as the hydrocarbons, combinations of hydrogen and carbon. In the late 1960s the phrase heavy organic chemicals came into use for such compounds as benzene, phenol, ethylene, and vinyl chloride. Benzene and phenol are related chemically, and they are also related to toluene and the xylenes, which can be considered together as part of the aromatic group of organic chemicals, the aromatic compounds being most easily defined as those with chemical properties like those of benzene.

Aromatic hydrocarbons. *Benzene.* Chemically, the hydrocarbon benzene, which forms the basis of the aromatics, is a closed, six-sided ring structure of carbon atoms with a hydrogen atom at each corner of the hexagonal structure. Thus a benzene atom is made up of six carbon (C) atoms and six hydrogen (H) atoms, and has the chemical formula C_6H_6. Benzene has long been an industrial chemical. Initially it was obtained from the carbonization (heating) of coal, which produces coke, combustible gas, and a number of by-products, including benzene. Carbonization of coal to produce illuminating gas dates back in England to the very early years of the 19th century. The process is still employed in some countries, but more and more use is being made of natural gas. The carbonizing process is also used (with slight modifications) to produce metallurgical coke, indispensable for the manufacture of iron and hence steel. The supply of benzene from the carbonizing process, however, is not sufficient to meet the demand. For every ton of coal carbonized only about two to three pounds (0.9 to 1.35 kilograms) of benzene are obtained.

The shortage of aromatics first became evident during World War I, when toluene was in great demand for the manufacture of trinitrotoluene, or TNT, the principal explosive used then. Methods were worked out to obtain toluene from petroleum. Much later, after World War II, benzene and all the other aromatics derived from it were needed in far greater quantities than metallurgical coke could supply, and by far the greater part of these aromatics now comes from petroleum.

Toluene. Toluene differs from benzene in that one of the hydrogen atoms is replaced by a special combination of carbon and hydrogen called a methyl group ($-CH_3$). The xylenes have two methyl groups in different positions in the benzene ring, and thus all aromatics are to some extent interchangeable. In fact, one of the uses for toluene is to produce benzene by removing the methyl group (dealkylation).

All of these hydrocarbons are useful as gasoline additives because of their antiknock properties.

Uses of solvents

Toluene is also used as a solvent. The expression "as a solvent," which occurs frequently in describing the uses for chemicals, covers a multitude of applications. The substance dissolved is usually also organic, and the process is used in coatings, adhesives, textiles, pharmaceuticals, inks, photographic film, and metal degreasing. An application that reaches the ultimate consumer is dry cleaning (although the solvent used here is not toluene, but other hydrocarbons or chlorohydrocarbons). Toluene has a multitude of other uses, such as in the polyurethane plastics and elastomers discussed below.

Xylene. The three isomeric xylenes (isomeric means that they have exactly the same number and kind of atoms but are arranged differently) occur together, and with them is another isomer, ethylbenzene, which has one ethyl group ($-C_2H_5$) replacing one of the hydrogen atoms of benzene. These isomers can be separated only with difficulty, but numerous separation methods have been worked out. The small letters *o-*, *m-*, and *p-* (standing for *ortho-*, *meta-*, and *para-*) preceding the name xylene are used to identify the three different isomers that vary in the ways the two methyl groups displace the hydrogen atoms of benzene. *Ortho*-xylene is used mostly to produce phthalic anhydride, an important intermediate that leads principally to various coatings and plastics. The least valued of the isomers is *meta*-xylene, but it has uses in the manufacture of coatings and plastics. *Para*-xylene leads to polyesters, which reach the ultimate consumer as polyester fibres under various trademarked names.

Benzene itself is perhaps the industrial chemical with the most varied uses of all. Figure 2 shows some in outline form; for example, several routes are shown to phenol, itself an important industrial chemical. In transforming benzene to the products obtained from it, other raw materials are required; for example, ethylene for the production of styrene, and sulfuric acid for the production of benzenesulfonic acid. It would have unduly complicated Figure 2 to attempt to show all of these; chlorine, however, has been shown entering at several places. Chlorine will be encountered in many operations discussed below.

The diagram of Figure 2 is drastically simplified. Many applications of benzene are not shown. In some cases,

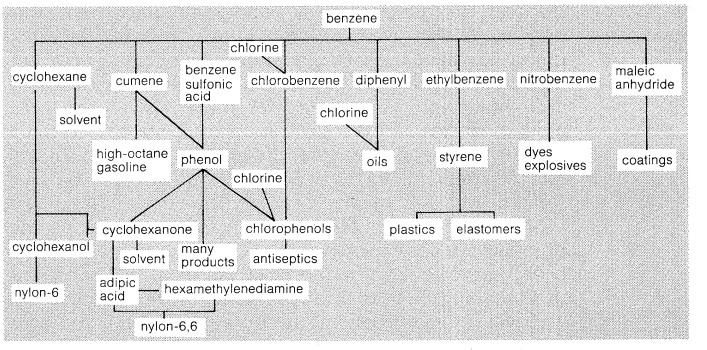


Figure 2: Intermediate and end products from raw benzene.

alternative starting points to the end product sidetrack benzene. For example, to obtain styrene from benzene the route passes through ethylbenzene; but ethylbenzene is found in a mixture with its isomers, the xylenes; the ethylbenzene that is separated from the xylene mixture can be used in the manufacture of styrene.

Figure 2 shows synthetic fibres (two kinds of nylon); coatings, plastics, and elastomers (synthetic products having rubberlike properties) are also mentioned. All these groups of substances have one thing in common—they are all polymers (substances composed of large molecules formed from smaller ones), produced by applications of a rapidly growing branch of chemistry, polymer chemistry, established in the early 1930s. The chemical process in-

dustries based upon polymers are covered in the major sections *Man-made fibres; Plastics and resins;* and *Paints and varnishes.* (See also the section *Polymers* in the article CHEMICAL COMPOUNDS.)

Aliphatic hydrocarbons. Because of the complicated interlocking network of the chemical industry, it will be helpful to return briefly to the original raw materials. Earlier the aromatic group of organic chemicals was described; contrasted with these are the aliphatics, of which a number of quite simple chemicals are of great industrial importance.

The simplest organic chemicals are the saturated hydrocarbons methane (CH_4), ethane (C_2H_6 or $H_3C - CH_3$), propane ($H_3C - CH_2 - CH_3$), and others. These are useful as fuels but are chemically rather unreactive, and so in order to process them to give further chemicals, they are "cracked" by a heat treatment to convert them to unsaturated hydrocarbons. These contain less hydrogen than the saturated hydrocarbons, and they contain one or more double valence bonds, or triple valence bonds, connecting carbon atoms. Some of the most important unsaturated hydrocarbons industrially are acetylene ($HC \equiv CH$), ethylene ($H_2C = CH_2$), propylene ($H_3C - CH = CH_2$), and butadiene ($H_2C = CH - CH = CH_2$). An idea of the raw materials for these hydrocarbons and a highly simplified diagram of their products are given in Figure 3.

Conversion to unsaturated hydrocarbons

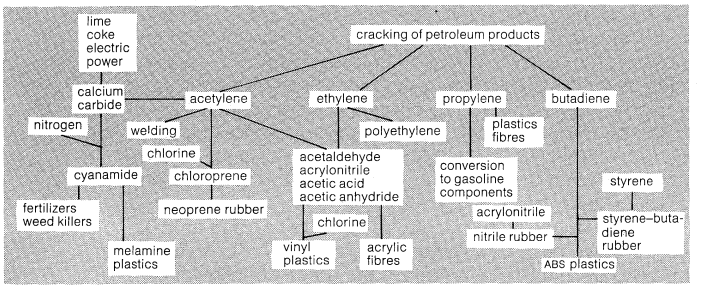


Figure 3: Some of the more important unsaturated hydrocarbons; their origin and final products.

Ethylene. Ethylene, one of the largest volume organic chemicals, can be produced either together with acetylene or with propylene. It gives rise to a large number of products, many in large volume. Some of the more important have been lumped together in a box (Figure 3): acetaldehyde, acrylonitrile, acetic acid, acetic anhydride, the list bringing together substances that have complex interrelations. These relations would come to light if this box were magnified and examined closely. These substances, however, can in general also be made from acetylene, and acetylene can also be made from a completely different source, calcium carbide.

Acetylene. The raw materials for calcium carbide are shown in Figure 3 as lime, coke, and electric power. Thus calcium carbide is a more suitable source of acetylene in a country that has hydroelectric power but lacks petroleum reserves. The largest producer of acetylene is Japan; Poland, the Soviet Union, and many other countries are also notable producers.

Calcium carbide generates acetylene when acted upon by water. This process can be a small-scale one to give acetylene suitable for illumination because of its extremely bright flame. Acetylene is also made on a large scale for chemical conversion, as shown in Figure 3. Acetylene is also used for oxyacetylene welding because when burned with oxygen it produces an extremely high temperature.

Acetylene and ethylene have been in competition for chemical industrial uses. In the 1950s, acetylene was widely used as a chemical raw material, and methods were worked out for obtaining it from hydrocarbon sources, as shown in Figure 3. Later ethylene became in general more economical, and the use of acetylene as a raw material has been declining. Calcium carbide, a raw material for acetylene, however, has other uses. When treated with nitrogen, it gives calcium cyanamide, valuable as a fertilizer and weed killer, and at the same time a raw material for the production of melamine, used in making some modern plastics (see on the left in Figure 3). Other products from acetylene, ethylene, and other unsaturated hydrocarbons marked, in their main outlines, in Figure 3 show that these processes provide a wide variety of raw materials for various plastic, elastic, and fibrous products.

Competition between acetylene and ethylene

Propylene. Propylene is not produced in as large volume as ethylene and is mostly used chemically. It is an important raw material for certain detergents. It leads to derivatives that are used in antiknock gasoline additives. It can also be polymerized to a product with uses generally similar to those of polyethylene. When made into a fibre, polypropylene is especially useful for carpets.

Elastomers. Butadiene (Figure 3) is used to produce plastics and elastomers, a group of substances related to plastics. The elastomers were at first thought of as synthetic substitutes for natural rubber. As has often happened with synthetic substitutes, however, a number of different varieties were developed; some were actually better than natural rubber in some ways and others better in other ways, and so it was soon realized that what was being developed was not so much a replacement as a supplement.

Interest in a synthetic material that could be used in automobile tires began in Germany as early as World War I, when supplies from the tropical, rubber-producing countries were cut off. A synthetic rubber of a sort was produced that could be used for tires, although the vehicle had to be jacked up when not in motion to prevent developing a flat spot on the tires. Much research in Germany and the United States led to the development, shortly before World War II, of several elastomers. The most important of these, and by far the best for tires, was made of a copolymer of 75 parts of butadiene and 25 parts of styrene. This synthetic was first known as GR-S (Government Rubber–Styrene) but later came to be called SBR—styrene–butadiene rubber. It is produced in far greater quantity than any of the other synthetics. It is better than natural rubber in some respects, but poorer in others. It is often used in blends with other rubbers.

Figure 3 also shows that acrylonitrile can be copolymerized with butadiene (roughly one-third acrylonitrile, two-thirds butadiene) to form nitrile rubber (NBR). This synthetic has different properties from other synthetics, and is used for rubber hose, tank lining, conveyor belts, gaskets, and wire insulation. Acrylonitrile and styrene, together with butadiene, form a terpolymer, called ABS, which is useful for high-impact-strength plastics.

Acrylonitrile contains nitrogen, and therefore is decidedly different in chemical constitution from natural rubber, which contains only carbon and hydrogen. Natural rubber has a repeating unit of five carbon atoms. By starting with the unsaturated hydrocarbon isoprene (C_5H_8), a polymer can be made with the spatial arrangement of the atoms the same as in natural rubber and with very similar properties. This polymer is sometimes referred to as synthetic natural rubber. Another hydrocarbon elastomer starts with isobutylene (C_4H_8) and gives butyl, a rubber characterized by resistance to oxygen and impermeability to gases, which is used widely in cable insulation and as a coating for fabrics.

Impermeability of butyl rubber

Figure 3 shows that acetylene is the raw material for chloroprene (C_4H_5Cl), which is converted into neoprene, another versatile elastomer of exceptional properties. There are also rubberlike products containing sulfur, known in the United States as the thiokols. A related group, containing carbon, sulfur, and oxygen, the sulfones, are tough plastic materials. Elastomeric materials are more fully treated in the major section *Rubber* below.

Film materials. Most of the above-mentioned groups of chemicals that can be used either as plastics or elastomers can also be made into the form of coherent films. In the more highly industrialized countries there is a very high demand for films for wrapping purposes, largely for food, and also in the building construction industry. The requirements for a film vary greatly. For many food products the wrapping film must have the ability to "breathe"; that is, it must have some permeability to water vapour and also to oxygen. Films can be developed with high permeability, or with none at all. In some applications the film should be self-sealing. Films can be made of any

thickness, and for some purposes extreme toughness is required. Paper, or treated paper, has of course been used for many of these purposes for many years, but it has such disadvantages as low strength, particularly when wet, and it is difficult to make it transparent. Cellophane was produced commercially starting in the 1920s; its transparency attracted attention at once, beginning a revolution in wrapping materials.

Cellophane is regenerated cellulose. It is like viscose rayon, except that it is extruded flat, instead of in the form of a fibre. It is still very popular, but is highly sensitive to water and to changing humidity. Many other polymers now supplement it and compete with it. Polyethylene makes fine, tough films; there is no sharp distinction between a thin extrusion, useful for a wrapping film, and thicker products used for nonbreakable bottles. Many vinyl products are used in films, as are polystyrene, polyesters, and nylon. A chemical derivative from natural rubber, chlorinated rubber, gives films of extraordinary stretchability.

From coherent films that can stand by themselves, it is a short step to one of the components of a paint. In the days before chemical technology, commercial paints were based on linseed oil as a film-former. Linseed oil and the pigment made a mixture that was too thick, so that it was normally thinned with turpentine.

The thinner in paint is the component that has undergone least change. Turpentine, obtained from pine trees, and sometimes as a by-product in the manufacture of paper, is still used. A petroleum distillate, however, is equally effective. The thinner completely evaporates very shortly after the paint is applied. In latex paints, the paint itself is in the form of minute droplets in water, and water is the thinner. (see the major section *Paints and varnishes* below.)

Carbon black. The outstanding black pigment is the versatile product known as carbon black. Carbon black is one of the most important industrial chemical products. Carbon black is considered a petrochemical because it is made from natural gas or petroleum residues. There are several processes involving either incomplete combustion (burning off the hydrogen of a hydrocarbon, such as methane, and leaving the carbon) or by externally applied heat in a furnace, splitting the hydrocarbon into hydrogen and carbon.

Major use of carbon black

The most important of all the uses of carbon black is in compounding rubber to be used in tires. An average tire of a passenger automobile contains about four pounds of carbon black. Carbon black is not only used as a pigment but also is employed in printing ink, an ink being little different from an applied coating. Carbon black creates the principal difficulty in recycling newsprint because no practical way has been found to destroy the black ink. A specialized use of carbon black is as an additive to phonograph records. A special form of carbon black, derived from acetylene, has its principal use in electrochemical dry cells.

Alcohols and their derivatives. *Methanol.* The important product methanol (Figure 1) is obtained from synthesis gas in the form of carbon monoxide and hydrogen (sometimes carbon dioxide and hydrogen). The terms methyl alcohol and methanol are synonymous, the former being used more in Great Britain and the latter expression universal in United States industry. The term wood alcohol, sometimes employed, refers to the fact that this alcohol was formerly obtained by the distillation of wood.

Methanol is a large-volume chemical; about half of the production goes to making formaldehyde (CH_2O), a very reactive chemical with a large number of uses. A small amount of formaldehyde comes from non-methanol sources, via the direct oxidation of hydrocarbons. Methanol also enters into the production of various plastics; leads to such useful derivatives as methyl chloride, a solvent for inks and dyes; and is used in the purification of steroidal and hormonal medicines.

Formaldehyde. The greatest uses of formaldehyde are in the formation of important groups of plastics, the urea–formaldehyde resins and the phenol–formaldehyde resins. In addition, it is used as a fungicide and as a preservative, in paper and textile treatments, and in the synthesis of further products.

Ethanol and its products. Methanol (CH_3OH) is the simplest of the alcohols. The next number of the series, called either ethanol or ethyl alcohol (CH_3CH_2OH), has two carbon atoms. It is most familiar as the active constituent of fermented beverages, but it is also widely used in industry. When intended for human consumption ethyl alcohol is always produced by fermentation of some suitable material to form beer, or wine, or distilled spirits of various kinds. For industrial use it is sometimes produced by fermentation from some cheap material, such as molasses, but more often is made from ethylene by causing it to combine with water under the influence of a catalyst, which may be sulfuric acid or phosphoric acid.

A major industrial use of ethanol is to convert it by oxidation into acetaldehyde (CH_3CHO). Ethanol could have been shown in Figure 3 between ethylene and the block containing acetaldehyde and several related chemicals. Ethanol is also used in the preparation of various derivatives, such as ethyl chloride (used in the production of tetraethyllead), in the course of making various plastics, and in the usual further syntheses.

Acetaldehyde made from ethanol is generally used in the next step by the same company, most often in the same plant, so that the ethanol is really an intermediate that is used at once. For other uses, ethanol is often shipped from one plant to another. Alcohol intended for human consumption, however, is in all countries subject to a tax, which would make the cost of ethanol prohibitive for any industrial use. Industrial alcohol, therefore, is denatured by the addition of small amounts of substances that are carefully chosen to be highly unpleasant in taste and hard to remove but that do not interfere with the intended industrial use.

Other alcohols. In the alcohols with three carbon atoms, there are two possible structures, or isomers. One is called *n*-propyl alcohol (or 1-propanol), the other isopropyl alcohol (or 2-propanol).

The alcohol 1-propanol, not manufactured in very large quantities, finds major use in printing inks. The alcohol 2-propanol, on the other hand, is manufactured on the million-ton scale. It is made from propylene by a process similar to that used to convert ethylene to ethanol, and manufacture of 2-propanol by this process initiated the petrochemical industry in the 1920s.

The principal use of 2-propanol is in the manufacture of acetone, which is used extensively as a solvent and as a starting material in the manufacture of numerous other organic compounds. Smaller amounts of 2-propanol are converted to other chemical products or used as a solvent, as rubbing alcohol, or as a denaturing agent for ethyl alcohol.

Manufacture of higher alcohols

Higher alcohols—that is, with more than three carbon atoms—too numerous to detail here, are also manufactured. Mention should also be made of the dihydric alcohol, ethylene glycol. This chemical is produced in large volume, and is made from ethylene by an indirect route. Its principal use is in antifreeze mixtures for automobile radiators. It is also used in brake fluids and has numerous derivatives used in resins, paints, and explosives and in the manufacture of polyester fibres. Similar reactions for propylene give propylene glycol, the principal use of which is as a moistening agent in foods and tobacco.

Soaps and detergents

Soaps and detergents are substances that, when dissolved in water, give it the ability to remove dirt from surfaces such as the human skin, textiles, and other solids. The seemingly simple process of cleaning a soiled surface is, in fact, complex, and consists of the following physical-chemical steps:

The chemical steps in cleaning

1. Wetting of the surface and, in the case of textiles, penetration of the fibre structure by wash liquor containing the detergent. Detergents (and other surface-active agents) increase the spreading and wetting ability of water by reducing its surface tension, that is, the affinity its molecules

have for each other in preference to the molecules of the material to be washed.

2. Absorption of a layer of the soap or detergent at the interfaces between the water and the surface to be washed and between the water and the soil. In the case of ionic surface-active agents (explained below), the layer formed is ionic (electrically polar) in nature.

3. Dispersion of soil from the fibre or other material into the wash water. This step is facilitated by mechanical agitation and high temperature; in the case of toilet soap, soil is dispersed in the foam formed by mechanical action of the hands.

4. Preventing the soil from being deposited again onto the surface cleaned. The soap or detergent accomplishes this by suspending the dirt in a protective colloid, sometimes with the aid of special additives. In a great many soiled surfaces the dirt is bound to the surface by a thin film of oil or grease. The cleaning of such surfaces involves the displacement of this film by the detergent solution, which is in turn washed away by rinse waters. The oil film breaks up and separates into individual droplets under the influence of the detergent solution. Proteinic stains, such as egg, milk, and blood, are difficult to remove by detergent action alone. The proteinic stain is nonsoluble in water, adheres strongly to the fibre, and prevents the penetration of the detergent. By using proteolytic enzymes (enzymes able to break down proteins) together with detergents, the proteinic substance can be made water-soluble or at least water-permeable, permitting the detergent to act and the proteinic stain to be dispersed together with the oily dirt. The enzymes may present a toxic hazard to some persons habitually exposed.

Enzymes

If detached oil droplets and dirt particles did not become suspended in the detergent solution in a stable and highly dispersed condition, they would be inclined to flocculate or coalesce into aggregates large enough to be redeposited on the cleansed surface. In the washing of fabrics and similar materials, small oil droplets or fine, deflocculated dirt particles are more easily carried through interstices in the material than are relatively large ones. The action of the detergent in maintaining the dirt in a highly dispersed condition is therefore important in preventing retention of detached dirt by the fabric.

In order to perform as detergents (surface-active agents), soaps and detergents must have certain chemical structures: their molecules must contain a hydrophobic (water-insoluble) part, such as a fatty acid or a rather long chain carbon group, such as fatty alcohols or alkylbenzene. The molecule must also contain a hydrophilic (water-soluble) group, such as $-COONa$, or a sulfo group, such as $-OSO_3Na$ or $-SO_3Na$ (such as in fatty alcohol sulfate or alkylbenzene sulfonate), or a long ethylene oxide chain in nonionic synthetic detergents. This hydrophilic part makes the molecule soluble in water. In general, the hydrophobic part of the molecule attaches itself to the solid or fibre and onto the soil, and the hydrophilic part attaches itself to the water.

Types of surface-active agents

Four groups of surface-active agents are distinguished:

1. Anionic detergents (including soap and the largest portion of modern synthetic detergents), which produce electrically negative colloidal ions in solution.
2. Cationic detergents, which produce electrically positive ions in solution.
3. Nonionic detergents, which produce electrically neutral colloidal particles in solution.
4. Ampholytic, or amphoteric, detergents which are capable of acting either as anionic or cationic detergents in solution depending on the pH (acidity or alkalinity) of the solution.

The first detergent (or surface-active agent) was soap. In a strictly chemical sense, any compound formed by the reaction of a water-insoluble fatty acid with an organic base or an alkali metal may be called a soap. Practically, however, the soap industry is concerned mainly with those water-soluble soaps that result from the interaction between fatty acids and alkali metals. In certain cases, however, the salts of fatty acids with ammonia or with triethanolamine are also used, as in shaving preparations, for instance.

HISTORY

Use. Soap has been known for at least 2,300 years. According to Pliny the Elder, the Phoenicians prepared it from goat's tallow and wood ashes in 600 BC and sometimes used it as an article of barter with the Gauls. Soap was widely known in the Roman Empire; whether the Romans learned its use and manufacture from ancient Mediterranean peoples or from the Celts, inhabitants of Britannia, is not known. The Celts, who produced their soap from animal fats and plant ashes, named the product *saipo*, from which the word soap is derived. The importance of soap for washing and cleaning was apparently not recognized until the 2nd century after Christ; the Greek physician Galen mentions it as a medicament and as a means of cleansing the body. Previously soap had been used as medicine. The writings attributed to the 8th-century Arab savant Jābir ibn Hayyān (Geber) repeatedly mention soap as a cleansing agent.

In Europe, soap production in the Middle Ages centred first at Marseilles, later at Genoa, then at Venice. Though some soap manufacture developed in Germany, the substance was so little used in central Europe that a box of soap presented to the Duchess of Juelich in 1549 caused a sensation. As late as 1672, when the German, A. Leo, sent Lady von Schleinitz a parcel containing soap from Italy, he accompanied it with a detailed description of how to use the mysterious product.

The first English soapmakers appeared at the end of the 12th century in Bristol. In the 13th and 14th centuries, a small community of them grew up in the neighbourhood of Cheapside in London. In those days, soapmakers had to pay a duty on all the soap they produced. After the Napoleonic Wars, this tax rose as high as three pence per pound; soap-boiling pans were fitted with lids that could be locked every night by the tax collector in order to prevent production under cover of darkness. Not until 1853 was this high tax finally abolished, at a sacrifice to the state of over £1,000,000. Soap came into such common use in the 19th century that Justus von Liebig, a German chemist, declared the quantity of soap consumed by a nation was an accurate measure of its wealth and civilization.

Early soap production. Early soapmakers probably used ashes and animal fats. Simple wood or plant ashes containing potassium carbonate were dispersed in water, and fat was added to the solution (see Figure 4). This mixture was then boiled; ashes were added again and again as the water evaporated. During this process a slow chemical splitting of the neutral fat took place; the fatty acids could then react with the alkali carbonates of the plant ash to form soap (this reaction is called saponification).

By courtesy of CIBA Review, Basel, Switzerland

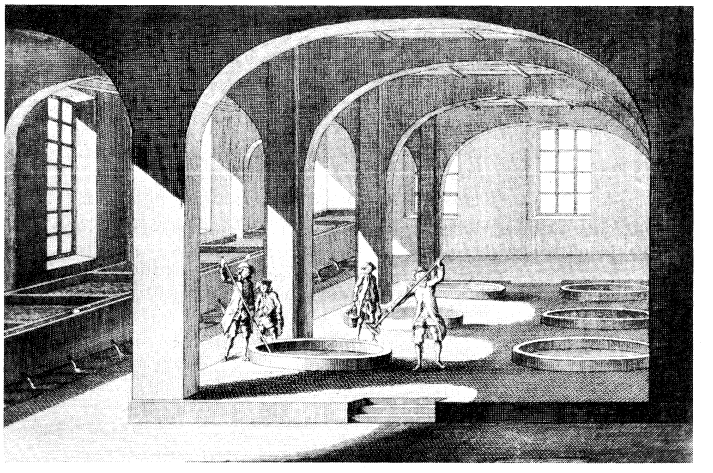


Figure 4: A French soap-boiling plant with the vessels for lye (far left) and the circular boiling pans; engraving published in Paris, 1771.

Animal fats containing a percentage of free fatty acids were used by the Celts. The presence of free fatty acids certainly helped to get the process started. This method probably prevailed until the end of the Middle Ages, when slaked lime came to be used to causticize the alkali carbon-

ate. Through this process, chemically neutral fats could be saponified easily with the caustic lye. The production of soap from a handicraft to an industry was helped by the introduction of the Leblanc process for the production of soda ash from brine (around 1790) and by the work of a French chemist, Eugène-Michel Chevreul, who in 1823 showed that the process of saponification is the chemical process of splitting fat into the alkali salt of fatty acids (that is, soap) and glycerin.

The method of producing soap by boiling with open steam, introduced at the end of the 19th century, was another step toward industrialization.

Early synthetic detergents. If turkey-red oil, *i.e.*, sulfated castor oil, still used in textile and leather industries today, is considered the first synthetic detergent, the industry began in the midst of the last century. The first synthetic detergents for general use, however, were produced by the Germans in the World War I period so that available fats could be utilized for other purposes. These detergents were chemicals of the short-chain alkylnaphthalene-sulfonate type, made by coupling propyl or butyl alcohols with naphthalene and subsequent sulfonation, and appeared under the name of Nekal. These products were only fair detergents, but good wetting agents, and are still being produced in large quantities for use in the textile industry.

In the late 1920s and early '30s, molecules consisting of long chain alcohols were sulfonated and sold as the neutralized sodium salts without any further additions except for sodium sulfate as an extender. In the early '30s molecules consisting of long chain alkylaryl sulfonates (with benzene as the aromatic nucleus and the alkyl portion made from a kerosene fraction) appeared on the market in the United States. Again, these were available as the sodium salts extended with sodium sulfates. Both the alcohol sulfates and the alkylaryl sulfonates were sold as cleaning materials but did not make any appreciable impression on the total market. By the end of World War II the alkylaryl sulfonates had almost completely swamped the sales of alcohol sulfates for the limited uses to which they were applied as general cleaning materials, but the alcohol sulfates were making big inroads into the shampoo and fine detergent fields.

Historically, synthetic detergents began as mainly a substitute for fat-based soap but developed into a sophisticated product, superior in many respects to soap.

Advantages of synthetic detergents over soap

Soap forms a scum or precipitate in hard water, leaving a ring around the bathtub, a whitish residue on glassware, and a sticky curd in the rinse water of the laundry tub. Not so easily perceived is the relation of this hard-water scum to a dull, lustreless condition of hair after shampooing, yellow spots on laundry after ironing, and a heavy usage of soap in the household. All these effects point to a serious defect of soaps, namely, their reaction with the calcium and other metal salts present in hard water to give a precipitate that constitutes the hardness of the water. Soaps also react with traces of acidic compounds to form a precipitate. On the other hand, the synthetic detergents generally are unaffected or very little affected by metal salts or acids; although they may react chemically with them, the resulting compounds are either soluble or remain dispersed in colloidal form in the solution. Other useful properties of the synthetics, such as solubility in cold water and flexibility in formulation, also contributed to their rapid replacement of soap products.

SOAP MANUFACTURING PROCESSES AND PRODUCTS

Hot caustic alkali solution, such as caustic soda (sodium hydroxide), acts on natural fats or oils, such as tallow or vegetable oil, to produce sodium fatty acid salt (soap) and glycerin (or glycerol). This saponification reaction is the basis for all soapmaking. If industrially produced fatty acids are used in place of natural fats or oils, the reaction with caustic soda yields soap and water instead of soap and glycerin.

Raw materials and additives. The major raw materials for soap manufacture are fat and alkali. Other substances, such as optical brighteners, water softeners, and abrasives are often added to obtain specific characteristics.

Alkali. Sodium hydroxide is employed as the saponification alkali for most soap now produced. Soap may also be manufactured with potassium hydroxide (caustic potash) as the alkali. Potassium soaps are more soluble in water than sodium soaps; in concentrated form, they are called soft soap. Though soft soaps are declining in importance, potassium soap is still produced in various liquid concentrations for use in combination with sodium soaps in shaving products and in the textile industry.

Certain alkaline materials (builders) are almost universally present in laundry soaps, functioning to increase detergency. The most important are sodium silicate (water glass), sodium carbonate (soda ash), sodium perborate, and various phosphates.

Fats and oils. Fatty raw materials for soap manufacture include animal and vegetable oils and fats or fatty acids, as well as by-products of the cellulose and paper industry, such as rosin and tall oil. Four groups of these raw materials can be distinguished according to the properties of the soap products they yield:

1. Hard fats yielding slow lathering soaps include tallow, garbage greases, hydrogenated high-melting-point marine and vegetable oils, and palm oil. These fats yield soaps that produce little lather in cold water, more in warm water; are mild on the skin; and cleanse well. This is the leading group of fats used in the international soap industry, with tallow the most important member.
2. Hard fats yielding quick-lathering soaps include coconut oil, palm-kernel oil, and babassu oil. (Palm-kernel oil is extracted from the kernel of the fruit of the oil palm, whereas palm oil, listed above in 1, is expressed from the pericarp or outer fleshy portion of the fruit.) These fats are not very sensitive to electrolytes, such as salt; thus, they are suitable for manufacture of marine soap, which must lather in seawater. This is the second most important group of soap fats, with coconut oil the most used.
3. Oils yielding soaps of soft consistency, such as olive oil, soybean oil, and groundnut (peanut) oil are most important here, and linseed and whale oils also belong to the group, as do some semi-drying or drying oils. Because these oils readily undergo changes in air or light or during storage, soaps made from them may become rancid and discoloured.
4. Rosin and tall oil (a resinous by-product of the manufacture of chemical wood pulp) form a group in themselves. Rosin is used in laundry soap, less expensive toilet soaps, and specialty soaps in various industries. Tall oil is mainly used in liquid soap.

Optical brighteners. Now an integral part of all washing powders, optical brighteners are dyestuffs absorbed by textile fibres from solution but not subsequently removed in rinsing. They convert invisible ultraviolet light into visible light on the blue side of the spectrum, causing the fibre to reflect a greater proportion of visible light and making it appear brighter. Furthermore, since the tone of the extra light reflected is on the blue side of the spectrum, this blue-violet tinge will complement any yellowishness present on the fibre to make it look whiter as well as brighter. The chemical structures of optical brightening agents are complicated; many formulas are trade secrets.

Laundry blueing

Though the action of optical brighteners resembles old-style laundry blueing in some ways, the two methods must not be confused. In the old method, a blue dye or pigment is adsorbed onto the fibre; this blue tends to absorb yellow light falling on it, reflecting light richer in blue. With blueing, however, the fabric absorbs some of the light falling on it and hence reflects less light than it receives. Thus, the fabric looks whiter, not brighter.

Sequestering or chelating agents. EDTA (ethylenediamine tetraacetic acid) or its sodium salt has the property of combining with certain metal ions to form a molecular complex that locks up or chelates the calcium ion so that it no longer exhibits ionic properties. In hard water, calcium and magnesium ions are thus inactivated, and the water is effectively softened. EDTA can form similar complexes with other metallic ions.

Abrasives. Water-insoluble minerals such as talc, diatomaceous earth, silica, marble, volcanic ash (pumice), chalk, feldspar, quartz, and sand are often powdered and added to soap or synthetic detergent formulations. Abra-

sives of an organic nature, such as sawdust, are also used.

Soap production processes. Several techniques are employed in making soap, most of which involve heat. Processes can be either continuous or on a batch basis.

Boiling process. Still widely used by small and medium-sized producers is the classical boiling process. Its object is to produce neat soap in purified condition, free from glycerin. Neat soap is the starting material for making bars, flakes, beads, and powders. The boiling process is conducted in a series of steps called changes; these occur in the kettle (called the pan in Great Britain).

In the first step, melted fats are placed in the kettle, and caustic soda solution is added gradually. The whole mass is then boiled with open steam from perforated coils within the kettle. The saponification reaction now takes place; the mass gradually thickens or emulsifies as the caustic soda reacts with the fat to produce both soap and glycerin.

Salting out

To separate the glycerin from the soap, the pasty boiling mass is treated with brine. Contents of the kettle salt out, or separate, into an upper layer that is a curdy mass of impure soap, and a lower layer, that consists of an aqueous salt solution with the glycerin dissolved in it. Thus the basis of glycerin removal is the solubility of glycerin and the insolubility of soap in salt solution. The slightly alkaline salt solution, termed spent lye, is extracted from the bottom of the pan or kettle and subsequently treated for glycerin recovery.

The grainy, curdy mass of soap remaining in the kettle after the spent lye has been removed contains any unsaponified fat (usually traces that escaped reaction during saponification) plus dirt and colouring matter present in the original oils. During the next step, called strong change, strong caustic solution is added to the mass, which is then boiled to remove the last of the free fat.

The final stage, called pitching and settling, transforms the mass into neat soap and removes dirt and colouring matter. After the strong change, the soap may be given one or more salt-water washes to remove free alkali, or it may be pitched directly. Pitching involves boiling the mass with added water until a concentration is attained that causes the kettle contents to separate into two layers. The upper layer is neat soap, sometimes called kettle soap, of almost constant composition for a given fat (about 70 percent soap, 30 percent water); the lower, called nigre, varies in soap content from 15 percent to 40 percent. Since colouring matter, dirt, salt, alkali, and metal soaps are soluble in nigre but relatively insoluble in neat soap, and since most of the impurities are dense and tend to settle, the nigre layer takes these from the neat soap.

Continuous soapmaking—the hydrolyzer process. The boiling process is very time consuming; settling takes days. To produce soap in quantity, huge kettles must be used. For this reason, continuous soapmaking has largely replaced the old boiling process. Most continuous processes today employ fatty acids in the saponification reaction in preference to natural fats and oils. These acids do not contain impurities and, as explained at the beginning of this section, produce water instead of glycerin when they react with alkali. Hence, it is not necessary to remove impurities or glycerin from soap produced with fatty acids. Furthermore, control of the entire process is easier and more precise. The fatty acids are proportionally fed into the saponification system either by flowmeter or by metering pump; final adjustment of the mixture is usually made by use of a pH meter (to test acidity and alkalinity) and conductivity-measuring instruments.

The continuous hydrolyzer process begins with natural fat that is split into fatty acids and glycerin by means of water at high temperature and pressure in the presence of a catalyst, zinc soap. The splitting reaction is carried on continuously, usually in a vertical column 50 feet (15 metres) or more in height. Molten fat and water are introduced continuously into opposite ends of the column; fatty acids and glycerin are simultaneously withdrawn. Next, the fatty acids are distilled under vacuum to effect purification. They are then neutralized with an alkali solution such as sodium hydroxide (caustic soda) to yield neat soap. In toilet-soap manufacture, a surplus of free fatty acid, often in combination with such superfatting agents

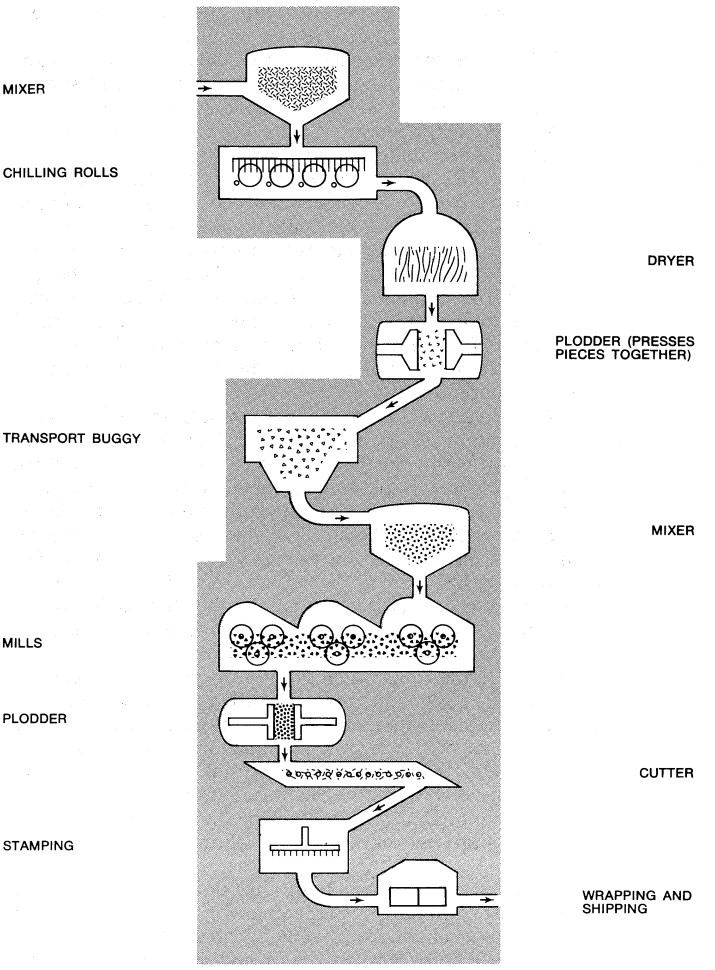


Figure 5: Steps in the manufacture of milled bar soap.
Drawing by D. Meighan based on N. Shreve, *Chemical Process Industries*, 3rd ed. (1967); McGraw-Hill Book Co.

as olive oil or coconut oil, is left or added at the final stage so that there is no danger of too much alkali in the final product. The entire hydrolyzer process, from natural fat to final marketable product, requires a few hours, as compared with the four to 11 days necessary for the old boiling process. The by-product glycerol is purified and concentrated as the fatty acid is being produced.

Cold and semiboiled methods. In the cold method, a fat and oil mixture, often containing a high percentage of coconut or palm-kernel oil, is mixed with the alkali solution. Slightly less alkali is used than theoretically required in order to leave a small amount of unsaponified fat or oil as a superfatting agent in the finished soap. The mass is mixed and agitated in an open pan until it begins to thicken. Then it is poured into frames and left there to saponify and solidify.

In the semiboiled method, the fat is placed in the kettle and alkali solution is added while the mixture is stirred and heated but not boiled. The mass saponifies in the kettle and is poured from there into frames, where it solidifies. Because these methods are technically simple and because they require very little investment for machinery, they are ideal for small factories.

Finishing operations. Finishing operations transform the hot mass coming from the boiling pan or from continuous production equipment into the end product desired. For laundry soap, the soap mass is cooled in frames or cooling presses, cut to size, and stamped. If soap flakes, usually transparent and very thin, are to be the final product, the soap mass is extruded into ribbons, dried, and cut to size. For toilet soap, the mass is treated with perfumes, colours, or superfatting agents, is vacuum dried, then cooled and solidified. The dried solidified soap is homogenized (often by milling or crushing) in stages to produce various degrees of fineness (see Figure 5). Air can

be introduced under pressure into the warm soap mass as it leaves the vacuum drier to produce a floating soap. Medicated soaps are usually toilet soaps with special additives, chlorinated phenol, xylenol derivatives, and similar compounds, added to give a deodorant and disinfectant effect. As mentioned above, shaving creams are based on potassium and sodium soap combinations.

Anionic detergents. Among synthetic detergents, commonly referred to as syndets, anionic-active types are the most important. The molecule of an anionic-active synthetic detergent is a long carbon chain to which a sulfo group ($-SO_3$) is attached, forming the negatively charged (anionic) part. This carbon chain must be so structured that a sulfo group can be attached easily by industrial processes (sulfonation), which may employ sulfuric acid, oleum (fuming sulfuric acid), gaseous sulfur trioxide, or chlorosulfonic acid.

Raw materials. Fatty alcohols are important raw materials for anionic synthetic detergents. Development of commercially feasible methods in the 1930s for obtaining these provided a great impetus to synthetic-detergent production. The first fatty alcohols used in production of synthetic detergents were derived from body oil of the sperm or bottlenosed whale (sperm oil). Efforts soon followed to derive these materials from the less expensive triglycerides (coconut and palm-kernel oils, and tallow). The first such process, the Bouveault-Blanc method of 1903, long used in laboratories, employed metallic sodium; it became commercially feasible in the 1950s when sodium prices fell to acceptable levels. When the chemical processing industry developed high-pressure hydrogenation and oil-hardening processes for natural oils, detergent manufacturers began to adopt these methods for reduction of coconut oil, palm-kernel oil, and other oils into fatty alcohols. Synthetic fatty alcohols have been produced from ethylene; the process, known as the Alfol process, employs diethylaluminum hydride.

Soon after World War II, another raw material, alkylbenzene, became available in huge quantities. Today it is the most important raw material for synthetic detergent production; about 50 percent of all synthetic detergents produced in the United States and western Europe are based on it. The alkyl molecular group has in the past usually been $C_{12}H_{24}$ (tetrapropylene) obtained from the petrochemical gas propylene. This molecular group is attached to benzene by a reaction called alkylation, with various catalysts, to form the alkylbenzene. By sulfonation, alkylbenzene sulfonate is produced; marketed in powder and liquid form, it has excellent detergent and cleaning properties and produces high foam.

Alkylbenzene sulfonates and foam

An undesirable effect of the alkylbenzene sulfonates, in contrast to the soap and fatty-alcohol-based synthetic detergents, has been that the large quantity of foam they produce is difficult to get rid of. This foam remains on the surface of waste water as it passes from towns through drains to sewers and sewage systems, then to rivers, and finally to the sea. It has caused difficulties with river navigation; and, because the foam retards biological degradation of organic material in sewage, it caused problems in sewage-water regeneration systems. In countries where sewage water is used for irrigation, the foam was also a problem. Intensive research in the 1960s led to changes in the alkylbenzene sulfonate molecules. The tetrapropylene, which has a branched structure, was replaced by an alkyl group consisting of a straight carbon chain which is more easily broken down by bacteria.

Processes. The organic compounds (fatty alcohols or alkylbenzene) are transformed into anionic surface-active detergents by the process called sulfonation. Sulfation is the chemically exact term when a fatty alcohol is used and sulfonation when alkylbenzene is used. The difference between them is that the detergent produced from a fatty alcohol has a sulfate molecular group ($-OSO_3Na$) attached and the detergent produced from an alkylbenzene has a sulfonate group ($-SO_3Na$) attached directly to the benzene ring. Both products are similarly hydrophilic (attracted to water).

Sulfonation and sulfation

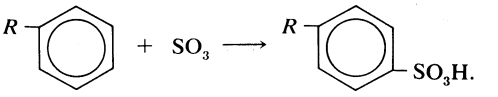
Recent sulfonation methods have revolutionized the industry; gaseous sulfur trioxide is now widely used to attach the sulfonate or sulfate group. The sulfur trioxide may be obtained either by vapourizing sulfuric acid anhydride (liquid stabilized SO_3) or by burning sulfur and thus converting it to sulfur trioxide.

The basic chemical reaction for a fatty alcohol is

$$R\text{CH}_2\text{OH} + \text{SO}_3 \rightarrow R\text{CH}_2\text{OSO}_3\text{H},$$

fatty alcohol + sulfur trioxide = fatty alkyl hydrogen sulfate

and for an alkylbenzene is



alkylbenzene + sulfur trioxide → alkylbenzenesulfonate

R in both reactions represents a hydrocarbon radical.

Following this, caustic soda solution is used to neutralize the acidic products of the reaction. Figure 6 shows the principles of this process.

Research on the part of the petrochemical industry has evolved new anionic synthetic detergents, such as directly sulfonated paraffinic compounds—alpha olefins, for example. Paraffins have been transformed directly into sulfonates by treatment with sulfur dioxide and air using a catalyst of radioactive cobalt.

Conversion of paraffins into sulfonates

Nonionic detergents. The most important nonionic detergents are obtained by condensing compounds having a hydrophobic molecular group, usually a hydroxyl (OH) group, with ethylene oxide or propylene oxide. The most usual compounds are either alkylphenol or a long chain alcohol having a hydroxyl group at the end of the molecule. During the condensation reaction, the ethylene oxide molecules form a chain which links to the hydroxyl group.

Drawing by D. Meighan

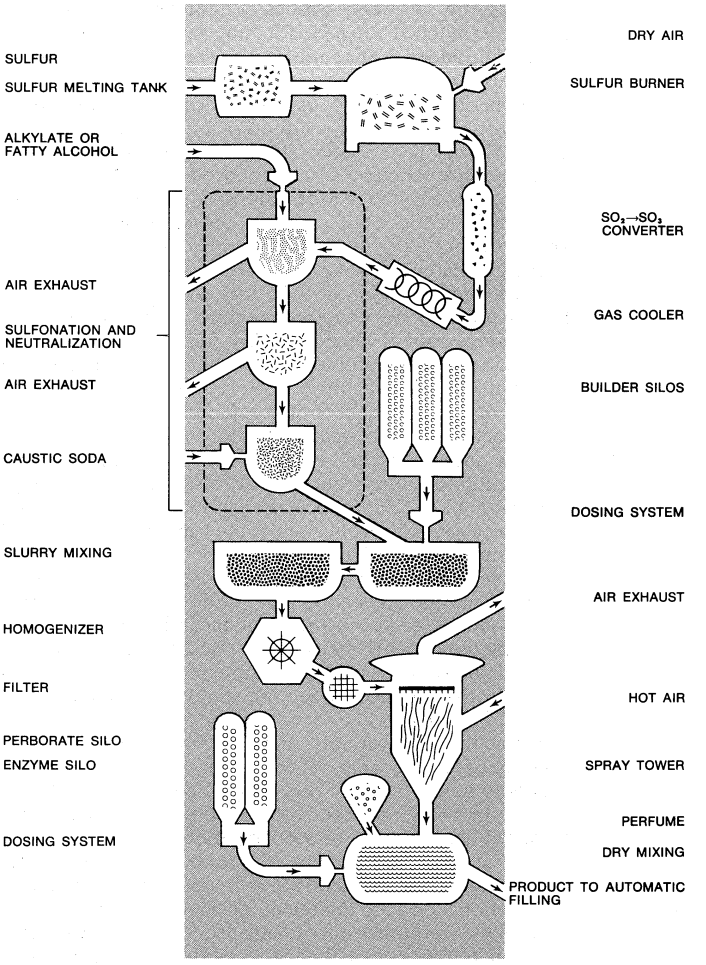


Figure 6: Steps in the manufacture of synthetic detergents.

The length of this chain and the structure of the alkylphenol or alcohol determine the properties of the detergent.

The reaction may take place continuously or in batches. It is strongly exothermic (heat producing), and both ethylene and propylene oxide are toxic and dangerously explosive. They are liquid only when under pressure. Hence, synthesis of these detergents requires specialized, explosion-proof equipment and careful, skilled supervision and control.

Other nonionic detergents are condensed from fatty acids and organic amines (alkylol amines). They are important as foam stabilizers in liquid detergent preparations and shampoos.

Some nonionic synthetic detergents may cause problems with unwanted foam in waste-water systems; the problem is not as serious as with anionic synthetic detergents, however.

Cationic detergents. Cationic detergents contain a long-chain cation that is responsible for their surface-active properties. Marketed in powder form, as paste, or in aqueous solution, they possess important wetting, foaming, and emulsifying properties but are not good detergents. Most applications are in areas in which anionic detergents cannot be used. Cationic-active agents are used as emulsifying agents for asphalt in the surfacing of roads; these emulsions are expected to "break" soon after being applied and to deposit an adhering coat of asphalt on the surface of the stone aggregate. These agents absorb strongly on minerals, particularly on silicates, and therefore make a strong bond between the asphalt and the aggregate. Cationic detergents also possess excellent germicidal properties and are utilized in surgery in dilute form.

Uses of cationic-active agents

Ampholytic detergents. Ampholytic detergents are used for special purposes in shampoos, cosmetics, and in the electroplating industry. They are not consumed in large quantities at present.

Finishing synthetic detergents. The largest quantities of synthetic detergents are consumed in the household in the form of spray-dried powders. They are produced from an aqueous slurry, which is prepared continuously or in batches and which contains all the builder components. Builders, consisting of certain alkaline materials, are almost universally present in laundry soaps. These materials give increased detergent action. The most important are sodium silicate (water glass), sodium carbonate (soda ash), and various phosphates; the latter have contributed to the problem of waste-water pollution by contributing nutrients which sustain undesirable algae and bacteria growth, and much work is being done to find acceptable builders which may replace, at least partially, phosphates. The slurry is atomized in heat to remove practically all the water. The powder thus obtained consists of hollow particles, called beads, that dissolve quickly in water and are practically dust free. Another portion of the syndets is transformed into liquid detergent products and used primarily for hand dishwashing. Though syndet pastes are seldom produced, solid products, manufactured in the same way as toilet or laundry soap, have been sold in increasingly greater quantity. Sodium perborate is sometimes added to the spray-dried beads to increase cleaning power by oxidation. Enzymes may be added as well. Many modern washing powders combine synthetic detergents, anionic and nonionic, with soap, to give maximum efficiency and controlled foam for use in household washing machines.

(A.S.D.)

Dyes

Most modern dyes are coloured substances synthesized from certain chemical compounds called benzenoid hydrocarbons, obtained from either coal tar or petroleum. Their most important use is in dyeing textiles. Wool or cotton fabrics may have dyes applied at any stage of manufacture from fibre to finished garment, depending on the nature of the textile, its subsequent processing, and the requirements in use. Dyeing, as the term is generally understood, occurs only when the dye is in solution, usually in an aqueous medium. Some dyes whose solubility is small are called disperse dyes, because they are dispersed rather than dissolved in the water. Some colorants, called pigments, are completely insoluble; they are melted together with synthetic resins to impart colour before the synthetic fibres are extruded. Synthetic textiles such as nylon are coloured in this way. (The term colorant is used to describe both dyes and pigments.) At the present time, dyeing from nonaqueous solvents, called simply solvent dyeing, shows considerable promise as a low-cost process that conserves water and reduces pollution.

Very few dyes of natural (animal or vegetable) origin are presently used by the commercial dyer; but before the mid-19th century they were the only colorants available. Their use over the centuries, especially from the Middle Ages on, laid the foundations of the art, and later the science, of the modern dyeing industry, which is of considerable complexity. More than 7,000 synthetic organic colorants are presently in commercial use, differing in fastness and other properties, and requiring different methods of application. Most of the dyestuffs are used by the textile industry, but the leather, paper, food, and cosmetic industries are also important users. Synthetic organic pigments are used in the manufacture of printing inks, paints, and automobile finishes, and for incorporation in plastics and rubber. Since each industry requires different properties in the dyes and pigments it employs, the large dye-making concerns usually provide technical service and advice to their customer industries on the selection and use of products. Despite the fact that colorants form only a small proportion of the cost of a finished article (less than 1.5 percent of the cost of a woman's dress and less than one percent of that of a woollen overcoat), the dye-making industry remains of considerable economic importance.

HISTORY

Natural dyes. The dyes used up to the middle of the 19th century, the greater part of them of vegetable origin, involved many problems, both in respect to sources and to processes. Tyrian purple mysteriously disappeared about the middle of the 15th century. The use of natural indigo involved protracted fermentation processes to liberate the dye in a soluble, colourless form. After cloth had been steeped in a vat, oxidation in air gave blue dyeings on cotton and wool. Many natural dyes have no affinity per se for textile fibres until such fibres have been treated with aluminum, iron, or tin compounds to receive the dye (mordanting). Chief among the mordant dyes was alizarin derived from madder root, widely cultivated at one time in Europe, Turkey, and India. It yielded a fast red on aluminum mordant, a purplish black on iron mordant; intermediate shades were obtained by using a mixture of mordants. The discovery of the New World and the establishment of trade routes to the Americas led to the introduction into Europe of new mordant dyes such as brazilwood, giving reds on aluminum, browns on iron, and a rose-pink on tin mordants, and logwood, or haematoxylin, giving fine blacks on a chromium oxide mordant. Scarlets were obtained either from kermes, which consists of the dried bodies of certain female insects, or cochineal, produced in Mexico from the dried female insect *Coccus cacti,* which lives on the Indian fig tree or nopal. Until 1954 the traditional scarlet of British Guards' uniforms was obtained by dyeing tin-mordanted wool with cochineal.

Synthetic dyes. The year 1856, in which William Henry Perkin, an English chemist, discovered the synthetic dye mauveine, marks the beginning of synthetic dyestuffs, which ultimately supplanted natural dyes. The synthetic-dye manufacturing industry was founded by Perkin in 1857, when he set up facilities near London for the commercial production of mauveine and, later, of other synthetic dyes. Other dye-making factories followed both in the U.K. and continental Europe, and new dyes began to appear on the market. Research was directed toward determining the structures of important natural dyes and then to synthesizing them. The first success was the synthesis of artificial alizarin from coal-tar anthracene in 1868. Indigo was first synthesized in 1880 though an economic manufacturing process was not devised until almost 20 years later. The cultivation of madder and indigo rapidly declined as the synthetic products came to be accepted by dyers. Organic chemistry flourished in the research centres

The discovery of mauveine

of Europe, and new dyes, many of them owing nothing to nature, were perfected by dye makers. These included azo dyestuffs, certain organic compounds having the general formula RN = NR′ (1862), which today comprise the largest group of synthetic dyes; congo red, the first direct cotton dye requiring no mordanting or fixing process (1884); azoic dyes, a group of organic dyes derived from azo benzene, mostly reds and yellows (1911); acetate dyes, mainly azo or anthraquinone dyes, highly dispersed to render them capable of penetrating acetate fibres (1922–23); and fibre-reactive dyes, those which combine chemically to form a link with the material (1956). Dyestuffs research continues, as indicated by the large annual volume of patent literature.

RAW MATERIALS FOR MANUFACTURE OF DYES

The primary raw materials from which the dye maker synthesizes dyestuffs and pigments are the coal-tar chemicals called aromatic hydrocarbons, chiefly benzene, toluene, orthoxylene, metaxylene, paraxylene, naphthalene, and anthracene. Dye manufacture thus involves the process of first deriving these intermediate chemicals from raw materials and then building up dyes from the intermediates.

Coal tar, the traditional source of dye intermediates, is obtained from bituminous coal by distillation. Coal tar is a complex mixture of substances, which, on distillation, is commonly separated into light oil (boiling point up to 170° C), middle oil (boiling point 170–230° C), heavy oil (boiling point 230–270° C), anthracene oil (boiling point above 270° C), and pitch (nonvolatile still residue). The light oil is redistilled and separated into fractions of narrower boiling-point range: 80–110° C, 110–140° C, and 140–170° C. These fractions are purified by successive contact with caustic soda (sodium hydroxide) solution, and concentrated sulfuric acid. The treated fractions are again distilled to yield respectively 90 percent benzol, 50 percent benzol, and solvent naphtha. By fractional distillation the 90 percent benzol may be separated into its constituents, benzene, toluene, and the xylenes. The commercial benzene thus obtained is not pure enough for successful synthesis into dye intermediates and must be purified even further. This is accomplished by a freezing process, followed by treatment with sulfuric acid. Toluene and the xylenes are brought to a similar degree of chemical purity by a combination of chemical and physical means.

Naphthalene is obtained from middle oil (170–230° C), from which it separates, on cooling, as crystals, which are pressed, treated with caustic-soda solution, then with sulfuric acid, and finally distilled or sublimed (vaporized and recondensed to solid form).

Anthracene is separated from anthracene oil in the form of crude crystals. These are then dissolved in naphtha, distilled in superheated steam, and recrystallized in pure form.

New sources of hydrocarbons

Although coal tar remains an important source of aromatic hydrocarbons, the petrochemicals industry is an increasingly important supplier; in the United States it is the principal source. Distillation of crude oil yields substances that can be reformed either by thermal or catalytic processes and suitable purification into the aromatic hydrocarbons. Largest users are the synthetic fibre, synthetic rubber, detergent, and plastics industries, but sizable quantities are consumed by the dye-making industry.

MANUFACTURING PROCESSES FOR INTERMEDIATES

All synthetic organic dyes and pigments contain a ring structure of atoms. The same formation is present in dye intermediates. It is usual also for the dye intermediate to have one or more of the groups of atoms that react chemically to form salts.

A simple example of an intermediate is β-naphthol, which is obtained from naphthalene by heating it with 96 percent sulfuric acid at 160° C (320° F), adding sodium sulfate, and volatizing the remaining naphthalene with steam. Sodium naphthalene-2-sulfonate, left behind after this steam treatment, is then fused with caustic soda to yield β-naphthol. The β-naphthol is separated from the mixture by diluting it with water and treating it with sulfuric acid. These reactions, known respectively as sulfonation and alkali fusion, are given the name unit processes.

Unit processes

A number of such unit processes are available to the chemical manufacturers, and by means of various permutations, it is possible to synthesize the several hundred intermediates needed for the manufacture of the modern range of colorants.

Nitration. In nitration, a nitro group is substituted for a hydrogen atom in an aromatic hydrocarbon molecule by the action of nitric acid. Benzene yields nitrobenzene by reaction with a mixture of nitric and sulfuric acids at a temperature not exceeding 50° C (122° F). Toluene, with mixed acid at 30–35° C (86–95° F), yields orthonitrotoluene, metanitrotoluene, and paranitrotoluene. Technical quality nitrotoluenes are obtained from the crude nitration product by physical separation methods including fractional distillation; 1-nitronaphthalene is obtained when naphthalene is nitrated.

Nitration is a strongly exothermic, or heat-releasing, reaction; in large-scale nitrations the temperature must be maintained between established limits and efficient means of cooling provided. Cast-iron, mild steel, or stainless-steel vessels are used, arranged to permit agitating the mixture. The addition of nitric acid is carefully controlled by instruments that detect incipient rises in temperature. Other safeguards stop the addition in case the stirring mechanism is halted by a power or mechanical failure.

Reduction. Amines, chemical compounds formed by substituting organic radicals for the hydrogen atoms of ammonia, may be obtained from nitro compounds by reduction; that is, replacement of oxygen by hydrogen. This is accomplished by mixing the nitro compound with iron borings and a minimum quantity of aqueous hydrochloric acid in a cast-iron reducer with a powerful agitator. The end products are aniline, or other amines, and ferric oxide. The aniline, which separates as an upper layer above the aqueous suspension of iron and oxide, is removed mechanically and purified by steam distillation followed by fractionation. Analogous processes are employed in making other amines from their corresponding nitro compounds.

Halogenation. Halogenation is the introduction, by direct or indirect means, of a halogen (fluorine, chlorine, or bromine) into molecules of dye intermediates. In many cases direct substitution can be effected using elemental chlorine or bromine with or without a catalyst. In other cases a diazotized amine is treated with cuprous chloride or bromide. Indirect means must be employed for introducing fluorine. Hydrogen fluoride (HF), for example, is used to displace chlorine in chloro compounds.

Amination. The amination, or conversion of chloro compounds to amines, can be accomplished by ammonolysis, a process that involves heating the chloro compound with aqueous ammonia in a steel autoclave (equipment for working at pressures above atmospheric). Chloronitrobenzene, for example, is transformed into nitroaniline by this process. Several industrial amination processes exist. In the Dow Process monochlorobenzene is converted into aniline by heating it with aqueous ammonia at 240° C (464° F) in the presence of a copper oxide catalyst. In the Halcon Process phenol undergoes ammonolysis to yield aniline.

Hydroxylation. Hydroxylation is the introduction of a hydroxyl, or OH, group into the dye-intermediate molecule. Among the methods used for this are fusion by sodium hydroxide, Bucherer reaction (conversion of aromatic amines into phenolic compounds by aqueous sulfite or bisulfite), hydrolysis of chloro compounds under various reaction conditions according to the reactivity of the chloro compound, decomposition of diazonium salts by hot aqueous sulfuric acid, hydrolysis of sulfo groups, and oxidation (described below); for example, the compound cumene yields phenol and acetone.

Phenol and acetone

Oxidation. Oxidation, the combination of a substance with oxygen, or, generally, any reaction in which an atom loses electrons, is a ubiquitous chemical reaction in dye chemistry; only a few examples can be given here. The conversion of methyl groups into carboxylic acids, and more especially into styryl compounds, is effected by various oxidants such as sodium hypochlorite, potassium

permanganate, and sodium dichromate. The catalyzed oxidation of naphthalene with air leads to phthalic anhydride. Other important reactions include oxidation of leuco compounds to dyes and the formation of complex polycyclic substances from simpler molecules.

Benzidine rearrangement. The benzidine rearrangement consists of the conversion of nitrobenzene and its derivatives into derivatives of biphenyl in a two-stage process. The first stage consists of alkaline reduction with zinc dust to the hydrazobenzene; and the second of treatment with hydrochloric acid.

The above and other unit processes are used in various combinations in order to produce the important intermediate chemicals that are, in turn, used to manufacture the dyes themselves.

H Acid. H Acid, which is 1-amino-8-hydroxynaphthalene-3,6-disulfonic acid, an important dye intermediate, is produced from naphthalene by a combination of the unit processes of sulfonation, nitration, reduction, and hydrolysis. H Acid is used in the manufacture of a large number of azo dyes.

Diaminostilbene disulfonic acid. 4,4′-diaminostilbene-2,2′-disulfonic acid is produced from *p*-nitrotoluene by a combination of the unit processes of sulfonation, oxidation, and reduction. This product is used in large quantities for manufacture of fluorescent brightening agents.

MANUFACTURING PROCESSES FOR DYES AND PIGMENTS

In the early synthetic dye plants, the equipment was primitive, consisting of simple iron vessels and wooden vats. In the latter half of the 19th century new processes arising from the discovery of an increasing number of new dyes led to improvements in the design of reaction vessels, autoclaves, pumps for the transfer of liquids, compressors, and vacuum pumps. The development of vitreous coating techniques, the use of lead and copper, and improvements in the centrifuge, drying stove, grinding, and mixing equipment also proceeded at that time and has continued to the present. The advent of stainless steel and the introduction of plastic and polymer materials and of the newer techniques, especially the rubber coating of mild steel, has resulted in the gradual disappearance of the wooden vat and has reduced the use of lead and copper as constructional materials. In modern plants multipurpose vessels constructed for easy linking to one another and to filters, separators, etc. are employed wherever possible. Rapid changes may be made from one combination to another for efficient production. Stainless steel and rubber-covered steel may be quickly and easily cleaned in contrast to the former laborious process of cleaning wooden vats. In the larger modern plants, flow rates, heating, cooling, and other services are usually automatically controlled. Inlet and outlet valves are actuated, usually pneumatically, by sensing probes that detect changes in temperature, pressure, and hydrogen ion concentration that have occurred in the reaction vessel or other part of the plant. Intermediates, dyes, and pigments are usually manufactured in batches, discontinuously, because the total annual requirements of each may not amount to more than 10–100 tons (9–90 metric tons). If an annual output greatly in excess of these quantities is required, semicontinuous or continuous processes may be adopted, relatively small amounts of reactants coming together at any one time. This increases the safety factor in strongly exothermic reactions.

Discontinuous batch operation

Many of the unit processes used to produce intermediates are also used to make dyes and pigments. In the processes described below, only the penultimate and final stages leading to the particular colorant are given. For precision, the Part I Reference, and the Part II number, where known, are given from the *Colour Index* (C.I.), the standard reference work published jointly by the Society of Dyers and Colourists in Great Britain, and the American Association of Textile Chemists and Colorists in the United States, with the cooperation of manufacturers in Belgium, Brazil, Chile, France, Germany, Greece, India, Italy, Japan, The Netherlands, Poland, Spain, and Switzerland. Part I of the *Colour Index* classifies dyes and pigments according to usage while Part II classifies them according to chemical constitution.

Nitroso dyes. By treating β-naphthol with nitrous acid, α-nitroso-β-naphthol is obtained, which reacts with an iron salt such as ferrous sulfate to give Pigment Green B, C.I. Pigment Green 8, 10006.

Nitro dyes. Celliton Fast Yellow RR, C.I. Disperse Yellow 1, 10345, results from the condensation of 2,4-dinitrochlorobenzene with 4-aminophenol. The process may be carried out in alcoholic solution. The final product, like all disperse dyes, must be ground to a particle size of 1–10 microns (1,000 microns = 1 millimetre) in an aqueous medium containing a dispersing agent.

Azo dyes. Almost all azo dyes are made by diazotization of a primary aromatic amine in an acid solution by using nitrous acid in the presence of ice. The diazo compound so formed is coupled with a suitable component, such as an aromatic amine, naphthol, or other phenolic substance, to form an azo compound.

These reactions are usually carried out in aqueous solution in rubber-lined steel equipment of capacities varying from 2,000 to 5,000 gallons (7,600–19,000 litres). Isolation of the final dye is usually achieved by addition of common salt in sufficient quantity to precipitate the dye, while leaving small amounts of undesirable, often dulling, by-products in solution. The dye precipitate is then separated from the solution in a filter press. As "press cake," it undergoes the finishing processes usual for most dyes, such as wet milling, drying, grinding, and reduction to a standard strength by mixing with diluents. Monoazo colorants contain a single azo group. Examples are the wool dye Acid Orange II, C.I. Acid Orange 7, 15510; and Permanent Red FRR, C.I. Pigment Red 2, 12310. Examples of diazo colorants, the molecules of which contain two azo groups, include Congo Red, C.I. Direct Red 28, 22120; and the solvent-soluble Fat Red B, C.I. Solvent Red 24, 26105. Triazo dyes contain three, and tetrakisazo dyes contain four, azo groups.

Acid alizarine dyes. Valuable greens and blues are derived by the process of sulfonation of diaminoanthraquinones to give dyes with the requisite degree of water solubility. Processes of sulfonation, nitration, reduction, alkylation, and other steps are also used to obtain certain disperse dyes from anthraquinone.

The more complex vat dyes are built up from simpler compounds by many of the common unit processes such as sulfonation, as well as more specialized chemical reactions.

Indigoid and thioindigoid dyes. The sodamide process (1901) for Indigo, C.I. Vat Blue I, 73000, begins with aniline, which, by reaction with formaldehyde and sodium cyanide, yields phenylglycine; the latter on fusion with a mixture of sodamide and caustic potash is changed into indoxyl, oxidation of which leads to indigo. The thioindigoid, Durindone Orange R, C.I. Vat Orange 5, 73335, is also synthesized by a complicated route.

Sodamide Indigo process

Triphenylmethane dyes. Crystal Violet, C.I. Basic Violet 3, 42555, has a typical structure containing a central carbon atom, which is introduced by various methods including the employment of phosgene, $COCL_2$, formaldehyde, HCHO, and intermediate chemicals containing the groups $=C=O$ and $=CH-OH$.

Cyanine dyes. Dyes in this class commonly contain two ring systems of atoms linked by conjugate chains ($N-C=C-C=N$, for example). There are many different methods of preparation, among which are the condensation of complex aldehydes with amines.

Fibre-reactive dyes. Fibre-reactive dyes react chemically to combine with and become part of the fibre substance. The simpler types are monochlorotriazinyl and dichlorotriazinyl compounds. An amino compound with an azo, phthalocyanine, or anthraquinone group in its molecule is condensed in an aqueous medium with cyanuric chloride.

Phthalocyanine pigments. Copper phthalocyanine, a blue pigment, is manufactured in hundreds of tons annually by heating phthalonitrile, or some other related substance derived from phthalic anhydride, with a copper (cuprous) salt. For a high-grade pigment quality, the raw pigment must be subjected to special conditioning processes to achieve the required physical form. Valuable green pigments are obtained from copper phthalocyanine by halogenation processes. (E.N.A.)

Pharmaceutical industry

The World Health Organization (WHO) defines a drug or pharmaceutical preparation as

> any substance or mixture of substances manufactured, sold, offered for sale, or represented for use in . . . the diagnosis, treatment, mitigation, or prevention of disease, abnormal physical state or the symptoms thereof in man or animal; [and for use in] . . . restoring, correcting, or modifying organic functions in man or animal.

The same organization defines a pharmaceutical specialty as "a simple or compound drug ready for use, and placed on the market under a special name or in a characteristic form."

The modern pharmaceutical industry began in the 19th century with the discovery of highly active medicinal compounds that could most efficiently be manufactured on a large scale. As these compounds replaced herbal medicines of earlier times, the occurrence and severity of such diseases as pernicious anemia, rheumatic fever, typhoid fever, lobar pneumonia, poliomyelitis, syphilis, and tuberculosis were greatly reduced. Pharmaceutical industry research has greatly aided medical progress; of the 66 most valuable drugs introduced since aspirin in 1899, 57 were discovered and then produced in industrial laboratories.

History. The earliest records of medicinal plants and minerals are those of the ancient Chinese, Hindu, and Mediterranean civilizations. In 2735 BC, the Chinese emperor Shen Nung wrote an herbal in which he described the antifever capabilities of a substance known as Ch'ang Shang, since shown to contain antimalarial alkaloids. The school of alchemy that flourished in Alexandria in the 2nd century BC could prepare white lead (lead carbonate) from litharge, arsenic from realgar (arsenic disulfide), and mercury by roasting cinnabar (mercuric sulfide) in a current of air. It is recorded in *De materia medica,* a book of the 1st century BC, that verdigris (basic cupric acetate) and cupric sulfate were prescribed as medicinal agents. Remarkably, cupric sulfate is still used in medicine today.

Many crude drugs still used now, such as ipecacuanha root (ipecac), were known and employed by the ancients. The Egyptians treated constipation with senna pods and castor oil and employed caraway and peppermint to relieve indigestion. The Greek physician Galen (*c.* AD 130–*c.* 200) included hyoscyamus, opium, squill (a plant drug used as an expectorant, cardiac stimulant, and diuretic), and viper toxin among other drugs in his apothecary shop. He also insisted on purity in drugs; *i.e.,* the right variety and age of the botanical specimens.

Pharmaceutical practice improved markedly in the 16th and 17th centuries. In 1546 the first pharmacopoeia, or collected list of drugs and medicinal chemicals with directions for making preparations from them, appeared in Nuremberg (Nürnberg, West Germany). Previous to this time, medical preparations had varied in concentration and even in constituents. Other pharmacopoeias followed, in Basel (1561), Augsburg (1564), and London (1618). Despite its name, the *London Pharmacopoeia* was mandatory for the whole of England and thus was the first example of a national pharmacopoeia.

The first apothecaries

In 1617 the Society of Apothecaries, London, was founded, marking the emergence of pharmacy as a distinct profession. The separation of apothecaries from grocers was authorized by King James I; only a member of the society could keep an apothecary's shop and make or sell pharmaceutical preparations. In 1841 the Pharmaceutical Society of Great Britain, London, was founded, placing the education and training of the pharmacist upon a proper scientific basis. Today, pharmaceutical practice and education are carefully supervised throughout the world.

Known as "Hoffman's drops," ether was first employed as an anesthetic in 1842; chloroform followed soon afterward in 1847. Alkaloid compounds were also isolated from plant sources during this period. Narcotine was obtained from opium by a French pharmacist in 1803 and was followed by morphine in 1806, emetine and strychnine (around 1817), brucine and piperine (1819), colchicine and quinine (1820), nicotine (1828), atropine (1833), cocaine (1860), and physostigmine (1867). Isolation of these potent compounds was a milestone in pharmaceutical progress for three reasons. First, accurate doses could now be administered; this had been impossible previously with crude drugs of unknown and variable composition. Second, toxic effects due to impurities in crude drugs could now be eliminated if pure compounds were used; and third, knowledge of the chemical structure of these drugs led to attempts at laboratory synthesis, which led in turn to discovery of valuable related compounds.

Joseph Lister, in England, opened the modern era of antiseptic surgery in 1865 when he used phenol (carbolic acid) to prevent infections. In 1869 the soporific properties of chloral hydrate, the first synthetic hypnotic (sleep-producing drug), were discovered. In 1874 it was found that organic nitrites relax the blood vessels, and, in 1875, salts of salicylic acid were introduced as remedies for fever. The year 1879 witnessed the introduction of saccharin, still in use today as a sweetening agent for diabetic patients. The simple compound acetanilide, introduced in 1886, was one of the first analgesic-antipyretic drugs (*i.e.,* reducing both pain and fever) to be used but was later replaced by the less toxic phenacetin in 1887, by aspirin in 1899, and all of these to some extent by paracetamol in 1956. The hypnotic sulfonal (sulfonmethane) was discovered about 1888, followed a few years later by barbital; this latter led to a whole series of barbiturates, of which phenobarbital is the best known.

Cocaine was the only known potent local anesthetic until about 1900, when the much simpler compound benzocaine was introduced. The closely related local anesthetic procaine followed early in the 20th century.

In 1909 arsphenamine, highly effective against syphilis, was introduced. Since arsphenamine is both insoluble in water and unstable, further work led to neoarsphenamine in 1912, a soluble derivative. Many other synthetic drugs followed, among which was the hormone progesterone, synthesized in 1934. Some thousands of similar compounds have since been prepared, some of which are used as oral contraceptives.

The sulfa drugs and antibiotics

In 1935 it was discovered that sulfanilamide (Prontosil) stopped the growth of bacteria. Over 6,000 derivatives of sulfanilamide—the sulfonamides, or sulfa drugs—were prepared and tested for their antibacterial properties. Today, the sulfonamides have partially been superseded by antibiotics, of which the first was penicillin, first isolated in 1941. In 1959, 6-aminopenicillanic acid was isolated for the first time; this led to production of many semisynthetic penicillins such as ampicillin, carbenicillin, cloxacillin, methicillin, oxacillin, and phenethicillin. Also used as antibiotics are compounds known as cephalosporins, the first of which was isolated in 1961.

Classification of pharmaceutical products. Drugs may be classified in one of three ways: by chemical group (*e.g.,* alkaloids, mentioned above); pharmacologically (*i.e.,* by the way they work in the body); and according to their therapeutic uses. Pharmacological and therapeutic classifications show considerable divergence, as drugs that act upon the body in different ways may bring about the same desired therapeutic result. Furthermore, classification by therapeutic usage is complicated by the fact that a drug may be used to combat more than one ailment. The antimalarial compound primaquine, for example, may also be employed to relieve arthritis. Some familiar drugs classified by therapeutic use include aspirin, an analgesic, or painkiller; benzocaine, a local anesthetic; magnesium carbonate, an antacid; charcoal, an antiflatulent (preventing production of flatulent gases); penicillin, used against syphilis; calcium lactate, a calcium supplement; hexachlorophene, a deodorant; phenolphthalein, a laxative; levulose, a nutrient; cascara sagrada, a purgative; phenobarbital, a sedative; and thiamine hydrochloride, a vitamin.

PRODUCTION OF PHARMACEUTICALS

Raw materials and their purity. Pharmaceutical raw materials may be plant, animal, or other biologic products; inorganic elements and compounds; or organic compounds. If the raw material is "official"—that is, if it is the subject of a monograph in a pharmacopoeia or national formulary—then the minimum acceptable degree

of chemical purity is specified. Very often, however, because some raw materials at specified levels may begin to decompose after a time, purification far exceeds these minimum requirements. If extra purification is inconvenient, a preservative may be added. If the raw material is not the subject of an official monograph, then physical or chemical specifications or both are drawn up by the manufacturer in accordance with the pharmaceutical requirements of the finished product, on lines similar to the official monographs.

Crude drugs. The term crude drug is usually applied to plant or animal organs or whole organisms or exudations of these, either in the fresh state or dried and either ground or unground. Some crude drugs such as acacia, belladonna, and starch are official; for these, rigid specifications are available. Plant-derived crude drugs may come from cultivated sources or may be collected in the wild. Cultivation and collection are carefully planned to obtain drugs with the highest possible therapeutic activity or content of active constituents. The harvested drug plant must then be cleaned to remove matter such as sand and dirt. Next, unwanted plants or plant parts are carefully removed. Some crude drugs, such as belladonna, undergo curing, a process that consists either of slow drying or sweating, during which enzymes bring about chemical changes whereby the content of active ingredient is increased. Alternatively, as with cascara sagrada, the drug is carefully stored for a year, during which time unwanted constituents slowly decompose. The final stage in crude-drug production consists of drying, accomplished in air or with artificial heat. Drying aids preservation, stops various chemical reactions that might weaken or destroy the substance, facilitates subsequent grinding, and reduces weight and bulk. Drugs containing volatile constituents are dried at temperatures near freezing. After drying, some drugs (such as belladonna and ipecac) are reduced to a fine powder.

Drying the drug

Inorganic elements and compounds. Production methods for inorganic elements and compounds are not always the same as those for industrial chemicals, since purification is usually carried to a much higher level. Industrial zinc oxide, for example, may contain as much as 10 percent of impurities, including small quantities of lead. The pharmaceutical substance, by contrast, must contain at least 99 percent zinc oxide and no lead. For this reason, industrial zinc oxide is manufactured by combustion of zinc in air, whereas pharmacopoeial zinc oxide is produced from zinc sulfate.

Organic compounds. Organic compounds used as pharmaceuticals are either extracted from natural sources or prepared by chemical synthesis (various methods are described below). Physical specifications used in purity control of inorganic and organic pharmaceuticals include particle size, colour, odour, solubility, homogeneity, and freedom from particulate impurities. Some common chemical specifications are melting point, boiling point, density, viscosity, and freedom from impurities.

Preparation from natural sources. *The alkaloids.* As mentioned above, the first pure pharmaceuticals isolated from natural sources were the alkaloids. Though the methods adopted for their extraction from plants vary in detail, all are based on three general characteristics of these compounds. First, most alkaloids are only slightly soluble in water but readily soluble in certain organic solvents such as benzene, chloroform, ether, and light petroleum. Second, alkaloids combine with acids to form salts that are usually freely soluble in water but only slightly so in organic solvents. Third, alkaloids are liberated from their salts by alkalies.

Extraction of alkaloids

Application of these general principles can be seen in the following generalized outline of extraction methods. The crude drug, ground to a suitable state of subdivision, is mixed with water, a water-soluble alkali such as lime, and some organic solvent that does not mix with water. The mixture separates into two layers: one contains water, lime, and impurities and is discarded; the other contains the alkaloid dissolved in the organic solvent. Fresh water and dilute acid are now added to the mixture. Again there are two layers, but the acid has caused the alkaloids to pass from the organic layer into the aqueous layer. The aqueous layer is now separated, and the alkaloidal salt or salts may be crystallized out by cooling or concentrating the solution. The process described above is used to obtain quinine sulfate from cinchona bark.

Glycosides. Glycosides such as digoxin are another important group of drugs obtained from plants. Generally, the glycoside is extracted from the crude drug with alcohol. Adding basic lead acetate to the alcoholic extract causes the impurities to fall to the bottom. After being poured through a filter, the alcoholic extract is concentrated, and the glycosides may crystallize out. Often, however, concentration of the extract is not sufficient to cause crystallization, and more complicated procedures must be employed.

Volatile, or essential, oils. Volatile, or essential, oils, also obtained from plants, may be extracted by distillation, steam distillation, expression, or by extraction with fats or organic solvents. In steam distillation, the most common method, the crude drug—either fresh or dried—is used in powder form. Water is mixed with the powder, serving the double purpose of preventing decomposition of plant material by excessive heat and of facilitating volatilization of the essential oil. The powder–water mixture is usually placed in a basket through which the steam penetrates. The distillate consists of a water–oil mixture; the oil forms a separate layer, which is run off. Anise, cinnamon, clove, coriander, fennel, and peppermint oils are all obtained by this method, as is the pharmaceutical substance camphor from camphor wood.

Fixed or fatty oils cannot be obtained by distillation, but only by expression or extraction. Castor, olive, and sesame oil, for example, are all obtained by expression. Cod-liver and halibut-liver oils, rich sources of vitamins A and D, are both extracted by passing steam into tanks containing the livers suspended in water, until the temperature reaches 70–80° C (158–176° F). The tissues disintegrate, and the oils float to the surface and are skimmed or centrifuged off.

Obtaining cod-liver oil

Insulin. Other drugs of animal origin primarily include glandular extracts containing hormones, such as thyroid and parathyroid, and purified hormones, such as insulin, obtained from the pancreas of a pig, sheep, or ox. For insulin production, the pancreas must either be absolutely fresh or frozen immediately after removal from the animal. Otherwise, enzymes will cause decomposition of the insulin. The glands are minced and immediately stirred with cold acidified alcohol, which both extracts the insulin and prevents its destruction by the enzyme trypsin. The gland residue is separated by centrifugation and further extracted with more acidified alcohol. The final alcoholic extract is passed through filter presses to remove the remaining fine particles of meat. The extract is now made sufficiently basic to cause precipitation of the inert proteins and allow their removal. Next, the extract is acidified again, and the alcohol is removed as quickly and with as little heat as possible. When the concentrated extract is cooled, fat separates and is filtered off. The crude insulin is then precipitated by adding common salt and is purified by repeated crystallization. For medical practice, insulin is almost always made up in the form of a solution for injection.

Antibiotics. Antibiotics are an extremely important group of drugs isolated from natural sources. Penicillin was originally produced by growing the *Penicillium notatum* mold in small containers; the mycelium, or vegetative portion, formed a mat on the surface of the medium that contained the penicillin in solution. The process was difficult to operate and required a large labour force to inoculate and harvest the containers.

Penicillin production was revolutionized by the discovery that a strain of *Penicillium chrysogenum* mold (isolated from an overripe cantaloupe) would produce high yields when grown in deep culture. The growth of a mold is aerobic (it requires air), and in surface culture the growth and consequent production of antibiotic is limited by the rate at which air can diffuse into the medium. In submerged or deep culture, the mold is grown in large tanks supplied with a continuous flow of sterile air.

Deep culture of penicillin

Though the medium used in fermentation varies for the particular antibiotic being produced, all contain a source of carbon (which may be lactose or glucose); nitrogen (in the form of ammonium salts); and trace elements (such as phosphorus, sulfur, magnesium, iron, zinc, and copper). If phenylacetic acid is added, the mold will utilize it and produce benzylpenicillin (penicillin G), whereas, if phenoxyacetic acid is added, phenoxymethylpenicillin (penicillin V) is obtained. In addition, corn-steep liquor may be added for penicillin production, and soybean meal and dried distiller's residues for streptomycin; these help increase the yield of product. (Corn-steep liquor is prepared by steeping cleaned corn grain in water for about 40 hours at about 48° C [118° F]; the liquor is drawn off and evaporated at reduced pressure to a suitable concentration—about 55 percent solids. Corn-steep liquor stimulates penicillin formation due to certain amino acids, minerals, and precursors that it contains.) Before use the medium—as well as the fermenter and associated equipment—is steam sterilized, as bacterial contamination can destroy the antibiotic.

A large volume of concentrated, actively growing fungal suspension is required for the main fermenting tanks, to keep the fermentation time to a minimum. This is obtained in three stages. First, the selected culture is transferred from cold storage to a culture medium (agar) to produce an initial inoculum. This inoculum is then cultured in shake flasks to give a suspension. Finally, the suspension is grown in seed tanks in the plant for 24–28 hours to the desired volume and concentration before transfer to the main fermenters. Fermentation is continued for three to five days, during which the vessel is cooled—to keep the temperature between 23–27° C (73–81° F)—and stirred and aerated with sterilized air. The introduction of large volumes of air causes frothing, which is controlled by the addition of antifoams such as lard oil, octadecanol, or silicones. When fermentation is complete, the mycelium is removed on a rotary filter and the penicillin extracted into an organic solvent (such as butyl acetate or methyl isobutyl ketone), after acidification. The free acids of penicillins are generally unstable and are converted into a metal salt by extraction with alkali, followed by freeze-drying of the extract, or by addition of a concentrated solution of a metal salt such as potassium acetate, whereby the potassium salt of the penicillin is precipitated. All products that are to be administered by injection are sterilized by passage through a sterilizing filter, followed by freeze-drying, precipitation, or crystallization under sterile conditions.

Vaccines. Vaccines, viral or bacterial preparations used to induce active immunity, may be divided into three groups: toxoids, bacterial vaccines, and virus vaccines. Toxoids, such as diphtheria and tetanus toxoid, can only be prepared from toxin-producing bacteria. To produce diphtheria toxoid, for example, diphtheria bacilli are cultured under artificial conditions in a suitable broth to produce a fluid rich in toxin. The bacilli are then filtered off, and formaldehyde is added to the crude toxin filtrate, which is incubated for two to three weeks, with the result that the substance is no longer toxic and is called a toxoid. The toxoid is then purified by precipitation with ammonium sulfate and is filtered off. If alum or some other aluminum compound is used to precipitate the toxoid, the product is relatively insoluble and thus more slowly absorbed, thereby possessing increased antigenic (causing the production of antibodies) properties.

Preparation of bacterial vaccines

The first step in the preparation of bacterial vaccines, such as typhoid vaccine, is similar in that an artificial culture of the organism is first prepared. In this case, however, a solid medium is usually used, such as nutrient agar. If a mixed bacterial vaccine is being made, the individual species of organisms are cultured separately and later mixed. The culture is incubated at a suitable temperature for at least 24 hours until sufficient growth has taken place. The bacteria are then collected under aseptic conditions (free from pathogens) by pouring sterile salt solution on the culture and gently shaking or scraping until the microorganisms have been suspended. If a liquid medium has been used for culture, it is centrifuged, and the bacterial sediment suspended in sterile isotonic saline. In either case, the crude product is shaken to produce a uniform suspension, and the bacterial concentration determined. The suspension is diluted with sterile isotonic saline, and the bacteria killed by a germicide or by heat; finally, a preservative is added.

In the preparation of virus vaccines, such as yellow-fever vaccine, the general principle is similar, with the important exception that most virus vaccines are weakened living virus preparations. The living virus is inoculated into the embryo region of hens' eggs that have been incubated for eight days. After a further six days' incubation, the embryo is removed, homogenized, and the vaccine separated. As these vaccines are unstable, they are freeze-dried and reconstituted at the time of inoculation. Poliomyelitis vaccine is usually prepared from viruses inactivated with formaldehyde. Some virus vaccines, such as measles and rubella, can now be prepared using human cells as the culture medium.

Human blood-plasma fractions. A number of proteins of therapeutic value can be concentrated from human plasma; these include albumin, immunoglobulins or gamma-globulins that contain antibodies to various viruses, and blood coagulants.

Careful addition of ethanol to protein solutions at temperatures in the range −8 to 0° C (18 to 32° F) causes precipitation without gross damage to molecular structure. The various proteins in plasma have their minimum solubilities at differing acid or base (pH) values and salt concentrations and can be separated from each other by a series of precipitations in which these factors are controlled. Ethanol concentrations, protein concentrations, and temperatures must also be adjusted to precise values. Because protein molecules are unstable, excessive frothing or agitation, local high ethanol concentrations, or local temperature rise (due to the heat of mixing of ethanol and water) must be avoided. Precipitates are separated in refrigerated centrifuges, and the final products are freeze-dried, which removes residual ethanol and makes them more stable for storage and distribution. Albumin and immunoglobulins are sufficiently stable to be redissolved and dispensed in solution.

Difficulties in processing human blood fractions

Processing is complicated for several reasons: proteins, being unstable in heat, can be sterilized only by filtration; large volumes of some products may be given intravenously, thus precluding the use of preservatives; blood coagulant substances are extremely unstable in plasma even at refrigerator temperatures, necessitating their preparation within a few hours after collection of the blood from the donors; and certain diseases, of which blood donors may be carriers, notably serum hepatitis, can be transmitted in the fractions. Albumin is, however, stable enough to withstand heating for 10 hours at 60° C or 140° F (which inactivates the hepatitis virus), and heated albumin preparations are now largely replacing whole plasma in transfusion therapy. It has been found that immunoglobulin fractions do not transmit this virus. Control of products includes tests for the presence of microorganisms, pyrogens, and toxic substances; for pH and salt content; for protein composition; and, where appropriate, for heat stability, antibody activity, or coagulation function.

Steroid hormones. Another important group of pharmaceuticals widely distributed in plants and animals is the steroid hormones (see also CHEMICAL COMPOUNDS: *Steroids*). The plant by-product called diosgenin is by far the most common starting material for the preparation of steroids. It is obtained from whole tubers of wild yams; these are broken up in a hammer mill, and the resultant mash is allowed to ferment for two days. It is then dried in the sun, hydrolyzed with mineral acid, and the diosgenin extracted with an immiscible organic solvent (one that will not mix with water), such as light petroleum or heptane.

The hormone estrone is usually manufactured from pregnant mares' urine, in which it is present almost entirely as the ethereal sulfate. This is hydrolyzed to estrone by acidification of the urine and heating, after previously covering with a layer of immiscible solvent, such as toluene. Estrogens are extracted from the solvent, and the estrone is separated from the other estrogens.

Progesterone was formerly obtained from the ovaries of

sows or whales. It is now usually manufactured synthetically by a complex series of chemical reactions from various materials such as stigmasterol (obtained from soybeans) or diosgenin. Testosterone is also manufactured from diosgenin.

Further medicinal steroids, such as ethisterone, methandienone, and methyltestosterone, are obtained from an intermediate of testosterone production.

Sources of cortisone

Cortisone is manufactured from ergosterol, from progesterone (itself obtainable from stigmasterol or from diosgenin), and from other plant by-products, by routes involving both chemical and microbiological transformations. Hydrocortisone, in turn, is produced from cortisone. Deoxycorticosterone is made as the acetate from progesterone.

Vitamins. These were formerly extracted from natural sources, but now the majority are made by synthetic methods. Not only are the synthetic vitamins cheaper and purer, but their production is easier than extraction of natural vitamins.

Preparation by chemical synthesis. Two or more of the methods discussed below are usually involved in the manufacture of any one drug. There are many different possible routes for the production of pharmaceuticals. It is not often that the simplest possible theoretical pathway is the initial choice of the manufacturer; many different methods are chosen by manufacturers depending upon the starting materials and equipment available to them.

Acetylation. This is a special case of acylation, a process in which alcohols, phenols, and primary and secondary amines yield acyl derivatives by displacement of the hydrogen atom of the hydroxy, primary amino, or secondary amino group by an acyl radical. The commonest acyl radical by far is the acetyl radical, CH_3CO-. The usual acetylating agent is acetic anhydride, $(CH_3CO)_2O$, but glacial acetic acid, CH_3COOH, and acetyl chloride, CH_3COCl, are also used. With acetic acid and alcohols or phenols, it is usual to use a dehydrating agent, such as concentrated sulfuric acid or anhydrous sodium acetate, in order to eliminate the water formed in the reaction. With acetyl chloride, a base such as dimethylaniline or pyridine is added to remove hydrogen chloride as it is formed.

Acetylation is often used in drug production since it reduces the toxicity of amines. Thus the drug paracetamol is much less toxic than *p*-aminophenol. Paracetamol is manufactured from *p*-nitrophenol.

Addition. This chemical reaction can only take place with unsaturated compounds, but the addition products vary greatly in nature. The mode of addition, where there are two possibilities, is determined by the structure of the unsaturated compound and the mechanism of addition.

The anesthetic drug halothane is manufactured from trichloroethylene (see below *Dehydrohalogenation*) by an addition reaction with hydrogen chloride. The product is treated with hydrogen fluoride; reaction with bromine then gives halothane.

The manufacture of the anesthetic drug trichloroethylene involves, as a first step, the addition of chlorine to acetylene, followed by dehydrohalogenation.

Alkylation and arylation. Alkylation is a general term for the introduction of an alkyl group into a compound and is more specifically known as methylation, ethylation, propylation, butylation, etc., depending upon the alkyl group inserted. Alkylation of an alcohol or a phenol results in an ether, while alkylation of a thiol gives a sulfide; primary and secondary amines can also be alkylated. The usual alkylating agents are alkyl sulfates and halides. The alcohol, phenol, thiol, or amine is dissolved in sodium hydroxide solution, and dimethyl sulfate, for example, is added to yield the alkylate.

In addition to the above types of compounds, any compound containing "active hydrogen" can be alkylated. Thus, compounds containing a methylene ($-CH_2-$) or a methyne ($=CH-$) group adjacent to a group such as carbonyl ($-CO-$) contain active hydrogen.

The important drug codeine, widely used as a cough suppressant and pain-killer, occurs naturally as an alkaloid in opium, together with morphine. Demand for codeine is such that much of it is manufactured by methylation of morphine. Morphine is dissolved in potassium hydroxide solution, and the solution is treated with methyl iodide to produce codeine.

The well-known constituent of tea and coffee, caffeine, used as a stimulant drug, is produced either from tea wastes or synthetically from the organic chemicals theophylline or theobromine. These three materials are closely related, and methylation of either theophylline or theobromine or a mixture of the two results in the fully methylated caffeine.

Synthesis of barbiturates

The barbiturates, widely used as sedative–hypnotic drugs, provide many examples of alkylation and arylation (introduction of one or more aromatic radicals, such as phenyl or naphthyl, to a molecule). The general method for their synthesis is as follows. Diethyl malonate (a malonic-acid ester, which contains an active methylene group) is treated with an alcoholic solution of sodium ethoxide, and the monosodium derivative produced is converted by reaction with an alkyl halide into an alkylmalonic ester. This is either used as such in the preparation of the barbiturate, or, as is usually required, a second alkyl group, either the same or different, is introduced in the same way. The alkylmalonic or dialkylmalonic ester is then condensed (see below *Condensation and cyclization*) with urea, in the presence of sodium ethoxide, to form the sodium salt of barbituric acid. This is either used as such, since sodium barbiturates are usually very soluble in water, or the free barbituric acid is liberated by acidification. Barbiturates containing aromatic groups (phenobarbital and mephobarbital are the only common examples) must be made by a modified route, because of the relative lack of reactivity of aryl halides. For both these drugs, ethyl phenylacetate is condensed with ethyl oxalate to form ethyl oxalophenylacetate, which, when distilled, decomposes to give diethyl phenylmalonate. Ethylation is then effected to yield diethyl ethylphenylmalonate. For the manufacture of phenobarbital, this ester is condensed with urea as in the scheme above.

Amination. Two important methods of amination (the process of adding an amino ($-NH_2$) group to a compound) are reduction and ammonolysis. There are many reduction methods, but the most common is the reduction of the corresponding nitro compound by means of iron and acid. An example of ammonolysis is the preparation of aniline by heating chlorobenzene with excess aqueous ammonia, in the presence of cuprous oxide, at about 200° C (392° F) under pressure. Amination enters into the synthesis of many pharmaceuticals; a good example is amphetamine. Phenylacetone is treated with ammonia and hydrogen in the presence of a nickel catalyst, and reductive amination takes place, forming amphetamine.

Arsonation. Arsonation applied to aromatic compounds is closely analogous to nitration (see below). For example, *p*-aminophenylarsonic acid (arsanilic acid) is made by heating aniline with concentrated arsenic acid. A method other than arsonation that achieves the same result is Bart's reaction, in which a diazonium salt is coupled with an alkali arsenite, usually in the presence of a catalyst such as copper powder. Sodium phenylarsonate, for example, can be made from benzenediazonium chloride by Bart's reaction. The drug acetarsol is manufactured from phenol by arsonation followed by nitration, reduction, and acetylation.

Carboxylation. This reaction is best exemplified by the manufacture of salicylic acid, used in external preparations as a skin softener. Dry sodium phenoxide is heated with carbon dioxide, under pressure, at about 130° C (266° F); the resulting sodium salicylate is dissolved in water, and the acid precipitated by acidification with hydrochloric acid. Salicylic acid is used in medicine as the free acid, as its acetyl derivative, aspirin (see below *Esterification*), and as its salts, of which the most important is sodium salicylate, used in the treatment of rheumatism. To obtain the latter product, a weighed amount of salicylic acid is made into a paste with water, and the equivalent quantity of pure sodium bicarbonate is added. The mixture is stirred and warmed until a clear solution is obtained, and this is evaporated to dryness; the sodium salicylate is purified by crystallization from alcohol.

Condensation and cyclization. Condensation is difficult

to define, but it involves the linking together of two or more organic molecules (identical or different), with or without the elimination of a simple molecule such as water, ethanol, or a hydrogen halide. Reactions involving condensations that can be defined more precisely, such as esterification, etherification, or polymerization, are usually not described as condensations.

Hexylresorcinol, used for destroying intestinal worms, provides a simple example of a condensation. It is produced by condensing resorcinol with hexanoic acid, in the presence of anhydrous zinc chloride, at about 130° C (266° F). Water is eliminated, and the resulting ketone is reduced by means of zinc amalgam and dilute hydrochloric acid. The sedative meprobamate involves two condensations in its production. Diethyl methyl-*n*-propylmalonate is reduced with lithium aluminum hydride, $LiAlH_4$, to the corresponding dihydric primary alcohol; this is condensed with carbonyl chloride, $COCl_2$, with elimination of hydrogen chloride to give a dichloroformate; condensation of this with ammonia, again with elimination of hydrogen chloride, gives meprobamate. Cyclization is a special case of condensation; internal condensation takes place in one molecule, with resulting formation of a cyclic molecule. The compound *o*-iminodibenzyl, from which the antidepressant drugs desipramine and imipramine are produced, illustrates cyclization. The iminodibenzyl is condensed, in the presence of sodamide, $NaNH_2$, with 1-chloro-3-methylaminopropane to give desipramine or with 1-chloro-3-dimethylaminopropane to give imipramine.

Production of meprobamate

Dehydration. This reaction, as the name indicates, involves the loss of the elements of water. If this takes place internally in a molecule, the result is usually an alkene or a cyclic compound. If dehydration takes place between two molecules (condensation), the product is usually an ether.

Diethyl ether (ordinary ether), used as an anesthetic, is produced by indirect condensation of ethanol. A mixture of concentrated sulfuric acid and ethanol is heated to 140° C (284° F), when ether distills over; more ethanol is run into the mixture at the same rate as the ether distills. The ethanol esterifies with the sulfuric acid to form ethyl hydrogen sulfate. This ester, on heating with ethanol at 140° C, gives ether and sulfuric acid back again.

Vinyl ether is also used as an anesthetic and provides a more complex example of dehydration. Ethylene chlorohydrin with sulfuric acid gives 2,2′-dichlorodiethyl ether, and this is dehydrohalogenated (see below) by means of solid potassium hydroxide to produce divinyl ether.

Dehydrohalogenation. This process may be defined as removal of the elements of hydrogen halides with the resultant formation of an unsaturated bond. It is usually effected by means of a base such as calcium oxide or hydroxide. Examples of its application are the production of the anesthetic trichloroethylene and the drug tetrachloroethylene (used for destroying intestinal worms), both involving the elimination of the elements of hydrogen chloride. Acetylene, readily available and comparatively cheap, is used as the starting material. The reaction with chlorine is dangerously explosive under ordinary conditions, but the reaction can be controlled to give acetylene tetrachloride. This readily yields trichloroethylene. In turn, trichloroethylene takes up one molecule of chlorine to form pentachloroethane, and this, treated in the same way as acetylene tetrachloride, produces tetrachloroethylene.

Esterification. The addition of an acid to an alcohol generally results in esterification; that is, the production of an ester plus water. The reaction is reversible, and the rate of the reverse reaction (see below *Hydrolysis*) will increase as the concentration of water (or ester) increases. Hence, both the anhydrous acid and the anhydrous alcohol are normally used in order to produce good yields. A catalyst such as dry hydrogen chloride speeds up the attainment of equilibrium (but will not affect the point of equilibrium). Concentrated sulfuric acid may be used, but it affects the point of equilibrium since it is a dehydrating agent as well as a catalyst; it may also dehydrate the alcohol and so must not be used in too high a concentration.

Production of acetylsalicylic acid

Acetylsalicylic acid (2-acetoxybenzoic acid, or aspirin) is the most widely used of all synthetic drugs. It is usually manufactured by the action of acetic anhydride on salicylic acid. As aspirin is only slightly soluble in water, it is easily separated by dilution of the reaction mixture with water. It is readily crystallized from aqueous alcohol, benzene, or light petroleum or from mixtures of these or other solvents with glacial acetic acid (to minimize hydrolysis).

Halogenation. This may involve reactions of addition, substitution, or replacement of a group of atoms. The catalysts used include halogen carriers such as iron, antimony, or phosphorus, all of which exert two valences as halides. They alternately add on and give up halogen to carry on the reaction. Since the halogens form unstable interhalogen compounds, chlorine, bromine, and iodine can also be used as catalysts. Under suitable conditions, substitution of alkenes with halogens takes place, and not addition. The synthetic oral estrogen, chlorotrianisene, for example, is produced from deoxyanisoin by reaction with 4-methoxyphenyl magnesium bromide, followed by subsequent dehydration and substitutive chlorination with a solution of chlorine in carbon tetrachloride. Gamma benzene hexachloride (gammexane), a delousing agent, is manufactured by the addition (not substitution) of chlorine to benzene. The product is a mixture of four of the nine possible stereoisomers, and the gamma isomer, which is the most active as an insecticide, is separated from the others.

Hydrolysis. This term is applied to reactions wherein water effects a double decomposition with another compound, hydrogen going to one component and hydroxyl to the other:

$$XY + H_2O = HY + XOH$$

The reversal of esterification is an example of hydrolysis. Catalysts that speed up esterification also accelerate hydrolysis; many enzymes will also catalyze hydrolytic reactions.

A pharmaceutical example is the production of propylene glycol. Propene is oxidized and converted by acid-catalyzed hydrolysis to propylene glycol. The pharmaceutical substance called wool alcohol is made by hydrolysis of wool grease.

Inorganic substances. Helium, nitrogen, oxygen, sulfur, and some inorganic compounds still find a place in medical practice. These are prepared by the general methods used in inorganic chemistry; however, they must comply with pharmacopoeial standards of purity, whereas those used for nonmedical purposes are not usually purified up to these standards.

An example is the manufacture of ferric ammonium citrate, used in iron-deficiency anemias. The first step is preparation of freshly precipitated ferric hydroxide by slow addition, with constant stirring, of a ferric-salt solution (*e.g.*, ferric sulfate) to an alkali solution such as sodium hydroxide:

$$6NaOH + Fe_2(SO_4)_3 = 3Na_2SO_4 + 2Fe(OH)_3.$$

The ferric hydroxide is collected and washed and, without being dried, is stirred with enough citric acid solution to dissolve almost all of it. A slight excess of ammonia solution is then added, and the residual ferric hydroxide filtered off. The clear, reddish-brown filtrate is evaporated to a syrup, a little ammonia being added occasionally to maintain a slight excess. The syrup is painted on glass plates, dried at a temperature below 40° C (104° F), and scraped off as scales. Ferric ammonium citrate is also produced in the form of granules. The medicinal substance is dark red in colour and contains about 21 percent iron. Green ferric ammonium citrate is also made; it contains about 16 percent iron.

Mercuration. This is usually effected by reaction with a solution of mercuric acetate in acetic acid or alcohol; substitution of a hydrogen atom by the acetoxymercuri group, $-HgOCOCH_3$, results. The topical antiseptic acetomeroctol, for example, is prepared by interaction of a complex organic chemical compound with the calculated quantity of mercuric acetate in a 50 percent aqueous ethanolic solution containing 5 percent acetic acid. The mixture is maintained at room temperature for one to two weeks.

Nitration. Medicinal nitro compounds are invariably aromatic carbocyclic or heterocyclic derivatives. These are

usually nitrated with a mixture of concentrated sulfuric acid and concentrated or fuming nitric acid (mixed acid). Other nitrating agents include oxides of nitrogen and acetyl nitrate, $CH_3CO_2NO_2$.

Nitrofurantoin is prescribed as a urinary antiseptic and provides an example of nitration. Furfural is reacted with acetyl nitrate, and the resulting triacetate is treated with pyridine to give 5-nitro-2-furaldehyde diacetate. This is condensed (with elimination of acetic acid) with 1-aminohydantoin to yield nitrofurantoin.

Oxidation. This reaction is very widely used in the preparation of chemical compounds. Oxidation may be defined as the loss of electrons or hydrogen or the gain of oxygen; reduction (see below) has the opposite meanings.

Oxidizing agents

There are many oxidizing agents available; the choice depends upon the extent to which oxidation is desired. Air, oxygen, ozone, peroxides, chromium trioxide, chlorates, dichromates, permanganates, nitric acid, and sulfuric acid are all common examples. Conversion of the alkaloid nicotine to the vitamin nicotinic acid provides an example of oxidation.

Polymerization. Polymerization reactions may be broadly divided into addition polymerization and condensation polymerization (polycondensation). The term polymerization strictly refers to addition polymerization but is often used to include all types of polymerization reactions. Addition polymerization may be defined as the joining together of two or more molecules of a compound in such a way that the molecular weight of the polymer produced is a multiple of that of the original compound. Addition polymerization is usually induced either by heat, pressure, or a catalyst or by a combination of these.

A pharmaceutical example is the preparation of paraldehyde, used as a sedative drug. A small proportion of concentrated hydrochloric acid is added to acetaldehyde, causing three molecules of the liquid to polymerize to form a molecule of paraldehyde, a saturated heterocyclic system. The reaction involves the mutual saturation of three unsaturated carbonyl groups.

Condensation polymerization

Condensation polymerization involves a similar joining together of molecules, but this time a simple molecule, such as water, is condensed out. Diethylene glycol, for example, can be made by heating glycol with a dehydrating agent. Higher polyethylene glycols can be made similarly, such as polyethylene glycol 400, with a molecular weight of about 400, and polyethylene glycol 4000, with an average molecular weight of about 3,400. These two polymers are employed in the preparation of a water-soluble ointment base.

Reduction. This reaction, the opposite of oxidation, has very wide applications in the production of pharmaceuticals. Examples of reducing agents are aluminum amalgam, hydriodic acid, hydrogen and a catalyst, hypophosphorous acid, lithium aluminum hydride, metal and acid, sodium and alcohol, stannous chloride, and zinc dust and water. Reduction is employed in the synthesis of racemic menthol, an anti-itching agent. It is produced by the catalytic hydrogenation of the aromatic analogue, thymol, or of the saturated ketone, menthone, or of the unsaturated ketone, pulegone.

Sulfation. This consists of direct linkage of a sulfate group to carbon, yielding an acid sulfate, $ROSO_2OH$, or a sulfate, $ROSO_2OR$. The most usual sulfating agents are concentrated sulfuric acid, oleum, sulfur trioxide, chlorosulfonic acid, or sulfamic acid, NH_2SO_3H.

The surface-active agent sodium lauryl sulfate, for example, is produced from commercial lauryl alcohol (mainly dodecanol) by reaction with chlorosulfonic acid. The resulting lauryl hydrogen sulfate is neutralized with sodium hydroxide.

Sulfonation. This consists of direct linkage of the sulfonic-acid group, $-SO_2OH$, or salts thereof, or of the sulfonyl halide group, $-SO_2X$, to a carbon or nitrogen atom. The latter are known as *N*-sulfonates or sulfamates. The same reagents are employed as are used for sulfation, above.

In the production of sulfonamides, the intermediate 4-acetamidobenzenesulfonyl chloride is made from acetanilide and chlorosulfonic acid. Most sulfonamides are made by the condensation of this intermediate with the appropriate primary amine, followed by removal of the acetyl group by alkaline hydrolysis. The hydrolysis is not usually difficult in practice, as the $-CO-NH-$ bond is more easily broken than the $-SO_2-NH-$ linkage. After the alkaline hydrolysis, the sulfonamide is precipitated by neutralization.

Complex chemical conversions. Ascorbic acid (vitamin C) provides an example of a complex synthesis. It is manufactured from D-glucose; this is converted by means of microbiological oxidation to a keto acid, which is catalytically reduced to L-idonic acid. Microbiological oxidation converts this to 2-OXO-L-gulonic acid. The 2-OXO-L-gulonic acid is converted to its methyl ester, which is isomerized and cyclized to L-ascorbic acid.

PREPARATION OF DOSAGE FORMS

Before describing individual types of dosage forms, some consideration of the process of size reduction or comminution is necessary. The degree of comminution varies according to the particular drug and the method of extraction used. Thin leaves and certain substances such as aloes or tolu need no treatment at all. Others may be cut, crushed, broken into small pieces, or sliced thin. Finally, the drug may be powdered or ground to one of five different degrees of fineness. A number of grinding machines are employed for powdering, the most important being the disintegrator, edge-runner mill, ball mill, and end-runner mill.

Process of comminution

Solutions. Many pharmaceuticals are simply solutions of a medicament in water, alcohol, ether, glycerin, or some other solvent. Various flavoured waters (*e.g.,* chloroform or peppermint water) may also be used as solvents. After mixing, many solutions require additional treatment such as sterilization or addition of antimicrobial agents, stabilizers, or buffers.

Spirits are solutions of a volatile substance in alcohol or a mixture of water and alcohol. Elixirs are clear, sweetened hydroalcoholic (alcohol and water solvent) liquids for oral use, usually containing potent or nauseous drugs. Simple elixirs are prepared by solution, while those made from plant drugs are produced by maceration or percolation (described below). Tinctures, solutions of a nonvolatile substance in alcohol or a mixture of water and alcohol, are manufactured by simple solution, maceration, or percolation.

In maceration, the drug, suitably prepared, is placed, together with the extracting solvent (known as the menstruum), in a closed vessel and left for a defined period, usually three to seven days, with occasional shaking. The product is then filtered, the residue on the filter (known as the marc) is pressed to avoid loss, and the mixed filtrates are clarified, either by subsidence (settling out) or filtration.

Percolation

In percolation, the powdered drug is first moistened with the extracting solvent, allowed to stand for a defined period, and then carefully packed in a conical or cylindrical vessel with a perforated false bottom and a tap at the base known as a percolator. Sufficient extracting solvent is added to saturate the drug, and, when liquid begins to drip from the percolator, the tap is closed and sufficient solvent added to form a layer above the drug. Maceration is then allowed to proceed for a defined period (usually 24 hours), followed by slow percolation, more solvent being added as necessary, until the desired volume of percolate is obtained, which is then clarified by subsidence or filtration.

Fluid extracts are solutions, in alcohol, water, ether, or a mixture of these, of the active part of a vegetable drug so prepared that one cubic centimetre of extract has the same strength as one gram of the dry drug. More concentrated than tinctures, fluid extracts are made by percolation or maceration.

Numerous other types of solutions are used in pharmacy, such as collodions (viscous liquids for external application) or liniments (liquids for external application, usually consisting of a drug in oil or soap solution).

Semisolid dosage forms. Ointments are preparations for external use, intended for application directly on the

skin or on a surgical dressing. An ointment base is chosen depending upon whether the final product is intended for absorption by the skin or not. Oily- or fatty-base ointments may have hard, soft, or liquid paraffin bases or mixtures of these in such proportions as will render an ointment of suitable consistency. Anhydrous lanolin (wool fat), also included in this group, is particularly suitable whenever a large volume of water or aqueous solution must be incorporated into the ointment, as anhydrous lanolin forms a stable water-in-oil emulsion. Anhydrous lanolin is readily absorbed by the skin and is often mixed with paraffins when both absorption and protection are desired. Emulsifying bases (in which minute globules of fatty substances are suspended) include anhydrous lanolin and hydrophilic petrolatum, which has similar properties. (The solid particles of a hydrophilic substance strongly attract and hold molecules of water.) Other ointment bases are emulsion bases, mixtures of oily substances and water; and water-soluble bases, generally prepared from polymers of ethylene glycol.

When the active medicament in the ointment is a solid, it is powdered as finely as possible, and then incorporated by the principle of geometric dilution (see below *Solid dosage forms*). The ingredients of ointment bases are often mixed together by fusion in a water bath, followed by stirring until cold, the drug being incorporated at a suitable stage. In manufacturing practice, ointments are usually improved by being passed through a roller mill; this ensures uniformity and smoothness. Creams, with the same therapeutic purposes as ointments, are solid or semisolid emulsions of the water-in-oil or oil-in-water type, giving greasy or nongreasy preparations, respectively. Pastes, similar to ointments, are usually stiffer, less greasy, and more absorptive, due to a higher proportion of powdered ingredients, such as calcium carbonate, talc, or zinc oxide. They are made by similar methods but with a glycogelatin, paraffin, or starch basis, or a mixture of these. Jellies used for similar purposes are made with an aqueous gel such as acacia, cellulose derivatives, chondrus, gelatin, gelatinized starch, or tragacanth as a base.

Solid dosage forms. Solid dosages, such as tablets, have many advantages over other types: greater stability, less risk of chemical interreaction between different medicaments, smaller bulk, accurate dosage, and ease of production.

Powders

Powders intended for internal use are usually mixtures of two or more ingredients. If two ingredients are present in unequal quantities, then the lesser ingredient (usually the drug) is mixed with an equal weight of the greater ingredient. Next, the resulting mixture is combined with an equal weight of the greater ingredient in steps until the mixture is complete. This process of geometric dilution (or trituration) is essential in order to produce a homogeneous powder. Powders may be sold in a bulk form or in individual packets. Granules, a dosage form related to powders, are particularly suitable for the preparation of solutions or mixtures of drugs, such as antibiotics, that are unstable in the presence of water. Cachets, occasionally used for administration of powdered drugs with an unpleasant taste, consist of shells made of gelatinized starch paste, the powder being placed inside, and the cachet swallowed. More common today is the hard capsule, in which the powder is enclosed in a shell of hard gelatin. Semiliquid and liquid drugs are often enclosed in a soft capsule with a soft gelatin shell.

Pills. Before the machine-made compressed tablet, pills were a very popular solid dosage form, being prepared at the dispensing bench by the pharmacist. Today pills are rarely prescribed, though some popular types are manufactured by machine. The powdered ingredients are mixed together with a binding agent, such as acacia or tragacanth, and are then made into a plastic mass by incorporation of any liquid drugs and addition of an inert liquid. The resulting mass, known as a pill mass, is then rolled into spheres and coated with talc, gelatin, or sugar.

Tablets. Tablets, by far the most common method of administration of drugs, are only rarely made by compression of the drug alone (*e.g.,* potassium bromide tablets). Usually, the drug is mixed with suitable diluents, such as dextrin, lactose, salt, starch, or synthetic substances, designed to ensure disintegration of the tablet in the body. To prevent sticking in the machine, a lubricant such as liquid paraffin, stearic acid, talc, or a synthetic substance is usually added. Furthermore, it is essential that the tablet machines are fed with the drug mixture in a free-flowing form to ensure complete filling of the molds. To achieve this, the drug mixture is customarily granulated by mechanically forcing pellets of the mixture through a sheet of perforated metal. The granulated mixture is fed into the tablet machine, which feeds the correct dose into a cavity, the mixture then being compressed by means of a punch that fits into the cavity. To be successful, the tablet maker must choose correct diluents and lubricants, prepare suitable granules, and obtain the right degree of compression in the tablet machine. Excessive compression may mean that the tablet will not disintegrate in the body; insufficient compression results in fragile tablets that may break, causing inaccurate dosage. Coatings of various types may be applied to the tablet—to protect the ingredients from deterioration, to hide the taste of certain drugs, to control the release of the drug from the tablet, or to produce a more attractive tablet. For sugarcoatings, a concentrated sucrose syrup containing suspended starch, calcium or magnesium carbonate, or other suitable substance is applied, each successive layer being dried before the application of the next. After the final layer is dried, it is highly polished to give an elegant finish. Sugarcoatings provide both protection and a sweet taste. The chief drawback to sugarcoating is the long time involved. This led to the development of film coating, in which a very thin transparent film, usually a cellulose derivative, is applied. Enteric coating is designed to resist solution in the stomach and to dissolve in the more alkaline intestinal fluid. Many substances have been used for enteric coatings, one of the more recent being cellulose acetate phthalate (cellacephate). In the manufacture of layered tablets, incorporating two or more drugs, a compressed tablet is fed to a second machine where another layer is compressed around it. In this way, drugs normally incompatible may be formulated in the same tablet.

Tablet coatings

Other solid dosages. Troches, also known as lozenges or pastilles, disintegrate or dissolve in the mouth, slowly releasing the active drug. The base usually consists of a mixture of sugar and gum or gelatin. Lozenges are generally manufactured by compression techniques, while pastilles are fabricated by fusion and the use of molds. Dry extracts are prepared by the methods described above for fluid extracts, followed by evaporation, usually under reduced pressure, either to a pilular consistency or to dryness. Dry extracts are usually granulated by being passed through a sieve and may be used for the preparation of tablets. Suppositories are solid, uniformly medicated masses designed for introduction into the rectum. Various bases are used in their preparation, but theobroma oil is the most common. Solid at room temperature, it melts a few degrees below body temperature. Suppositories are manufactured by the use of molds, together with fusion of the suppository mass or cold compression. Pessaries are suppositories intended for introduction into the vagina, while bougies are designed for insertion into the urethra, nostrils, or ears.

Sterile medicaments. Some drugs cannot be given orally, because they are decomposed by the digestive processes; these drugs must be administered by injection, intravenously, intramuscularly, or subcutaneously (under the skin). It is imperative that such injected solutions be free from micro-organisms or toxic agents; other pharmaceutical preparations that must be sterile include eye drops, eye lotions, eye ointments, implants, powders to be applied to wounds or body cavities, and any solution or preparation to be used in surgical operations, such as surgical dressings, ligatures, and sutures.

Sterilization methods

There are several main methods for preparing sterile products. Steam sterilization is carried out in an enclosed chamber (autoclave). The material is distributed into its final containers, which are sealed to exclude microorganisms. The containers are then placed in the autoclave and subjected to saturated steam under pressure at a temperature of 121° C (250° F) for 15 minutes. Substances stable

in heat may be sterilized at higher temperatures, for which a shorter heating period is sufficient; conversely, lower temperatures require longer sterilization periods.

Pharmaceuticals in which the solvent is not water cannot be sterilized by steam, so dry heat at temperatures of 140° C (284° F) or higher is used, with exposure times ranging from one to four hours. The temperature–time relationship is similar to that for steam sterilization. Another technique is gaseous sterilization, in which the material is exposed to a vapour or gas such as ethylene oxide; this method is much used for foods, biologicals, and medical equipment (*e.g.,* cotton wool, syringes, needles, and tubing). Radiation sterilization generally involves exposure to ultraviolet or gamma radiation or high-energy electrons; its use is limited mainly to the production of sterile medicaments and apparatus on a large scale. The final method, aseptic manipulation, is not actually a sterilization process. Here, the separate ingredients of the pharmaceutical are all available in sterile form and merely require compounding without microbial contamination.

Injections are aqueous or oily solutions, suspensions or emulsions, prepared by normal methods, with special care taken to remove all extraneous particulate matter. Injections must be sterilized by one of the methods given above. Some aqueous injections are not stable and so are prepared at the time of use by adding sterile water to the sterile drug.

Eye drops, eye lotions, and eye ointments are all prepared by general methods, but they must be sterile. The pH value (acid–alkaline content) of eye drops and eye lotions is important; a compromise must be made between the ideal value of 7.4 (almost neutral) and the value of greatest stability or therapeutic activity. Implants or pellets are prepared either by fusion and molding of the drug or by compression of the sterile crystals.

Dispersions. Aerosols were formerly defined as colloidal systems consisting of very finely subdivided liquid or solid particles dispersed in a gas. Today, the term aerosol, in general usage, has become synonymous with a pressurized package. For pharmaceutical purposes aerosols may be divided into two types. Space sprays disperse the medicament as a finely divided spray with particles not exceeding 50 microns (0.05 millimetre, or 0.002 inch) in diameter. Surface-coating aerosols produce a coarse or wet spray and are used to coat surfaces with a residual film. Propellants used in aerosols are of two main types: liquefied gases and compressed gases. The former consist of easily liquefiable gases such as halogenated hydrocarbons. The drug is dissolved in the liquefied gas or in a mixture of the gas and a suitable solvent. When these are sealed into the container, the system separates into a liquid and a vapour phase and soon reaches an equilibrium. The vapour pressure pushes the liquid phase up the standpipe and against the valve. When the valve is opened by pressing down, the liquid phase is expelled into air at atmospheric pressure and immediately vaporizes, leaving an aerosol of the drug. The pressure inside the container is maintained at a constant value as more liquid changes into vapour. When compressed gases are used as the propellant, the pressure falls steadily as the contents of the aerosol are used, and for this reason liquefied gases are used whenever possible. Pharmaceutical aerosols include solutions, suspensions, emulsions, powders, and semisolid preparations. The products include inhalation aerosols, spray-on bandages, creams, and ointments. The application of these latter to wounds and burns is obviously advantageous, as rubbing is eliminated, and the fine film produced promotes rapid absorption. Inhalation aerosols often include a metering valve, so that measured quantities of drug may be administered; these are rapidly replacing old hand sprays.

Aerosols

Sprays, solutions of drugs in aqueous or oily solutions, are applied by means of an atomizer to the mucous membranes of the nose or throat. Oily solutions are no longer considered desirable, and the ideal spray is an aqueous solution isotonic (equal in osmotic pressure) with nasal secretions and of the same pH.

Since some drugs are insoluble in all solvents suitable for medicinal use, they must be administered either as a solid dosage form or as a suspension. Suspensions are chemically more stable than solutions. Apart from aerosols, pharmaceutical suspensions almost always consist of a finely divided solid dispersed in a liquid. The state of subdivision varies from colloidal particles to particles that only slowly subside on standing. Suspensions should not cake on standing, and the solid phase should readily redisperse on shaking; suspensions should pour fairly easily, so the viscosity must not be too high. Gels are special suspensions, namely, those of hydrated drugs in an aqueous medium, in which the particle size is colloidal or nearly colloidal. Lotions, such as calamine lotion, are suspensions for external use only. Magmas and milks are thick, viscous, aqueous suspensions of insoluble inorganic compounds; the particle size is usually larger than in gels. Bentonite magma, for example, is produced by hydration of bentonite, a colloidal hydrated aluminum silicate, and is used as a suspending agent, as, for example, in calamine lotion. Milk (cream) of magnesia, a magnesium hydroxide mixture, is an aqueous suspension of magnesium hydroxide. Mixtures, in the wider sense, are any combination of drugs prescribed and dispensed for internal use but, in the narrow sense, are official mixtures of standard composition. Official mixtures are liquid preparations of one or more solid drugs dissolved or suspended in water; they often contain a thickening or suspending agent such as tragacanth. Thus, most mixtures may be regarded as a special type of suspension.

Suspensions

Mucilages are thick, viscous, aqueous solutions of gums, frequently containing a preservative such as chloroform or benzoic acid. As they are colloidal in nature, they fall between true solutions and suspensions. Acacia and tragacanth mucilages are the best known examples and are used to aid in suspending insoluble solids in liquids.

Emulsions consist of one liquid dispersed in another. Pharmaceutically, they are intended for internal use and consist of small globules dispersed in water. Oil-in-water emulsions will mix with water, whereas water-in-oil emulsions only mix with oils. However well two immiscible liquids are mixed together, on standing they will separate into two layers. To prevent separation, an emulsifying agent is used. Emulsifying agents can be divided into three groups: finely divided solids such as bentonite and magnesium aluminum silicate; natural emulsifying agents such as cholesterol, gelatin, acacia, methylcellulose, pectin, and tragacanth; and synthetic emulsifying agents such as the anionic sodium lauryl sulfate, the cationic benzalkonium chloride, and the nonionic polyethylene glycol 400 monostearate. A preservative such as chloroform is usually added to emulsions. The production of emulsions depends on the emulsifying agent used. Equipment includes a wide variety of agitators, colloid mills, homogenizers, and ultrasonic devices.

Radioactive dosage forms. A radiopharmaceutical is a medical product incorporating a radioactive isotope. Radiopharmaceuticals are widely used for various diagnostic tests and to a lesser extent as therapeutic agents. Radioactive iodine in the form of sodium iodide has been extensively used in the diagnosis of thyroid disorders. The iodine isotopes used are iodine-131 and iodine-125, orally administered as a sodium iodide solution; the thyroid uptake of radioactive iodine is measured with a measuring instrument placed close to the thyroid gland. Vitamin B_{12} containing cobalt-57, cobalt-58, or cobalt-60 is employed for the diagnosis of pernicious anemia. Chlormerodrin injection containing mercury-203 is used for renal (kidney) scanning and brain scanning. (L.M.At.)

Explosives

An explosive is any substance or device that can be made to produce a volume of rapidly expanding gas in an extremely brief period. There are three fundamental types: mechanical, nuclear, and chemical. A mechanical explosive is one that depends on a physical reaction, such as overloading a container with compressed air. Such a device has some application in mining, where the release of gas from chemical explosives may be undesirable, but otherwise is very little used. A nuclear explosive is one in

which a sustained nuclear reaction can be made to take place with almost instant rapidity, releasing large amounts of energy. Experimentation has been carried on with nuclear explosives for possible petroleum extraction purposes. Military applications of nuclear explosives are discussed in WAR, THE TECHNOLOGY OF: *Nuclear weapons.* This section is concerned with chemical explosives, which account for virtually all explosive applications in engineering.

TYPES OF CHEMICAL EXPLOSIVES

Basically, chemical explosives are of two types: (1) detonating, or high, explosives and (2) deflagrating, or low, explosives. Detonating explosives, such as TNT and dynamite, are characterized by extremely rapid decomposition and development of high pressure, whereas deflagrating explosives, such as black and smokeless powders, involve merely fast burning and produce relatively low pressures. Under certain conditions, such as the use of large quantities and a high degree of confinement, some normally deflagrating explosives can be caused to detonate.

Detonating explosives are usually subdivided into two categories, primary and secondary. Primary explosives detonate by ignition from some source such as flame, spark, impact, or other means that will produce heat of sufficient magnitude. Secondary explosives require a detonator and, in some cases, a supplementary booster. A few explosives can be both primary and secondary depending on the conditions of use.

BLACK POWDER

It may never be known with certainty who invented the first explosive, black powder, which is a mixture of saltpetre (potassium nitrate), sulfur, and charcoal (carbon).

Chinese fireworks

The consensus is that it originated in China in the 10th century, but that its use there was almost exclusively in fireworks and signals. It is possible that the Chinese also used black powder in bombs for military purposes and there is written record that, in mid-13th century, they put it in bamboo tubes to propel stone projectiles.

There is, however, some evidence that the Arabs invented black powder. By about 1300, certainly, they had developed the first real gun, a bamboo tube reinforced with iron, which used a charge of black powder to fire an arrow.

A strong case can also be made that black powder was discovered by the English medieval scholar Roger Bacon, who wrote explicit instructions for its preparation in 1242, in the strange form of a Latin anagram, difficult to decipher. But Bacon read Arabic, and it is possible that he got his knowledge from Arabic sources.

Some scholars attribute the invention of firearms to an early 14th-century German monk named Berthold Schwarz. In any case they are frequently mentioned in 14th-century manuscripts from many countries, and there is a record of the shipment of guns and powder from Ghent to England in 1314.

Not until the 17th century was black powder used for peaceful purposes. There is a doubtful claim that it was used in mining operations in Germany in 1613, and fairly authentic evidence that it was employed in the mines of Schemnitz, Hungary (modern Banská Štiavnica, Czechoslovakia), in 1627. For various reasons, such as high cost, lack of suitable boring implements, and fear of roof collapse, the use of black powder in mining did not spread rapidly, though it was widely accepted by 1700. The first application in civil engineering was in the Malpas Tunnel of the Canal du Midi in France in 1679.

Composition of black powder

For 300 years the unvarying composition of black powder has been approximately 75 percent saltpetre (potassium nitrate), 15 percent charcoal, and 10 percent sulfur. The saltpetre was originally extracted from compost piles and animal wastes. Deposits found in India provided a source for many years. During the 1850s tremendous quantities of sodium nitrate were discovered in Chile, and saltpetre was formed by reaction with potassium chloride, of which there was a plentiful supply.

Chilean nitrate was not at first considered satisfactory for the manufacture of black powder because it too readily absorbed moisture. Lammot du Pont, an American industrialist, solved this problem and started making sodium nitrate powder in 1858. It became popular in a short time because, although it did not produce as high a quality explosive as potassium nitrate, it was suitable for most mining and construction applications and was much less expensive. To distinguish between them, the potassium nitrate and sodium nitrate versions came to be known as A and B blasting powder respectively. The A powder continued in use for special purposes that required its higher quality, principally for firearms, military devices, and safety fuses.

Manufacture of black powder. Manufacture of black powder was accomplished originally by hand methods. Ingredients were ground together with a mortar and pestle. The next step was to use crushing devices of wood (wooden stamps), also operated by hand, in wooden or stone bowls. The stamping process was gradually mechanized and, about 1435, the first powder mill driven by water power was erected near Nuremberg, Germany.

Metallic crushing devices, introduced in the early 1800s, slowly and steadily replaced the wooden stamp mills.

In the modern process, charcoal and sulfur are placed in a hollow drum along with heavy steel balls. As the drum rotates, the steel balls pulverize the contents; this device is called a ball mill. The saltpetre is crushed separately by heavy steel rollers. Next, a mixture of several hundred pounds of saltpetre, charcoal, and sulfur is placed in a heavy iron device shaped like a cooking pan. There it is continuously turned over by devices called plows, then ground and mixed by two rotating iron wheels, which weigh from 10 to 12 tons each. The process takes several hours; water is added periodically to keep the mixture moist.

The product of the mills is next put through wooden rolls to break up the larger lumps and is then formed into cakes under high pressure—namely, from about 210 to 280 kilograms per square centimetre (3,000 to 4,000 pounds per square inch) of pressure. Coarse-toothed rolls crack the cakes into manageable pieces and the corning mill, which contains rolls of several different dimensions, reduces them to the sizes desired.

Glazing (the next operation) consists of tumbling the grains for several hours in large wooden cylinders, during which friction rounds off the corners, and, aided by forced air circulation, brings the powder to a specified moisture content. The term glazing derives from the fact that graphite is added during this process, forming a thin film over the individual powder grains. Glazed powder flows more readily than unglazed powder and is more moisture resistant.

After glazing the powder is graded by sieves into different sizes and packaged, usually in kegs.

Because the burning of black powder is a surface phenomenon, a fine granulation burns faster than a coarse one. Grain sizes are designated as F, 2F, etc., up to 7F, which is the finest, and from C up as the grains become larger. For the A powder the letter indicating the fineness becomes 3FA, etc., and if the powder is glazed, this is followed by the letter *g; e.g.,* 3FAg. For many years the B blasting material was offered in pellet as well as granular form. Four pellets, each five centimetres (two inches) in length and from 2.75 to 6.25 centimetres (1.1 to 2.5 inches) in diameter, were packed in waxed paper cartridges. Each pellet had a hole through its centre to accommodate a safety fuse or an electric device used to ignite the powder. Pelleted powder was used almost entirely in underground coal mines, but now regulations generally prohibit both it and the granular type.

Ignition of black powder. Black powder is relatively insensitive to shock and friction and must be ignited by flame or heat. In the early days such devices as torches, glowing tinder, and heated iron rods were used to ignite the powder and, in most cases, a train of the powder was led to the main charge in order to give the firer time to get to a safe place.

In cannons a small touchhole was drilled into the breech and filled with fine powder. Ignition was usually by means of a slow-burning punk. The same principle was employed in flintlock muskets and rifles except that ignition resulted from sparks produced by contact between flint and steel.

Percussion methods of firing guns have long been in universal use. In the most common procedure, pulling the trigger releases a hammer, which strikes an impact-sensitive explosive mixture. This explosion then ignites the black powder or other powder charge.

Some black powder is still used as the propellant in guns in spite of the superiority of smokeless powder. Besides antique gun experts, who employ it mostly with hand-loaded shells and cartridges, hunters in South and Central America still use guns that require black powder.

In mining, a succession of crude means for ignition (fuses) included straws filled with pulverized black powder, reeds in which the pith was scooped out and replaced with a paste of powder and water (later bound with string and dried), or powder paste spread on wool threads. All of these fuses were ignited either by a piece of wool yarn impregnated with sulfur, called a sulfur mannikin, or some equivalent slow-burning device. A later, and extremely popular, type of fuse was formed of goose quills. The quills were cut so that they could be inserted one into the other and then filled with powder. Quill fuses could be ignited directly, that is, without any delaying element such as the sulfur mannikin. Unfortunately, their reliability was not high, and they often burned erratically.

Invention of safety fuse by William Bickford

Safety fuse. A major contributor to progress in the use of explosives was William Bickford, a leather merchant who lived in the tin-mining district of Cornwall, England. Familiar with the frequency of accidents in the mines and the fact that many of them were caused by deficiencies inherent in the quill fuse, Bickford sought to devise an improvement. In 1831 he conceived the safety fuse: a core of black powder tightly wrapped in textiles, one of the most important of which was jute yarn. The present-day version is not very different from the original model. The cord is coated with a waterproofing agent, such as asphalt, and is covered with either textile or plastic.

The safety fuse provided a dependable means for conveying flame to the charge. Its timing (the time required for a given length to burn) was amazingly accurate and consistent, compared to that of its predecessors, and it was much better from the standpoints of resistance to water and abuse.

Underground coal mining was formerly by far the largest consumer of black powder. From a performance standpoint, it is probably the best explosive for that purpose. Its relatively gentle, heaving action gives a high yield of lump and leaves the coal in good position for rapid loading. Before the advent of oil, gas, and electric heating and cooking, coal was produced in tremendous quantities for household use and lump demanded a premium price. But black powder has a dangerous tendency to ignite coal gas (mostly methane) and coal dust, and many mine explosions occurred. Around 1880 several European governments, seeking to develop safer substitutes for black powder, set up testing stations. Similar action was taken in the United States a few years later. The result was a series of special dynamites approved for use in gassy and dusty coal mines when used in the specified manner. Their blasting action was not as good as that of black powder, but they were very much safer. These dynamites are discussed below.

The use of black powder in underground coal mines is no longer allowed in most countries. As a result, black powder production has decreased tremendously. Further, black powder is now more expensive than dynamite and is used only for special purposes. There is, for example, no substitute for black powder in certain military applications and nothing equal to it has yet been found for the manufacture of safety fuse. The fact that black powder is relatively nonshattering is of value in blasting certain types of stone.

NITROGLYCERIN

Nitroglycerin, another chemical explosive, was discovered by an Italian chemist, Ascanio Sobrero, in 1846. Though he first called it pyroglycerin, it soon came to be known generally as nitroglycerin, or blasting oil. Because of the risks inherent in its manufacture and the lack of dependable means for its detonation, nitroglycerin was largely a laboratory curiosity until Immanuel Nobel and his son Alfred made extensive studies of its commercial potential in the years 1859–61. In 1862 they built a crude plant at Heleneborg, Sweden; Alfred, a chemist, was basically responsible for the design of this factory that was efficient and relatively safe considering the state of knowledge of the times. Nevertheless, it exploded in 1864 and killed, among others, Alfred's youngest brother Emil Oskar. Though deeply affected by the accident, Alfred continued work, at first on a barge that he moored in the middle of a lake. In 1865 he erected a plant at Krümmel, Germany, and another in Sweden at Vinterviken near Stockholm. A third plant was built a year later in Norway. Nobel was granted a patent for the manufacture and use of nitroglycerin in the United States, in 1866, and since importation on a large scale was impractical, he visited the United States in an effort to interest local capital. The victim of a number of unscrupulous businessmen, he finally sold his American holdings in 1885 for only $20,000.

Even today most experts regard Nobel's invention of the blasting cap, a device for detonating explosives, in 1865, as the greatest advance in the science of explosives since the discovery of black powder. Combined with Bickford's safety fuse, the blasting cap provided a dependable means for detonating nitroglycerin and the many other high explosives that followed it. After a number of attempts that were only partially successful, Nobel settled on a charge of mercury fulminate, which had been known for many years, in a copper capsule. With one or two minor changes, this blasting cap remained in general use until the 1920s.

The blasting cap

DYNAMITE

The second most important of Nobel's inventions was dynamite, in 1867. He coined the name from the Greek *dynamis,* "power." The basis for the invention was his discovery that kieselguhr, a porous siliceous earth, would absorb large quantities of nitroglycerin, giving a product that was much safer to handle and easier to use than nitroglycerin alone.

Dynamite No. 1, as Nobel called it, was 75 percent nitroglycerin and 25 percent guhr. Shortly after its invention, Nobel realized that guhr, an inert substance, not only contributed nothing to the power of the explosive but actually detracted from it because it absorbed heat that otherwise would have improved the blasting action. He turned, therefore, to active ingredients such as wood pulp for an absorbent, and sodium nitrate for an oxidizing agent. By varying the ratio of nitroglycerin to these "dopes," as they came to be called, Nobel not only improved the efficiency of dynamite but also was able to prepare it in varying strengths, termed straight dynamites. Thus 40 percent straight dynamite contained 40 percent nitroglycerin and 60 percent dope.

Nobel patented the use of active ingredients in dynamite in 1869. Several others obtained similar patents at about the same time, however, and the result was that no one could establish a clear-cut claim to the invention.

Nobel's next outstanding contribution was his invention of gelatinous dynamites in 1875. There is a legend that he hurt a finger and used collodion, a solution of relatively low nitrogen content nitrocellulose in a mixture of ether and alcohol, to cover the wound. Later, unable to sleep because of the pain, Nobel went to the laboratory to find out what effect collodion would have on nitroglycerin. To his great satisfaction, he found that after evaporation of the solvents, there remained a tough, plastic material. He discovered that he could duplicate this by the direct addition of 7 to 8 percent of collodion-type nitrocotton to nitroglycerin and that lesser quantities of nitrocotton decreased the viscosity and enabled him to add other active ingredients. He called the original material *blasting gelatin* and the dope mixtures *gelatin dynamites.* The principal advantages of these products were their high water resistance and greater blasting action power than the comparable dynamites. This added power resulted from a combination of higher density and a degree of plasticity that allowed complete filling of the borehole (the hole that was bored in the coal seam or elsewhere for implantation of the explosive).

Invention of gelatinous dynamites

The first large-scale manufacture of nitroglycerin in the United States is attributed to George Mowbray, a chemist of considerable ability who had followed the work of Sobrero and others in Europe with great interest. Mowbray published an advertisement offering to supply nitroglycerin. This led to an invitation to manufacture it for completion of the Hoosac Tunnel at North Adams, Massachusetts. Mowbray's plant was built near North Adams in the latter part of 1867. Most of its product went to the tunnel, but a substantial amount was shipped, frozen, throughout the eastern United States and Canada. Pure nitroglycerin, relatively insensitive in frozen form, freezes at about 11° C (52° F) and is, therefore, easy to keep frozen by packing it in ice. Before closing his plant down because of patent difficulties, Mowbray made about 450,000 kilograms (1,000,000 pounds) of nitroglycerin without accidents in either manufacture or shipment.

One of the earliest major uses of nitroglycerin in the United States was in blasting oil wells to increase the flow of oil. E.A.L. Roberts in that country obtained a patent covering this procedure and later acquired the right to manufacture and use nitroglycerin under the Nobel patents. Theoretically, this gave him a monopoly on shooting oil wells, and his company dominated the field, but many of his competitors ignored his patent rights.

After 1883 the use of nitroglycerin was, with a few unimportant exceptions, restricted to oil-well shooting. In recent years more efficient means have been developed for increasing oil flow. Nitroglycerin is still used occasionally because it is more economical in small wells.

Major tunnels blasted by explosives

Three tunnels stand out as benchmarks in the history of the use of explosives: first is Mont Cenis, a 13-kilometre (eight-mile) railway tunnel driven through the Alps between France and Italy in 1857–71, much the largest construction job with black powder up to that time; second was the 6.4 kilometre (four-mile) Hoosac, also a railway project, during the construction of which (1855–66) nitroglycerin first replaced black powder in large-scale construction; third was the Sutro mine development tunnel in Nevada (1864–74) where the switch from nitroglycerin to dynamite for this type of work started.

Ammonium nitrate. After the straight dynamites and gelatins, the next important advance in dynamite was the substitution of ammonium nitrate for part of the nitroglycerin to give a safer and less expensive product. The use of ammonium nitrate in explosives had been patented by others in Sweden in 1867, but it was Nobel who made the new "extra dynamites" successful by devising gelatins that contained from 20 to 60 percent ammonium nitrate.

During the period 1867–84, many people worked to develop nongelatinous ammonium nitrate mixtures, but nothing of value resulted, largely because ammonium nitrate is too hygroscopic; that is, it picks up moisture too readily. In 1885 R.S. Penniman, an American, found a solution to the problem by coating the ammonium nitrate with a small percentage of paraffin, or some similar substance, prior to use. With this development, a series of ammonia dynamites soon became popular. Coating was discontinued when other, safer means were developed to handle the moisture problem.

Explosives used in underground mines

All major underground-coal-mining countries have similar explosives and regulations. In the United States explosives that have been approved by the U.S. Bureau of Mines for use in underground coal mines are called permissibles. Besides passing the Bureau's safety tests, these explosives must be used in a manner specified by the Bureau. In England the explosives are known as permitted; in France, *explosifs antigrisouteux;* in Belgium, *explosifs S. G. P.* (*sécurité, grisou, poussière*), and in Germany, *schlagwettersichere Sprengstoffe.* Almost without exception, the major ingredient in these explosives is ammonium nitrate, chosen because of its low explosion temperature, and nearly all of them contain a cooling agent such as sodium chloride (common salt) or ammonium chloride to prevent the heat of their explosion in a mine from igniting underground gases such as methane, or a combination of them and coal dust, and causing a fire or disastrous secondary explosion. The sensitizer is usually a small amount of nitroglycerin but in some cases it is TNT, trinitrotoluene (discussed later); for example, it is said that a typical Russian permissible would be 68 percent ammonium nitrate, 10 TNT, 20 sodium chloride, and 2 powdered bark.

Synthetic ammonia and its products

As synthetic ammonia became less expensive, because of improvements in manufacture and a raw material change from coal to natural gas, the explosives industry concentrated its efforts on substituting ammonium nitrate for nitroglycerin. Two important products were (1) low-density ammonia dynamites and (2) semigelatins. Prior to their development, the density of most dynamites was about the same and was quite high. Strength was changed in the different grades by varying the amount of explosives used. The new concept was to employ the strongest formula possible, with a minimum of nitroglycerin and a maximum of ammonium nitrate, and to dilute it systematically with suitable low-density ingredients such as bagasse (the pulp remaining after extraction of sugar from the cane) so that one stick of the new product would give the same blasting action as one of the old. This provided a substantial saving to the user because the cost per stick of the new product was much lower.

The only difference between the low-density ammonia dynamites and the semigelatins is that the latter are partially gelatinized through the use of nitrocellulose and a higher nitroglycerin content. This gelatinization provides good water resistance and a degree of plasticity that is desirable in loading holes prior to blasting.

Means are available to obtain a moderate amount of water resistance in the ammonia dynamites without resorting to gelatinization of the nitroglycerin. The most common involve the use of water-repellents such as calcium stearate, and ingredients that form a water gel on the surface of the dynamite that slows down the further penetration of water. Examples of the latter are pregelatinized starch products and rye flour.

Low-freezing dynamite. Attempts to reduce the freezing point of nitroglycerin began shortly after the Nobels introduced it commercially. Frozen dynamite is very insensitive, sometimes so much so that it will not give dependable performance, and it is difficult to use, since it cannot be punched for the insertion of a blasting cap or slit and tamped into a borehole. Consequently, almost all of it had to be thawed for use and careless thawing methods caused many accidents. Not until 1907 was a reasonably successful procedure for producing low-freezing dynamite developed. This involved adding 20 to 25 percent of the liquid isomers (molecules with identical formulas but different structure) of TNT to the nitroglycerin. This was replaced for a short time by a nitrated solution of sugar in glycerin. In 1911 a practical way to manufacture diglycerin (a glycerin polymer) was discovered. Its nitration product, tetranitrodiglycerin, when mixed with nitroglycerin, reduced its freezing point materially.

The ultimate solution to the freezing problem was found in 1925, when synthetic ethylene glycol became available. The explosive properties of ethylene glycol dinitrate are practically identical with those of nitroglycerin and its low-freezing qualities are extremely good. Dynamite containing a mixture of it and nitroglycerin was stored in the open at Point Barrow, Alaska, for four years without freezing.

OTHER EXPLOSIVES

Chlorates and perchlorates. Interest in the chlorates and perchlorates (salts of chloric or perchloric acid) as a base for explosives dates back to 1788. They were mixed with various solid and liquid fuels. Many plants were built in Europe and the United States for the manufacture of this type of explosive, mostly using potassium chlorate, but so far as can be determined, all of them either blew up or burned up, and no chlorate explosives have been manufactured for many years.

Sprengel explosives. In England in 1871, Hermann Sprengel patented combinations of oxidizing agents such as chlorates, nitrates, and nitric acid with combustible substances such as nitronaphthalene, benzene, and nitrobenzene. These differed from previous explosives in that one of the ingredients was liquid and the mixture was made just prior to use. Sprengel explosives were quite popular in Europe but consumption in the United States was rel-

atively small except for the spectacular Hell Gate blast in New York harbour in 1885, in which a combination of 34,000 kilograms (75,000 pounds) of No. 1 dynamite and 110,000 kilograms (240,000 pounds) of potassium chlorate–nitrobenzene were used to remove "Flood Rock," a menace to navigation. Cloth bags of the chlorate were soaked in the nitrobenzene and loaded directly from the soaking tank into the boreholes.

Liquid oxygen explosives. In 1895 the German Carl von Linde introduced carbon black packed in porous bags and dipped in liquid oxygen. This, which was a Sprengel-type explosive, came to be known as LOX. Because of the shortage of nitrates, LOX was widely used in Germany during World War I. Little if any was used in World War II, however, because ample supplies of nitrates could be obtained from synthetic ammonia.

Because the manufacture of liquid oxygen requires complicated and expensive equipment, the use of LOX was limited to areas that could consume very large quantities. In the United States several of the tremendous strip coal mines in the Midwest met this requirement. Maximum consumption of LOX explosive was about 10,190,000 kilograms (22,465,000 pounds) in 1953, but it fell to zero in 1968. Cheap as LOX is, it cannot compete with ammonium nitrate–fuel oil mixtures.

Nitrostarch explosives. Nitrostarch, which is closely related to nitrocellulose, attracted early attention, but it was not until about 1905 that it proved possible to produce it in a stable form. In general nitrostarch explosives are similar to the straight and ammonia dynamites except that nitrostarch is used in place of nitroglycerin. Disadvantages are its relatively low strength, mediocre water resistance, and the fact that it cannot be transformed into gelatinous products. Nitrostarch explosives, however, do not produce the headaches from skin contact that are characteristic of mixtures containing nitroglycerin. For that reason they are still marketed.

Nitramon and Nitramex explosives. An important advance in explosives technology was the development by du Pont in 1934 of Nitramon, a canned product with a typical formula of 92 percent ammonium nitrate, 4 percent dinitrotoluene, and 4 percent paraffin wax. Some grades contain metallic ingredients such as aluminum and ferrosilicon. Nitramon is insensitive to the action of a line of detonating cord, a commercial blasting cap, shock and friction, or the impact of small-calibre ammunition. A large primer is required for its detonation and the one normally used is known as a Nitramon primer. This is also a canned product with Nitramon at each end but a centre section of amatol that can be detonated by either detonating cord or a blasting cap. The cans are provided in varying sizes. A minimum diameter of 10 centimetres (four inches) for regular Nitramon is necessary to insure proper explosive effect if individual cans in a column become separated by some material such as a rock. Special grades are made for use in seismic exploration for gas and oil in 5- and 6.4-centimetre (2- and 2½-inch) diameters. In this case, however, the cans are threaded and intimate contact is assured because the column is screwed together.

Nitramex is similar to Nitramon but is much stronger because it contains TNT and a metallic ingredient such as aluminum. Both it and Nitramon have been largely replaced by the water gels, which are described later.

Removal of a mountaintop

So far as is known, the largest commercial, non-nuclear blast in North America was made on April 5, 1958, in Seymour Narrows, which lies between Vancouver Island and the mainland of British Columbia. The object of the blast was to remove the top of a submerged twin-peak mountain, known as Ripple Rock, which was only 2.7 metres (9 feet) below the surface at low tide. More than 120 vessels had been lost because of this obstacle. In preparing for the blast, a shaft was sunk on shore to the proper depth. From it a tunnel was driven to a point directly under the twin peaks, from which a vertical shaft finally was driven to the desired depth below the peaks. A series of small horizontal drifts and pockets was prepared for placement of the explosives, consisting of 1,253,000 kilograms (2,756,000 pounds) of Nitramex 2H and a special primer, fired by means of detonating cord.

After the blast the top of the rock was a minimum of 15 metres (50 feet) below the surface and no longer a menace to navigation.

MODERN HIGH EXPLOSIVES

The year 1955, marking the beginning of the most revolutionary change in the explosives industry since the invention of dynamite, saw the development of ammonium nitrate–fuel oil mixtures (ANFO) and ammonium nitrate-base water gels, which together now account for at least 70 percent of the high explosives consumption in the United States. The technology of these products is far more advanced in the U.S. than it is in other countries; so, at the present time, they have not replaced nearly as much of the older explosives in the rest of the world. In addition to a variety of packages, both ANFO and water gels are delivered in bulk by special trucks and loaded directly into boreholes.

Ammonium nitrate–fuel oil mixtures. In 1955 it was discovered that mixtures of ammonium nitrate and fine coal dust would give very satisfactory blasting results in the large (about 22.5-centimetre, 9-inch) holes used in open-pit coal mines to remove the rock and soil covering the coal. Polyethylene bags for this material both stretched to fill the holes and provided a moderate amount of water resistance.

Shortly thereafter, ANFO was evaluated in the open-pit iron mines of Canada and the United States, with a high degree of success. From there ANFO spread to other open pits, such as copper, and to construction work such as road building. It was then found that the mixture could be air blown into holes 5 centimetres in diameter, or even smaller, with excellent results. This led to its adoption in many underground mines.

Prills

ANFO applications were based on prilled rather than crystallized ammonium nitrate. Prills, or free-flowing pellets, were developed for the fertilizer market, which requires a coarse product that has little tendency to set and can be spread easily and smoothly. A small amount of kieselguhr is generally added to improve the flowing properties. Prills are made by allowing droplets of ammonium nitrate that is almost molten to fall freely from a high tower. When they reach the bottom, they are dry and solidified, and slightly porous, which allows them to absorb and hold a greater amount of oil and gives a more sensitive product. ANFO is almost universally prepared by mixing 94 percent of prills with 6 percent of No. 2 fuel oil. The latter imparts some water resistance and, if that is not enough, polyethylene bags can often be used to give the necessary protection.

Water gels. Water gels, or slurries, were introduced in 1958. These were, at first, mixtures of ammonium nitrate, TNT, water, and gelatinizing agents, usually guar gum and a cross-linking agent such as borax. (Cross-linking is a form of chemical bonding.) Later, aluminum and other metallic fuels were sometimes used and vastly better gelatinizers were discovered. In addition nonexplosive sensitizers were developed that could replace the TNT if desired. When the highest possible concentration of strength is needed, however, large quantities of TNT are still used.

Water gels have many advantages. Among them are a high concentration of strength, a high degree of water resistance, plasticity that permits them to displace air or water and completely fill the borehole, economy, ease of handling and loading, and good safety characteristics.

NITROCELLULOSIC EXPLOSIVES

When Christian Schoenbein invented nitrocotton (guncotton) in 1845 by dipping cotton in a mixture of nitric and sulfuric acids and then removing the acids by washing with water, he hoped to obtain a propellant for military weapons. It proved, however, to be too fast and violent. About 1860 Maj. E. Schultze of the Prussian army produced a useful nitrocellulosic propellant. He nitrated small pieces of wood by placing them in nitric acid and then, after removing the acid, impregnated the pieces with barium and potassium nitrates. The purpose of the latter was to provide oxygen to burn the incompletely nitrated wood. Schultze's powder was highly successful in shotguns but was too fast for cannon or even most rifles.

In 1884 a French chemist, Paul Vieille, made the first smokeless powder as it is now known. He partially dissolved nitrocellulose in a mixture of ether and alcohol until it became a gelatinous mass, which he rolled into sheets and then cut into flakes. When the solvent evaporated, it left a hard, dense material resembling horn. This product gave satisfactory results in all types of guns.

In 1887 Nobel introduced another of his revolutionary inventions, which he called Ballistite. He mixed 40 percent of a lower nitrogen content, more soluble nitrocellulose and 60 percent of nitroglycerin. Cut into flakes, this made an excellent propellant and it continued in use for over 75 years. The British refused to recognize Nobel's patent and developed a number of similar products under the generic name cordite.

The progress of smokeless powder in the United States was much slower than it was in Europe. Long-continued work, principally by E.I. du Pont de Nemours & Company, finally resulted in a material that was excellent for guns of all types and sizes. It was first marketed about 1909 and was the most important type of smokeless powder used by the Allies in World War I. It was made from a nitrocotton of relatively low nitrogen content, called pyrocellulose, because that type is quite soluble in ether–alcohol. A small amount of diphenylamine was used as a stabilizer and, after forming the grains and removing the liquid, a coating of graphite was added. The smokeless powder most widely used in the United States at the present time is much the same. Other popular types are mostly double-base and may contain from about 20 to 35 percent nitroglycerin. Cotton linters for nitration have been almost, if not entirely, replaced by purified wood cellulose.

BLASTING CAPS

Nobel's original fuse-type blasting cap remained virtually unchanged for many years, except for the substitution of 90–10 and 80–20 mixtures of mercury fulminate and potassium chlorate for the pure fulminate. This did not affect the performance materially and provided a substantial economy. Mercury fulminate is an example of an explosive that can be both primary and secondary. In its more compressed form it is a high density base charge; less compressed, a low density primer charge. Hexanitromannitol (nitromannite) functions in the same manner and is used that way in a very successful blasting cap.

Mercury fulminate

Extensive work was carried out on replacements for the costly mercury fulminate; by 1930 very little of it remained in use, and by the 1970s it had disappeared from commercial use. Experience has shown that the cheaper replacements are actually superior.

The dominant base-charge materials are now pentaerythritol tetranitrate (PETN) and cyclotrimethylenetrinitramine (RDX). These are as strong as nitroglycerin, quite safe to manufacture and handle, and relatively inexpensive. In addition to low density nitromannite, diazodinitrophenol, lead styphnate, and lead azide are widely used as ignition-primer charges. One other departure from Nobel's blasting cap is the fact that aluminum has now almost entirely replaced copper as the material used for the shell.

Electrical firing. The principal advantages of electric over fuse firing are exact control of the time when the blast is initiated, the simultaneous firing of a number of shots, if that is desired, and the ability to obtain a very high degree of water resistance. Attempts to make electric blasting caps date back to the 1700s, but nothing of a really practical nature was developed until late in the 19th century. There were two separate problems, the cap and the means to fire it.

Blasting machines. The first satisfactory electrical blasting machine was invented by H. Julius Smith, an American, in 1878. It comprised a gear-type arrangement of rack bar and pinion that operated an armature to generate electricity. When the rack bar was pushed down rapidly, it revolved the pinion and armature with sufficient speed to obtain the desired current. This current was released into the external, or cap, circuit when the rack bar struck a brass spring in the bottom of the machine. Smith's blasting machine was improved and made in a range of capacities; also, a small twist-type machine that employed basically the same principles was introduced. These machines are still in widespread use, although they have been replaced to a considerable extent by power firing, and capacitor-discharge blasting machines. The latter have a battery power source for energizing one or more capacitors and a safe, dependable means for discharging the stored energy. They have high capacity for their weight and size and are rapidly displacing the other firing systems.

Ignition systems. Except for the means of firing, there is little difference between electric and fuse-type blasting caps. With minor variations, the explosives used are the same.

It was in the 1880s that the forerunner of the modern electric blasting cap was first assembled. In contrast to the spark-type ignitions previously used, it employed a fine, high-resistance wire soldered between two insulated leg wires and embedded in, or coated with, an ignition mixture. The resistance wire was either platinum or one of its alloys, and the ignition mixture was based on mercury fulminate. The leg wires were insulated with two layers of cotton thread, wound in opposite directions. Except for coal-mine caps, the wire was then run through a bath of molten asphalt. Paraffin wax was used for the coal-mine caps because its white colour provided good contrast with the black coal. Sulfur, or a mixture of sulfur and mica or graphite, was used to hold the leg wires in place and seal the cap. Sulfur was well suited for this purpose because its melting point is very low and it is compatible with the explosive ingredients. Later, to obtain better water resistance, part of the sulfur was replaced by asphalt.

In 1939 the du Pont company introduced a revolutionary new type of ignition system. Nylon plastic was substituted for the cotton insulation, a rubber plug to hold the leg wires replaced the sulfur plug, and the bridge wire was welded to the leg wires instead of soldered. By that time alloys such as nichrome had largely replaced the platinum bridge wires. The shell was crimped tightly to the rubber plug, with the result that the cap could withstand a substantial amount of water pressure. All electric blasting caps are now made substantially in this way. Polyvinyl chloride is widely used for the leg wire insulation and plastic is sometimes substituted for rubber in the plug.

Introduction of nylon insulation

Match-head ignition, very popular in Europe, is used less widely in the United States. The ignition device consists of a piece of cardboard with a thin sheet of metal glued to each side. A bridge wire is soldered to these sheets, around the end of the cardboard, and this part of the assembly is dipped in a slurry of ignition mixture, usually based on copper acetylide. After drying, the match head is given a protective coating and is then soldered to the leg wires.

Most countries require explosives in underground coal mines to be fired electrically but prohibit the use of aluminum-shell electric blasting caps. This is because aluminum burns with a very hot flame and is much more likely than copper to ignite coal gas. Otherwise, almost all electric blasting-cap shells are made of aluminum.

Delay systems. Delay, or rotational, shooting has many advantages over instantaneous firing in almost all types of blasting. It generally gives better fragmentation, more efficient use of the explosive, reduced vibration and concussion, and better control of the rock. For these, and sometimes other reasons, most blasting operations are now conducted with a delay system.

It is probable that the first use of delay firing was in tunnels. The centre was shot out first and then successive rings around it until the desired tunnel dimensions were reached. The procedure was to cut all of the fuses to the same length and then trim them toward the centre; for example, the outside ring of fuses would be full length, the next ring a few centimetres shorter, and so on. In addition, the fuses were lit from the centre out, causing a little more delay in the desired direction. This method of shooting could not be used until Bickford's safety fuse, which had a very uniform burning speed, became available.

Delay electric blasting caps are the most commonly used means for obtaining rotational firing. They are of two types: (1) the so-called regular delay, which has been

Types of delay caps

in use since the early 1900s; and (2) the short-interval, or millisecond, delay, which was introduced about 1943. Except for a delay element placed between the ignition and primer charges, they are the same as instantaneous electric caps.

A typical series of regular delays would comprise 14 periods ranging from a few milliseconds to about 12 seconds. To avoid overlapping and because there is some variation in the burning speed of the delay element, the intervals are made longer in the higher periods; for example, the delay between periods 1 and 2 might be 0.8 seconds, whereas for 13 and 14 it might be 1.5 seconds. Ordinary delays have been largely replaced by short-interval delays but are still used to a considerable extent for such purposes as driving tunnels and sinking shafts.

The periods in short-interval delays are usually separated by 25 milliseconds up to 200 milliseconds, by 50 up to 500, and by 100 up to 1,000 (one second). This close spacing gives improved fragmentation, the ability to fire many holes with hardly any more vibration or concussion than would be obtained with one hole, less chance that the detonation of one hole will cut off an adjacent hole, and a reduction in the quantity and cost of explosives. Short-interval delays are used above ground in such work as excavating and quarrying, and for almost all types of underground mining. Their development is one of the major advances in explosives.

Delay elements for electric blasting caps function in about the same way as black powder in safety fuse, except that the chemical mixtures used are much faster. At times the delay mixture is simply pressed on top of the primer mix. Usually, however, it is put in the centre of a metallic tube in lengths that will give the desired delay interval.

DETONATING CORD

Detonating cord (detonating fuse) resembles safety fuse but contains a high explosive instead of black powder. The first successful one, patented in France in 1908, consisted of a lead tube, about the same diameter as safety fuse, filled with a core of TNT. It was made by filling a large tube with molten TNT that was allowed to solidify. The tube was then passed through successively smaller rolls until it reached the specified diameter. In France the product was called *cordeau détonant,* elsewhere shortened to cordeau. Its velocity was about 4,900 metres (16,000 feet) per second.

Cordeau

In 1936 the Ensign-Bickford Company, Simsbury, Connecticut, the American manufacturers of cordeau, developed Primacord, based on French patents and comprising a core of PETN covered with various combinations of textiles, waterproofing materials, and plastics. The velocity is approximately 6,400 metres (21,000 feet) per second. Many types of Primacord are available, for both military and commercial use, but the industrial varieties generally contain from 25 to 60 grains of PETN per 0.3 metres. RDX is sometimes used in place of PETN for high temperatures, because the melting points are, respectively, 203.5° and 140° C (398.3° and 284° F).

Detonating cord has many applications in blasting. Any number of holes can be connected with it in just about any desired pattern. Attached to the blasting charge and knotted to a trunk line, it is fired by means of either a fuse-type or electric blasting cap. Sequential shooting may be obtained by cutting the trunk lines and inserting delay connectors, which have delay periods ranging from about 5 to 25 milliseconds.

MILITARY EXPLOSIVES

Military requirements for high explosives differ in many respects from those for commercial users. Military explosives must have insensitivity to shock and friction and must be unlikely to detonate from small-arms fire and yet have excellent shattering power. They must have the ability to withstand long periods of adverse storage without deterioration and must be able to be fired in projectiles or dropped in aerial time bombs without premature explosion. Some types are required to possess almost unlimited water resistance. Many types must have complex fuzes for detonation.

TNT. Trinitrotoluene (TNT) is the most useful military high explosive. Although it had been known for many years and was used extensively in the dye industry, it was not employed as an explosive until 1904. It is an excellent military explosive in itself, but its most valuable property is that it can be safely melted and cast either alone or as a slurry with other explosives. This is because there is a wide spread between its melting point and its decomposition temperature.

It has two shortcomings: first, it is extremely insensitive in the cast form, and second, it is difficult to cast without air holes. The first problem can be overcome by drilling a hole, about 2.5 centimetres (1 inch) in diameter, the length of the charge in the shell and filling it with trinitrophenylmethylnitramine (tetryl); the second, by using a mixture of 40 percent trinitroxylene (TNX) and 60 percent TNT. This mixture not only casts perfectly but can be detonated with a smaller tetryl booster. There is no indication that any TNX was used in World War II; it is believed to have been replaced by PETN and RDX.

Picric acid and ammonium picrate. Picric acid was used as a shell explosive in Europe during the 1880s and carried through World War I on a very large scale. Quantities of it were made in the United States but the Army and Navy used mainly TNT.

Ammonium picrate (Explosive D) has exceptional value as a charge for armour-piercing projectiles. Loaded in a shell with a suitably insensitive primer, it can be fired through 30 centimetres (12 inches) of armour plate and made to detonate on the far side. These armour-piercing shells were used in both World Wars.

Early in World War I it was found that mixtures of molten TNT and ammonium nitrate were almost as effective for shell loadings as pure TNT. The mixtures most commonly used were 80–20 and 50–50 AN and TNT, known as amatol. Their principal advantages were that they made the supply of TNT go further and were considerably cheaper. In World War II the amatols were used in aerial bombs as well as artillery shells.

Amatol

To conserve TNT in World War I, a nitrostarch-base composition was also developed for loading hand grenades and trench-mortar shells.

Several explosives, although previously known, only came into use during World War II. The most important of these were RDX, PETN, and ethylenediaminedinitrate (EDNA), all of which were cast with varying amounts of TNT, usually 40 to 50 percent, and used where the highest possible shattering power was desired. For example, cast 60–40 RDX-TNT, called cyclotol, develops a detonation pressure of about 270,000 atmospheres (4,000,000 pounds per square inch). Corresponding mixtures of PETN and TNT have almost as much shattering effect. The EDNA mixtures, or ednatol, were used only to a limited extent and for special purposes. Probably the most powerful of all nonatomic military explosives are the cast mixtures containing aluminum. The torpedo warhead torpex, for example, is a cast mixture of RDX, TNT, and aluminum.

A series of plastic demolition explosives with great shattering power, designated Composition C-1 to Composition C-4, has had considerable publicity. These contain about 80 percent RDX combined with a mixture of various oils, waxes, and plasticizers. The only significant difference is in the temperature range through which they remain useful. C-3 stays plastic to −29° C (−20° F) and does not exude oil below 49° C (120° F). In contrast, C-4 remains plastic to −57° C (−70° F) and does not leak below 77° C (170° F).

Plastic explosives

Shaped charges. The shaped charge, principally the hand-fired rocket, is another highly publicized product introduced during World War II. A shaped charge normally consists of a cone made of metal or glass surrounded by a high-strength, high-density explosive and means to obtain the proper standoff, or distance to the target.

When the explosive is detonated, the cone is collapsed and vaporized, forming a small, high-temperature jet containing particles of liner material moving at 3,050 to 9,100 metres (10,000 to 30,000 feet) per second. This strikes the target with such heat and force that the target simply flows radially from the point of impact leaving a deep, nearly

round hole. Spectacular as the results are, only about 15 percent of the explosive energy is focussed.

Only two commercial applications have been found for shaped charges. One is for drawing off molten iron or slag from open hearth and blast furnaces in the steel industry, and the other for perforating casings in the oil industry. In the first case, the principal advantage is safety as compared with oxygen lancing. In the second, shaped charges are almost always more efficient than the bullets they replace.

OTHER INDUSTRIAL APPLICATIONS

Explosive rivets. Blind rivets are needed when space limitations make conventional rivets impractical. One type of these is explosive; it has a hollow space in the shank containing a small charge of heat-sensitive chemicals. When a suitable amount of heat is applied to the head, an explosion takes place and expands the rivet shank tightly into the hole. The shank is normally open but can be sealed to eliminate noise and the ejection of metal fragments. Most explosive rivets are aluminum but they can be obtained in stainless steel and certain other metals. Their use is mainly in aircraft.

Bonding metals

Explosive bonding. Explosives are sometimes used to bond various metals to each other. For example, when silver was removed from United States coinage, much of the so-called sandwich metal that replaced it was obtained by the explosive bonding of large slabs, which were then rolled down to the required thickness. These slabs are placed parallel to each other and approximately 6.4 millimetres (0.25 inch) apart. An explosive developed especially for the purpose is placed on the top slab and its detonation slams the slabs together with such force that they become welded. Stainless steel is often joined to ordinary steel in this manner. One especially valuable feature of explosion cladding is that it can frequently be applied to metallurgically incompatible metals, such as aluminum and steel or titanium and steel.

Finally, the very fine industrial-type diamonds used for grinding and polishing are produced by the carefully controlled action of explosives on carbon. (N.G.J.)

Rubber

Rubber is an organic substance—obtained from natural sources or synthesized artificially—that has the desirable properties of extensibility, stretchability, and toughness. Previously known as caoutchouc, a term that has become limited to the chemically pure form of the substance, rubber acquired its name from its ability to erase pencil marks.

Initially, all rubber was natural. Formed in a living tree, it consisted of the solids separated (coagulated) from the milky fluid latex, which occurs under the bark of many tropical and subtropical trees and shrubs. Today, more than half the rubber on the market, apart from one specialty rubber, is manufactured synthetically by means of chemical processes that were partly known in the 19th century but had not been applied successfully until World War II.

NATURAL RUBBER

History. On his second voyage to the New World in 1493–96, Christopher Columbus saw Indians in Haiti play a game with balls made from the gum of a tree. The first mention of rubber being used for purposes other than sport was made in 1615, when a Spaniard related how the Indians, having gathered the milk from incisions made in various trees, brushed it onto their cloaks and obtained crude footwear and bottles by coating earthen molds and allowing them to dry.

The first serious accounts of rubber production and the primitive native system of manufacture were given in the 18th century when a member of a French geographical expedition sent to South America in 1735 described caoutchouc as the condensed juice of the *Hevea* tree. The first accurate account of the botanical characteristics of the *Hevea* species was given in 1775.

At the outset, not chemical knowledge but the ability to devise a suitable method of rubber manipulation was of paramount importance to the development of the rubber industry. Two Frenchmen, a physician and a chemist, attempted to discover a cheap and effective solvent for crude rubber brought from South America and introduced the use of turpentine and ether as solvents.

Macintosh and Hancock. Important progress came at the beginning of the 19th century from the separate experiments of a Scottish chemist, Charles Macintosh, and an English inventor, Thomas Hancock. Macintosh's contribution was the rediscovery, about 1820, of coal-tar naphtha as a cheap and effective solvent. His experiments were successful; he placed a solution of rubber and naphtha between two fabrics and in so doing avoided the sticky or brittle surfaces that had been common in earlier single-texture garments treated with rubber. Manufacture of these double-textured, waterproof cloaks, henceforth known as "mackintoshes," began soon afterward.

Invention of the "mackintosh"

The work of Hancock, who became Macintosh's colleague and partner, is of even greater importance. He first attempted to dissolve the rubber in turpentine, but his hand-coated fabrics were unsatisfactory in surface texture and smell. He then turned to the production of elastic thread. Strips of rubber were cut from the imported lumps and applied in their crude state to clothing and footwear. In 1820, in an effort to find use for his waste cuttings, Hancock invented a masticator. Constructed of a hollow wooden cylinder equipped with teeth in which a hand-driven, spiked roller was turned, this tiny machine, originally taking a charge of two ounces of rubber, fulfilled Hancock's greatest hopes. Instead of tearing the rubber to shreds, it produced enough friction to weld the scraps of rubber into a coherent mass that could be applied in further manufacture.

Goodyear and vulcanization. Early manufactured rubber softened with heat and hardened with cold (particularly annoying in the U.S., where climate was more extreme than in England). It also was tacky, odorous, and perishable. These fundamental weaknesses were removed by the invention of vulcanization in 1839 by the U.S. inventor Charles Goodyear. Developing a compound of rubber, lead, and sulfur and a heat treatment (or curing) process, Goodyear created a product, at first called fireproof gum, afterward vulcanized rubber, that exhibited impressive durability.

Vulcanization made the modern rubber industry possible by permitting use of the substance in combination with machinery and in tires for bicycles, and later for automobiles. Though subsequent discoveries have refined Goodyear's original techniques, the vulcanization process remains fundamentally the same as it was in his day.

Rise of the rubber industry. Most European countries had established rubber industries by the decade 1820–30. Vulcanization and the increased use of steam and electric power brought a large demand for rubber. A pneumatic wheel was patented as early as 1845. The invention was ahead of its time, and it was left to John B. Dunlop, a British veterinary surgeon, to found the tire industry by patenting and developing pneumatic tires for bicycles and tricycles in 1888. His work had a great influence on popularizing cycling, both as a hobby and as a means of transport. At the end of the century some of the largest tire companies were commencing operations manufacturing solid rubber tires for horse-drawn carriages.

Founding of the tire industry

In 1876 the Englishman Sir Henry Wickham collected rubber seeds from the wild trees of the Amazon jungles and had them planted in Kew Gardens. The young trees were subsequently transferred to Ceylon and the Malay Peninsula where they formed the base of the natural rubber plantation industry that produces nearly 3,000,000 tons a year.

Rubber and the automotive industry. The first decade of the 20th century saw the establishment of the motorcar in Europe and the U.S. The dependence of the automotive industry for its tires, as well as many other components, on the drops of latex exuding from trees halfway across the world seems extraordinary in retrospect. Yet, despite the rapid expansion of the auto industry, especially in the U.S., it remained entirely dependent on natural rubber until about 1940. The change from steel wheels to rub-

ber tires on agricultural tractors in the early 1930s added still more to the demand for rubber in the years before World War II.

When World War II began, Germany and the Soviet Union were the only two countries with the means of making synthetic rubber; all the other combatants were dependent upon natural rubber from southern Asia. After Japan entered the war, Asian sources, except for Ceylon, were cut off. The U.S., faced with necessity, developed a synthetic rubber industry almost overnight, achieving a production of 800,000 tons per year. At the war's end, with natural rubber again available, the U.S. synthetic rubber industry went into a sharp decline. By the early 1950s, however, a superior and more uniform synthetic became available. This new synthetic rubber, exported to the British and other European markets, stimulated development of a synthetic rubber industry in England. Synthetic rubber equalled natural in world volume in the early 1960s and moved ahead thereafter.

Development of a synthetic rubber industry

Since World War II, rubber—natural or synthetic—has retained its place as one of the world's most important commodities. While plastics have made some inroads into markets previously monopolized by rubber, new uses and expansion in traditional applications have brought steady expansion of the industry.

The rubber tree. Commercially, natural rubber is obtained almost exclusively from *Hevea brasiliensis,* a tree indigenous to South America, where it grows wild to a height of 36 metres (120 feet). Cultivated in plantations, however, the tree grows only to about 24 metres (80 feet) because carbon, necessary for growth, is also an essential constituent of latex. Since only atmospheric carbon dioxide can supply carbon to the plant, that element has to be rationed between the two needs when the tree is in active production. And with foliage limited to the top of the tree (to facilitate tapping), the intake of carbon dioxide is less than in a wild tree. Because other latex-producing trees do not compare with *Hevea brasiliensis* for efficiency, industry botanists have concentrated their efforts exclusively upon this species.

Other natural sources of rubber

Because of their historical importance, however, other natural sources of rubber of considerable value in pre-synthetic days deserve brief mention. *Castilla,* a plant of the family Moraceae of the order Urticales (elm), yielded a large quantity of rubber at one tapping, then required a rest for several months. *Manihot,* a genus of the family Euphorbiaceae, though best known as the source of tapioca, provides another variety of latex. Outside Brazil, producers have constantly searched for alternate sources of rubber. the watery content of many plants contains some rubber; two of importance are guayule (*Parthenium argentatum*), a shrub grown mainly on the U.S.–Mexican border, its chief drawback being its high gum content (up to 50 percent); and the Russian dandelion (*kok-saghyz*), yielding a latex with a 30 percent dry-rubber content. Both were exploited during World War II.

Modern production. *Hevea brasiliensis* has long been cultivated. Seeds are carefully selected, often from widely scattered areas to avoid inbreeding. Widespread planting provides a greater selection of young plants with elimination of weak growths. In the scientific layout of plantations, the natural contours of the land are followed and the trees protected from wind. Precautions against disease include inoculation. Cover crops planted adjacent to the rubber trees hold rainwater on sloping ground and help fertilize the soil by fixing atmospheric nitrogen. Standard horticultural techniques, such as nursery growing of hardy rootstocks and grafting on top of them, hand pollination, and vegetative propagation (cloning) to produce a genetically uniform product, are also employed.

Plantations. Rubber trees grow only within a well-defined area of the tropics—about 10 degrees north or south of the Equator—and in most types of soil. Heavy annual rainfall of about 2,500 millimetres (100 inches) is essential, with emphasis on a wet spring. In consequence, growing areas are limited. Southeast Asia is particularly well situated for rubber culture; so are Liberia and Nigeria in West Africa. Malaysia is the biggest producing area and is considered the centre of the world industry because of the concentration of commercial and technical interests in the capital city of Kuala Lumpur.

Malaysian rubber production

The plantations of Malaysia achieved prominence about 1900; from that time their products steadily replaced the wild rubber of Brazil. Yield initially was about 335 kilograms per hectare per year. By effective management, research, and the use of chemical stimulants, the average has been increased more than five times. About 28 grams (one ounce) of rubber is obtained per tree per tapping before the latex coagulates and flow stops. If this coagulation can be prevented, the latex continues to flow; a chemical, Ethrel, applied to the bark has been shown to be capable of doing this. In consequence, the small collection cup that was emptied every day or two is being replaced by bags that require attention only every 10 to 14 days.

Tapping. The standard method of tapping a rubber tree for latex is to score it with a knife to a depth of one millimetre for half the circumference of the trunk, slanting the cut down from the left to right at an angle of 30 degrees starting at the highest point convenient to the tapper; each subsequent cut is made immediately below its predecessor (Figure 7). Trees are often rested for a period after heavy tapping. Production commences when a tree is from five to seven years old; with care, the tree's life may extend to more than 30 years.

By courtesy of UniRoyal, Inc., New York

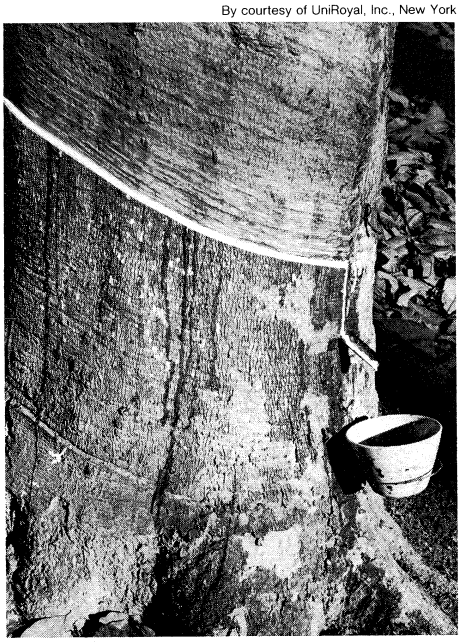


Figure 7: Trunk of a rubber tree showing latex flowing from tapping cut into the collection cup on a plantation in Malaysia.

The watery latex from the scored tree oozes out from between the outside cork layer of the bark and the cambium inner layer, where new cells are formed. Although the function of latex in maintenance of the trees has not been explained satisfactorily, a great deal is known of its composition. Three types of particles are suspended in the watery medium: one rubber and two nonrubber. Of the two latter, the more abundant is the neutral gray particle known as a lutoid, and the other, more highly coloured and named after its discoverer, is called the "Frey Wyssling" particle.

Coagulation. After collection, the latex is sieved to remove foreign objects; then water is added. Next, the rubber is coagulated from the latex by the addition of dilute acetic or formic acid and is deposited on aluminum partitions in horizontal tanks having aluminum partitions that allow the rubber to be removed in slabs rather than in a single lump. When necessary, the slabs are washed by passing them between horizontal rollers, moving at different speeds and flushed with a copious supply of water. Excess water is removed by passing the rubber through a succession of rollers. Subsequent treatment depends upon the type and grade of rubber desired.

Treating slabs of rubber

Wild rubber was originally recovered by heating latex over a fire while stirring it with a stick. The water was removed by evaporation, leaving behind a solid ball of rubber. This method was not welcomed by the market, however, because the solid balls were hard to store and, in the early days, because of the fear that stones and metals would be included to falsify weight.

Rubber types and grades. Rubber may be processed into pale crepe, smoked sheet, or skim rubber. The various ways of grading rubber are subject to controversy.

Pale crepe. After emerging from the coagulation tanks, the wet slabs of rubber are run through a succession of creping mills that roughen the rubber's surface and gradually reduce its thickness to facilitate drying. The sheets are then laminated to the desired thickness and allowed to air-dry. Pale crepe, which varies in colour from off-white to deep cream, is used wherever colour is an important factor. Impurities in the rubber give it its colour, which can be controlled by regulating the strength of the acid when the first rubber is deposited by coagulation; the stronger the acid, the more impurities will be precipitated. The lighter coloured rubber can be sold at a premium. A bleaching process is also used. In general, pale crepe is used only where colour is decisively important, as some of its physical properties (tensile strength, for example) are inferior to those of sheet.

Smoked sheet. Slab rubber can sometimes be sheeted out—that is, run through sets of plain rolls until the last set, which prints a simple pattern, such as crisscross. The sheet is then hung up to dry in smoke from burning wood or oil to obtain the "smoked sheet" of commerce. The best quality has an amber or brown shade.

Skim rubber. Normal coagulation methods leave 10 percent of the rubber behind in the liquid of the coagulation tank. That residual rubber can be recovered by skimming, but it contains a much higher proportion of impurities and so is less desirable.

Pale crepe, smoked sheet, and skim rubbers are generally made up into bales, most frequently weighing 102 or 113 kilograms (225 or 250 pounds) and bulking as a cube about 60 centimetres (two feet) to a side. The rubber is coated with clay or similar material to prevent sticking, and the bales are bound with metal strapping.

International grading standards. The three types of rubber described above are graded according to international standards published by the Rubber Manufacturers Association, Inc. of New York. Blending different grades of the same type or of entirely different types of rubber is quite a common practice of manufacturers to obtain the best properties for some specific purpose. The user also may blend bales from different deliveries of the same type of rubber to obtain a more uniform product. It is uneconomical to wash and strain low grades to raise their quality.

Grading has traditionally relied upon visual inspection and, for that reason, has received serious criticism over the years. Consequently, efforts have been made to find a more scientific basis for grading. In 1950, the Technically Classified Rubber system was devised, employing nine grades based on the physical properties, viscosity and rate of cure. But because viscosity of natural rubber varies with age, industry would not accept the system. Furthermore, a premium was charged for grading; and the plan was initially confined to top-grade rubber. By the time it was extended to the lower grades, it was too late to halt declining interest. Rubber-growing interests always regard top grade as standard and others as off-grades, whereas important sections of the manufacturing industry, such as tire makers, purchase the lower grades.

The Technically Classified Rubber system

In 1965 Standard Malaysian Rubber, a much more ambitious and practical system, based on technical specifications and defined packing requirements, came into use. With some later modifications and extensions, the fundamental principle has taken hold, and rapidly increasing quantities of rubber have been marketed under this system, amounting in 1970 to about 250,000 tons, or 10 percent of the total natural rubber produced (more than 20 percent of the Malaysian output). Grading is based on chemical analysis and controlled by the Rubber Research Institute of Malaysia.

Fundamentally, classification is by dirt content, but other chemical standards are imposed. For each grade there are maximum amounts allowable for various impurities. The Plasticity Retention Index has proved the most controversial quality criterion of the Standard Malaysian Rubber system. That index measures the chemical degradation of the rubber over time; the better the rubber, the less it should break down and, consequently, the higher the Plasticity Retention Index. Though the originators of the Standard system now feel that sufficient experience exists to accept the Index as an essential part of their standard, major rubber growers disagree. They say they have been able to provide first-class rubber that fails this test and samples of poor rubber that pass. Nevertheless, the Standard system has been a great advance over visual inspection and has formed the cornerstone of technical rubber grading.

Chemical and physical properties. Rubber has the empirical formula C_5H_8, indicating a molecule in its simplest form with five carbon and eight hydrogen atoms. Many of these simple molecules commonly combine together, producing a structure whose atomic weight varies within wide limits, depending upon conditions. The composition of raw rubber has been chemically determined both by breaking down the molecule into its simplest form, or C_5H_8 monomer, and by the reverse process of building up rubberlike materials from monomers (polymerization).

Rubber has certain peculiar and useful properties of which the most outstanding is its ability to stretch seven to eight times its original length. During extension, rubber gives out heat; as it contracts, it absorbs heat. The extent to which a piece of rubber can contract after being extended or stretched depends on temperature.

The behaviour of rubber in certain solvents is peculiar. In most organic solvents it swells, part going into solution readily (the sol portion) and part resisting (the gel portion). Mastication of rubber encourages its diffusion in the solvent appreciably, producing "rubber solution," the viscosity of which is substantially reduced by further mastication or by the addition of small amounts of other liquids.

Rubber in solvents

When defining the properties of rubber, care must be taken to distinguish between the chemically pure substance and the commercial product. Commercial latex contains a significant proportion of nonrubber material, much of which remains in the coagulated solid. Dried rubber typically contains 92 to 94 percent chemically pure C_5H_8, 0.13 to 1.20 percent water, 2.50 to 3.20 percent acetone extract, 2.50 to 3.50 percent nitrogenous substances, and 0.15 to 0.50 percent ash.

The nonrubber substances in commercial rubber affect its properties to a greater extent than their proportion would suggest. A high moisture content causes mold to form, damaging market value where appraisal is based on appearance, even though usefulness is not impaired.

The acetone extract contains unsaturated fats and fatty acids, which have a softening effect and also accelerate vulcanization even when present in quite small proportions. The yellow colouring material of pale crepe is found to some extent in the acetone extract. Antioxidants occur in minute amounts but have a very beneficial effect on the aging qualities. Proteins, present in varying degrees, have little effect except upon vulcanization, which they accelerate. Complicated alcohols and other organic substances occur in small proportions.

The ash found in latex contains all the mineral matter and varies in quantity and nature according to circumstances. A certain amount of sand and soil is inevitable. More harmful are the traces of copper and manganese, which act as catalysts for chemical reactions that result in the degradation of rubber. Their action depends upon their state of combination. For example, Cambodian rubbers are always high in manganese, but because the manganese is combined organically it has no deleterious effect.

Chemical examination of rubber is a tedious operation requiring great skill. Hence, it is usually done only where absolutely required as, for example, in arbitration disputes. Of greater practical value and far easier to perform is the physical testing of commercial rubber. It is carried out

on vulcanized compounds and varies in type according to the proposed use of the product (*i.e.,* resistance to or conduction of electric current, resistance to water, acids, and other chemicals). Two important physical tests involve measurement of modulus of elasticity and tensile strength. Modulus of elasticity, defined as the response of rubber to a force applied in a given direction, is generally measured by stretching a piece of rubber of known cross section, the stress being the force applied and the strain, the extension, which, of course, will vary with temperature. Modulus may be measured at 100, 200, 300, 500, or 600 percent extension or at any other appropriate elongation. The result is recorded in kilograms per square centimetre or pounds per square inch. Modulus is normally a function of the state of cure, and as such is considered more important than tensile strength, which is the ultimate stress required to break a standard specimen, again expressed in the same units. Elongation at break is usually measured at the same time.

Physical tests of rubber

One other physical property of cured rubber is frequently quoted because of its simplicity, namely, hardness. It is generally measured by pressing the point of a cone-shaped instrument called an indenter into the rubber and measuring the depth of penetration on an International Rubber Hardness scale.

For specific uses, special laboratory tests have been devised. For example, tire-tread compounds are tested for abrasion and skid resistance. Dynamic testing (*i.e.,* with the rubber in motion) attempts to simulate service conditions; but until it is accepted more widely, full-scale service testing, which is expensive and time-consuming, will continue.

Distinctive uses of natural rubber. Though natural rubber performs well in most applications, some of the newer synthetics surpass it for specialized purposes. For example, acrylonitrile rubber has much better oil resistance, butyl rubber is much less permeable to air, and silicone rubber withstands higher temperatures.

Natural rubber is still preferred in applications that demand elasticity, resilience, tackiness, and low heat buildup. Natural rubber also is chosen when it has no particular technical advantage over synthetics but is competitive for price or availability. Natural rubber is indispensable for the treads of tires for racing cars, bus and truck tires, airplane tires, and any other form of transport where conditions are particularly severe and where a buildup of heat inside the tire could cause a failure. Broadly speaking, the larger the tire, the greater the proportion of natural rubber used in its construction, to minimize heat buildup.

The resilience of rubber is utilized in engine mounts and suspension units of automobiles, and its ability to withstand severe loads for long periods in compression has led to its use in building foundations. These properties, in addition to good cold-temperature resistance, also make it useful in bridge bearings.

Latex represents rubber in its most tractable form. Criticism has long been made of the time and effort spent at plantations in removing water from latex to obtain solid rubber and the subsequent effort made at its destination to render the rubber processable. Dried latex, in fact, has some better physical properties than conventional solid rubber, but the cost of transporting a material containing 60 percent water across the world would be excessive. Concentration of latex to a content of 60 to 70 percent dry rubber is performed either at the source or in the user country by means of centrifuging, evaporation, or a process known as creaming, in which an agent is added to the latex that causes the rubber particles to swell and rise to the surface of the liquid.

Creaming as a concentration process

Latex is an excellent adhesive and is widely used in its natural state, but for most industrial uses it needs additives and vulcanization. Accelerating agents for vulcanizing and other essential fillers are added. One of the oldest uses of latex is in the production of such dipped goods as rubber gloves and prophylactics. The process consists of dipping formers of the correct size and shape into compounded latex, then drying, repeating the operations, and vulcanizing.

Rubber (elastic) thread also is produced from latex. Initially, strips were cut from a thin sheet to form a thread of square cross section. Later, the much superior round thread was obtained by extruding compounded latex into a bath of coagulant through glass nozzles.

Foam rubber has been one of the most important latex products since its discovery in the 1920s. The confinement of thousands of gaseous bubbles in natural rubber cells provides ideal resilience. In the simple manufacturing process, excess ammonia is removed from the latex, and, after a vigorous stirring with dispersions of soap, sodium fluorosilicate is added to the frothy mixture as a gelling agent, along with accelerating agents, zinc oxide, and antioxidant. After the liquid has stood in a mold for several minutes to permit gelling, it is vulcanized. In practice, the process requires the most careful control and the judgment of experienced personnel.

Foam products were made originally in individual molds, which progressed along a conveyor in series and into each of which a predetermined quantity of foam was inserted, the mold closing during its progress to the oven. Washing, drying, and finishing completed the process. For large-scale production, a new method has supplanted this early method, which is still in active service. The process consists of filling a fixed mold with a metered quantity of foam, sealing it with its lid, and evacuating the air; the foam expands to partly fill the resulting vacuum, and then is frozen; next, carbon dioxide gas is passed in to fill the remaining vacuum in the foam, and the temperature is raised to 104° C (219° F) for curing.

The carpet industry has made increasing use of latex not only as a separate foam rubber underlay but as an undercoating on the carpet itself and as an anchoring matrix for tufted carpets. The latex may be either a vulcanizable natural rubber compound or a synthetic self-curing compound.

Foams frequently comprise a blend of natural and styrene-butadiene rubbers, the latter reducing the cost with some reduction in resilience.

SYNTHETIC RUBBER

History. In 1826 Michael Faraday assigned the empirical formula C_5H_8 to rubber. For the next 100 years or so chemists endeavoured to reproduce this molecule synthetically without success. It was only when the search for chemical equivalence was abandoned and adequate physical properties were emphasized that synthetic rubber came into being. The choice fell upon butadiene, C_4H_6, a compound related to natural rubber, to form the basis for synthetic rubber. Though Germany produced synthetic rubber (called methyl rubber) during World War I, it was an inferior substitute by present-day standards. Subsequently, Germany produced Buna rubbers, the first two letters "bu" standing for butadiene, and the last two, "na," for the chemical symbol for sodium, the catalyst used for polymerization.

Buna rubber and its successors

In Germany during World War II, product control remained in the hands of the chemical manufacturer, who gave first priority to quality and ease of production, without much regard for the subsequent difficulties of the user. In the U.S., the rubber producer was also the user; hence a more workable product was developed.

Wartime U.S. synthetic rubber was made from butadiene and styrene. Originally known also as Buna, its name was changed to GR-S (Government Rubber–Styrene) when the government took control. This was the forerunner of modern synthetic rubber. Although ideas for better or cheaper products developed, the government held to the GR-S formula for the duration of the war to avoid complications. Meantime, German chemical engineers perfected low-temperature polymerization, producing a more uniform product. The process, which has come into universal use, is known as redox, coined from the words reduction and oxidation. The two products are often referred to, respectively, as "hot" rubber and "cold" or low-temperature polymer synthetic rubber.

Types of synthetic rubber. In addition to styrene–butadiene, there are many specialized rubbers, sometimes difficult to classify because of the varying systems of nomenclature. Table 1 gives the equivalents that have been used: Some of the newer rubbers—*e.g.,* silicone,

Table 1: Types of Synthetic Rubber

common name	original code name	American government code	current code	chemical description
Synthetic (unspecified)	Buna S	GR-S	SBR	styrene–butadiene
Butyl		GR-I	IIR	polyisobutylene–polyisoprene
Butadiene			BR	polybutadiene (solution polymer–stereospecific)
Neoprene (trade name)		GR-M	CR	polychloroprene
Nitrile	Buna N	GR-N (previously GR-A)	NBR	acrylonitrile–butadiene
Polyisoprene			IR	polyisoprene (solution polymer–stereospecific)
Thiokol (trade name)		GR-P		polysulfide
EPR				ethylene–propylene
EPDM				ethylene–propylene terpolymer

polyurethane, and Hypalon—have no official or generally agreed symbols.

Styrene–butadiene rubber. This general-purpose synthetic rubber, a polymer of butadiene and styrene, exceeds all others in consumption. Its symbol, SBR, covers a range of materials in which proportions of the two ingredients vary. Most styrene–butadiene rubber is manufactured by the emulsion process, in which the styrene and butadiene are brought together in a water solution of soap that acts to disperse, or emulsify, the materials in the solution. Other materials in the solution include catalysts, which initiate the reaction, and stabilizers, which prevent deterioration of the final product. A "short stop" agent is added to stop the reaction at the optimum point, when about 60 percent of the ingredients are converted to synthetic rubber. A modifier produces a material of the desired plasticity, and a coagulant is added to deposit the rubber from the latex.

Styrene, a liquid, boils at 145° C (293° F), but butadiene, a gas at normal temperatures, boils at −4° C (25° F); butadiene may be stored in liquid form under pressure. Both liquids are pumped continuously into reactors and mixed with water, soap, and catalyst by constant agitation at controlled temperature. After the short stop is added, the unconverted butadiene and styrene are recovered and re-used. Then the antioxidant is added and the coagulant to deposit the rubber.

Varieties of SBR, each with its own properties, can be produced either by altering the ratio of the two ingredients; by using hot or cold polymerization; or by varying other elements in the manufacturing process. For example, with 23.5 percent SBR, fatty acid–soap emulsifying agents give a faster curing rubber, while the rosin acid–soap types yield a less tacky final product. Oils of different grades added to the reaction mixture of styrene and butadiene determine the final product's toughness.

A convenient feature of styrene–butadiene rubber is its final packaging in 34-kilogram (75-pound) polyethylene plastic bags, shaped for easy handling and shipping. If the minimum temperature in the user's mixing machine approximates 140° C (284° F), the rubber, bag and all, can be tossed into the mixer where the polyethylene will soften and disappear.

For many purposes SBR directly replaces natural rubber, the choice depending simply on economics. Its particular qualities include abrasion and crack resistance and generally better aging properties. Its limitations are poor strength (without such reinforcing fillers as carbon black), low resilience, low-tear strength (particularly at high temperatures), and poor tack (*i.e.*, it is not tacky or sticky to the touch). These characteristics determine its use in tire treads; its proportions decrease as need for heat resistance increases, until 100 percent natural is reached in the heaviest and most severe uses, such as tires for buses and aircraft. SBR, however, is used in great quantities for tire carcasses; generally some natural rubber is admixed to produce the necessary tack in assembly.

Butyl rubber. Butyl rubber is essentially polyisobutylene, a polymer of isobutylene, which is an isomer of butylene. (Isomers are molecules with the same number of atoms but with different molecular structure and properties.) Isoprene is added during manufacture. Marketed in 1943, after ten years of research work, butyl rubber has enjoyed widespread acceptance because of some of its special properties: low permeability to gases, excellent resistance to oxygen and ozone at normal temperatures, and good resistance to vegetable oils. Butyl rubber is not readily attacked by acids, alkalies, and other chemicals harmful to natural rubber. In addition, it is elastic, odourless, nontoxic, light in colour, and can be worked on normal rubber machinery. Its electrical insulating properties are excellent.

Properties of butyl rubber

Butyl rubber has several disadvantages, especially its incompatibility with natural rubber and the main synthetics. In addition, butyl rubber has a strong affinity for foreign matter and must be carefully stored and screened before manufacture. Finally, it stiffens at low temperatures.

The polymerization of isobutylene is carried out at −95° C (−139° F), with aluminum chloride as a catalyst (Figure 8). Isobutylene and the desired proportion of isoprene are fed continuously into a medium of methyl chloride. After antioxidant is added, crumbs of butyl rubber form

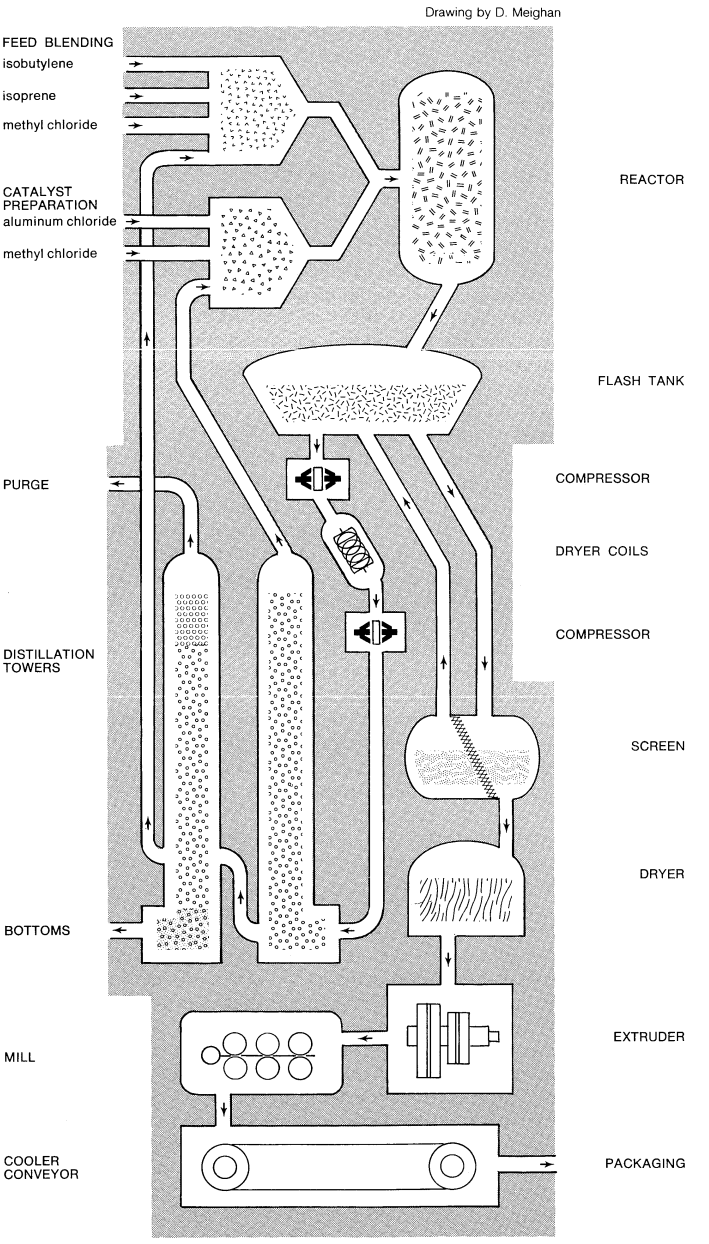


Figure 8: Principal steps in processing butyl rubber.

directly, without any latex stage. Zinc stearate (to prevent agglomeration, or formation of clumps of butyl rubber) also goes into the mixture. Excess reactants are recovered and re-used.

Uses of butyl rubber

Because of its excellent air retention butyl rubber quickly took over from natural rubber for inner tubes in all but the largest sizes. Butyl rubber also plays an important part in the inner liners of tubeless tires. Despite excellent road-holding properties under wet conditions, and quiet, comfortable riding qualities, all-butyl tires have not proved commercially successful because of their poor tread durability. Butyl rubber, however, is used for many automobile components, such as window strips, because of its excellent resistance to oxidation. In its resistance to heat, butyl rubber also plays an indispensable part in tire manufacture, forming the container for the hot water or steam required to vulcanize the inside of tires.

Polybutadiene. Polybutadiene is produced by a hydrocarbon solution process, in contrast with the water-emulsion method for styrene–butadiene rubber. Indeed, in making polybutadiene, water must be rigorously excluded to enable the catalyst to work. The resulting polybutadiene has better physical properties, particularly tensile strength and resilience, than styrene–butadiene rubber.

The main outlet for polybutadiene is in tire treads. Its high resistance to abrasion, low heat buildup, and resistance to cracking are strongly in its favour, particularly when blended with natural or styrene–butadiene rubber in giant tires. Its lower skid resistance when wet and lower tear resistance limit its proportions in car tires. Polybutadiene can absorb large proportions of oil and carbon black, thus reducing the cost of the final product and improving skid resistance appreciably. It is increasingly used to add strength and flexibility to plastics, highly resilient foams, and footwear; its low-temperature performance favours outdoor uses.

Polychloroprene (neoprene). Polychloroprene is the scientific name for the synthetic that bears the proprietary name Neoprene (which, however, has become generic). One of the first synthetic rubbers to be discovered, neoprene was first marketed in 1931. It is prepared by treating acetylene (C_2H_2) with cuprous chloride to form monovinyl acetylene, which, with hydrochloric acid, yields chloroprene. Competitive methods of production exist. It is manufactured in solid form (chips) and as latex.

Because of its general properties of high tensile strength, high resilience, good resistance to oxygen, ozone, and tearing, as well as an oil resistance only slightly inferior to nitrile (see below), neoprene is used extensively in the wire and cable industries, hose, extruded automobile parts, and for protective clothing. It also has excellent resistance to flame.

Nitrile. The original Buna N of German war production, the name nitrile (its symbol NBR) covers a range of copolymers of butadiene and acrylonitrile, the members differing from one another in their acrylonitrile content. The "standard" NBR material contains from 33 to 36 percent acrylonitrile, and others from 20 to 50 percent. The outstanding property of nitrile is its oil resistance, which increases in pace with its acrylonitrile content—but at the expense of low-temperature flexibility. Production is by an emulsion process quite similar to the styrene–butadiene rubber process; it can be made as solid sheets or liquid latex. Nitrile requires the addition of carbon black or other reinforcement to produce satisfactory physical properties, thus reducing the overall cost of nitrile, which is more expensive than conventional rubbers.

Nitrile's oil resistance

Nitrile is mostly used where high oil resistance is required, as for seals, gaskets, or other items subject to contact with hot oils. The rolls for spreading ink in printing and hoses for oil products are other obvious uses. Nitrile's reaction to oil is selective, however; it has excellent resistance only to some types. Nitrile applied to woven and unwoven fabrics improves finish and provides waterproofing. It blends readily with the plastic polyvinyl chloride, a mixture whose use for artificial leather and for paper coating is increasing.

Polyisoprene. The first true equivalent of natural rubber was prepared in 1955 by the polymerization of isoprene. Its manufacture follows the hydrocarbon solvent method for making polybutadiene. Synthetic polyisoprene can replace natural rubber for most purposes, except when the nonrubber constituents of natural become advantageous. The decision is therefore usually based on economics, and since the basic monomer isoprene is expensive, either synthetic polybutadiene or natural rubber is usually favoured.

Silicone. A specialty rubber, silicone is limited in use by its cost; but its exclusive ability to retain its physical properties at extreme temperatures, from −74° C (−101° F) to 315° C (599° F), makes up for its high cost. It is the first "inorganic" rubber; the polymer has a backbone of silicon and oxygen instead of the conventional carbon chain.

Although the raw material silica is cheap and plentiful, the manufacturing process is difficult and expensive. A range of products is available, their properties largely influenced by the molecular structure of the silicone. Its retention of electrical properties, resistance to oxygen and ozone, and inertness at extremely high temperatures recommend uses in aircraft and outer space equipment, and in the automobile field. Its inertness and nontoxicity encourage use for food and surgical applications.

Polysulfide rubber (Thiokol). Polysulfide rubber is the technical term for the substance bearing the trade name Thiokol, which has become generic. Originally synthesized in 1927, it is known for its excellent resistance to organic solvents and is marketed as a liquid, a putty, and in solid form. Its tensile strength does not compare with conventional rubbers, but it offers excellent resistance to weathering and permeability over a wide range of temperature.

Ethylene–propylene rubber. A versatile new synthetic, ethylene–propylene rubber exhibits excellent resistance to oxygen, ozone, acids, alkalies, and other chemicals over a wide range of temperature. It has good electrical properties and is quite lightweight, with an enormous capacity for acceptance of oil, thereby reducing the final cost. The addition of carbon black used as a filler, along with the oil, acts to raise the low tensile strength.

Polyurethane rubber. Polyurethane rubber includes a large range of materials of varied compositions, properties, and uses. In this group of rapidly growing importance, some of the polymers help significantly to bridge the gap between rubber and plastics. All are based on isocyanates, organic compounds that combine with other organic compounds such as alcohols, esters, ethers, and amides. The wide variety of materials available, when properly controlled, offers a series of polymers with a broad range of properties and forms.

Casting, mixing, and milling of polyurethanes

Polyurethanes may be cast from the liquid or semiliquid state, or mixed, milled, and vulcanized like other rubbers. Although they have high tensile strength and good resistance to abrasion and to oil and other solvents, they are not suited for normal tire service because of their rapid drop in abrasion resistance as temperature increases. Polyurethane is, however, finding increased use in small, solid tires. Probably the most popular uses of polyurethane are as foams and surface coatings: rigid foams for lining refrigerators and flexible foams as a substitute for natural rubber and durable finishes for floors and furniture.

Chlorosulfonated polyethylene (Hypalon). Chlorosulfonated polyethylene is normally known by its trade name of Hypalon. It is made by introducing first sulfur dioxide into a solution of polyethylene under strictly controlled conditions and then chlorine at higher temperature. Chlorine content may vary between 20 and 45 percent and sulfur from 0.5 to 2.75 percent. The physical properties depend much more upon the sulfur than the chlorine content. Rubbery compounds are produced with low sulfur and metallic oxide curing. Hypalon is outstanding in its resistance to weathering, ozone, heat, flame, and chemicals. Cost alone restricts its wider adoption.

Fluorocarbon polymers. Probably the best known fluorocarbon polymer is PTFE—polytetrafluoroethylene—although similar products share polytetrafluoroethylene's outstanding qualities of heat resistance, minimal friction, and freedom from surface adhesion. Originally these materials could not be extruded or vulcanized, but rubbery polymers have been developed that can be worked on

conventional equipment. A recent use that has popularized this polymer is as a lining for cooking utensils; its nonsticking quality facilitates removal of cooking residues by simple washing.

Reclaimed rubber. Reclaimed rubber, or vulcanized rubber reworked to render it suitable as raw material, has become an important element in the industry. The starting point in reclaim rubber manufacture is the cured rubber scrap, natural or synthetic in origin, which is segregated into separate incompatible rubbers and then graded according to quality and intended use. Reclamation methods differ in sophistication, so that an appreciable range of reclaims is available.

It is not profitable to use reclaim unless it costs no more than half as much as virgin rubber, and when that new rubber is low-cost synthetic, the market for reclaim suffers.

Digestion of scrap rubber

Reclaim is manufactured in various ways. The simplest is digestion of scrap (to which oil has been added) either in caustic soda or in zinc chloride (the more modern method) to remove nonrubber products; it then is milled. Under the newer, continuous "reclaimator" method, any fabric is first removed mechanically; the rubber then is ground, mixed with oil, and extruded.

COMPOUNDING

Blending rubbers in prescribed proportions and adding chemicals in accordance with a "formula," or schedule, is called compounding. Initially, these formulas were guarded with great secrecy, often unnecessarily. Such secrecy has largely disappeared; most suppliers disclose the contents of their products. Although some manufacturers prefer to show each constituent of a formula as a percentage of the whole, the more usual practice is to list ingredients in relationship to 100 parts of rubber hydrocarbon.

The materials added to rubber are known variously as drugs, pigments (not necessarily confined to colours), and fillers. Since all rubbers, natural and synthetic, with the exception of some already containing fluorine, require addition of chemicals, a filler is not to be mistaken for an adulterant or cheapener. Of course, the vulcanizing agent, sulfur, must be added. But by itself it is slow, so that it needs accelerators, which, in turn, can be speeded up with secondary accelerators, or activators. To delay the decomposition of rubber products, antioxidants and antiozonants are required. (These and all other such protective materials are termed antidegradants.) To increase strength, rubbers need reinforcing agents—*e.g.*, carbon blacks, which are available in a wide range and impart a black coloration. Other materials may be added, as well as oils for processing, plasticizing, or softening. Finally, inert materials may be added for bulk, thus reducing cost.

Reinforcing and filling materials. The outstanding material used for the last 50 years or more for improving elasticity and tensile strength of rubber has been carbon black. Early carbon black was made by burning natural gas in multiple fine jets and collecting the product on cold metal. Under the modern method, "furnace type" black is made by burning oil in a controlled atmosphere. The many grades of blacks have been distinguished by initials; HAF (high abrasion furnace) has been the mainstay. Every pound of rubber used in tire treads requires at least one-half pound of black; tubes require more, and carcasses only slightly less.

Carbon-black filler

To abate the dust nuisance associated with the use of carbon black, alternatives have been investigated, especially silica in finely divided condition. Though considerably more expensive than carbon black, it has had success as a partial replacement in treads for large off-the-road tires.

Inert inorganic fillers basically add substance, though clays and whiting have some technical value; organic fillers, more akin to rubber, are now increasingly used.

Colouring agents. Pigments for colouring rubber must withstand curing temperatures and not interfere with or be affected by other ingredients. In high-temperature cures, therefore, the choice is limited to inorganic substances: zinc oxide and some titanium dioxide, plus a trace of ultramarine (to offset any creamy tinge) to achieve whiteness; iron oxide, to produce browns and reds. For rubbers curing at lower temperatures, a wide choice of organic dyes is available. For colour work, any added antioxidant must be nonstaining.

Plasticizers and softeners. Adding mineral oils to aid processing is as old as the industry. Usually 5 to 8 percent of oil was added to natural rubber. The introduction of synthetics, with their greater affinity for oil, has led to more scientific standards.

Studies have shown that increasing the oil content of natural rubber improves certain of its properties, particularly traction of tires on snow-packed roads. In addition to mineral oils, a wide variety of softeners has been used, including coal-tar by-products such as bitumen, and Vaseline, fatty acids, and wood resins. Pine tar has always been popular with natural rubber because of its value in dispersing carbon black.

Vulcanizing agent. The outstanding vulcanization agent and the one longest used is sulfur. Cheap, effective, and relatively free from problems, it is added in proportions of up to 3 percent, and much higher for hard rubbers. Sulfur has one peculiarity: regardless of how little or how much is used in vulcanizing, some remains uncombined and is known as free sulfur, which migrates to the surface forming an unattractive sulfur bloom. Insoluble sulfur, considerably more expensive, can be substituted to avoid blooming.

Sulfur bloom

Accelerators of vulcanization. Accelerators or catalysts form an ever-growing list. Mercaptobenzothiazole, which revolutionized vulcanization in the 1920s, is still in use, although much safer and more efficient products have been developed since. As a class, accelerators have enabled curing times and temperatures to be reduced significantly, rendering the operation more efficient and permitting the introduction of valuable materials that would be destroyed under the earlier conditions.

Antioxidants. These, as their name implies, resist the action of oxygen. They form an important section of the group of materials known as antidegradants, which encompasses all materials resisting the deterioration of rubber. The amount of antioxidant added per pound of rubber has steadily risen, including the amounts indigenous to natural rubber and the amounts introduced as stabilizers in styrene-butadiene and other synthetics during manufacture.

Antioxidants may be divided into two main classes: amine derivatives, among which the more powerful are generally complicated condensation products with aldehydes or ketones, which cannot be used in light-coloured stocks because of staining; and phenol derivatives, which stain less and are more efficient. A third and smaller class of antidegradants fulfill such specific purposes as resisting the catalytic action of copper in breaking down rubber.

Derivatives of *p*–phenylenediamine occupy a prominent position among modern materials. Despite their excellent resistance to oxygen and ozone and good flex resistance, great care has to be taken to minimize staining and more particularly to ensure freedom from skin irritation.

Zinc oxide. Versatile zinc oxide belongs in a category by itself. Mention has been made of its use as a colouring agent. Although its use as a reinforcing agent has declined, zinc oxide finds its way into most formulas, primarily because of its activating powers. It makes accelerators more efficient and generally gives stability to products. Its high density makes it an expensive addition because volume rather than weight is the criterion in filling molds. Stearic acid should be associated with zinc oxide, for although as a fatty acid it might be considered a softener, it is normally added in such proportions as to convert the zinc oxide to zinc stearate, which is thereby made more effective.

Other compounding materials. Resins fulfill a variety of important functions. Mention has been made of wood resin softeners, and also of phenolic resins for high-temperature curing. Other resins are added to rubber to make it tacky and still others to stiffen it.

Several waxes may be added either as softeners or to form protective blooms on the surface of the product.

With the introduction of faster accelerators in manufacture, retarders have been developed; they are designed to make processing safer without slowing down vulcanization.

Peptizers, organic sulfur compounds that save mixing

time by breaking down the huge molecular structures of the raw rubbers, are effective in small amounts. They are selective, however, and their performance is dependent on the presence or absence of other materials. Their relatively high cost must be justified by the mixing time saved.

Polyac and Elastopar

Where butyl rubber is used in compounding, the material Polyac serves as a promoter, a substance a minute quantity of which when heated with the polymer speeds subsequent action with the accelerator. A somewhat similar substance is Elastopar, used with styrene–butadiene rubber, but not effective below 149° C (300° F). Both promoters act vigorously and because of the difficulty of control are not always acceptable. To overcome that difficulty, they are sold mixed with an excess of inert material.

Lubricating additives cover a wide range. Dusting powders applied to uncured rubber prevent pieces from sticking together and ease their removal from molds. The standard materials are talc or french chalk. Zinc stearate is much more expensive but has the outstanding advantage of not causing cracking because it is soluble in rubber and, therefore, does not act as a foreign substance at the point of contact between two rubber surfaces. Liquid lubricants fall broadly into two categories: suspensions of talc, clay, or soap in water or alcohol; and the higher costing silicones, which have proved their effectiveness as mold release agents.

Vast quantities of solvents are consumed by the industry. Basically, they are petroleum distillates often known as naphthas, added in the factory to provide or restore tack to uncured rubber. Other solvents include aromatics obtained from coal tar and such chlorinated hydrocarbons as trichloroethylene and carbon tetrachloride.

PROCESSES

Mastication. Mastication is the process of reducing the size of the large organic molecules in natural rubber to increase its plasticity and thus facilitate subsequent mixing. Mastication may be accomplished in a specially designed machine, on open mills, or by merely giving the rubber an extra two minutes in a mixer before the addition of other ingredients. The process, therefore, is not precise but depends upon the condition of the rubber. The necessity of mastication may be minimized by the prior controlled storage of rubber at elevated temperatures, the use of low-viscosity rubber, or the use of special plasticizers.

Mixing. Open mills originally performed mixing. They were machines with two adjacent rollers mounted horizontally and rotating in opposite directions; the clearance between the rollers (the nip) was adjustable. Open mills, however, are slow, hazardous, and not amenable to temperature control; their replacement by internal mixers (see below) was accelerated by the increasing incorporation of powdery carbon black. Open mills still are used for warming up rubber before extrusion or sheeting, for low-temperature mixing, and other specialized work.

Internal mixers

An internal mixer consists of two horizontally mounted rotors with blades that intermesh and shear the compound against the lining, with most of the mixing done in the small clearance between the outside of the rotors and the lining of the mixer. Mixing is regarded as complete at the end of a given time period, after a given temperature is attained, or after a certain amount of power is consumed. The mixed batch, normally in the form of discrete lumps, is sheeted out by dropping it on a mill mounted immediately below the mixer.

Forming. Uncured solid rubber not required in sheets is shaped or formed, generally in an extruder or tuber. The apparatus is frequently compared with a sausage machine because raw material is fed into a hopper at one end and forced through an aperture at the other, along a barrel containing a tightly fitting screw. Temperature is controlled by hot or cold water circulating around the barrel, and a variety of dies can be fitted to the head to produce the desired shapes. The rubber must be consistent in plasticity, for the cut of the die is correct for only one set of conditions, including temperature and speed. Extruders, defined by screw diameter, range from 38 millimetres (1.5 inches) up to as much as 305 millimetres (12 inches) for very large tire treads.

The compounded rubber is warmed up before being shaped, generally on open mills. The process is extravagant in use of time and material for setting up before the specified dimensions are obtained; and when the product is cut, it shrinks on cooling, with "belled ends," sections that have to be discarded. Extruders may be used in twin or triple formation, each with its own rubber stock, to provide hot unions—*e.g.*, a tough tread on top of a more resilient base or a tread with flexible sidewalls.

Injection molding. The waste and delays associated with extrusion have led to more efficient processes such as cold feed mixer-extruders and injection molding machines similar to those used for plastics. With modification to guard against incipient cure in the barrel, they extrude the rubber, not through a die, but directly into a mold where it is given a very short cure. Success depends on extremely long runs, which are not common. Multiheads that allow four molds to be operated from one machine increase output.

Calendering. Calenders have two main functions: one, to produce sheet rubber of accurate and consistent gauge up to about 3 millimetres (0.12 inch) or more by plying up (using multiple plies); and, two, to rubberize fabric with great accuracy. In its crudest form, a calender is essentially a two-roller mill, although the rollers are normally mounted vertically. Modern calenders comprise three or four rollers ground very accurately.

Rolling plies and coating fabrics

In the typical three-roller calender for rubberizing fabric, all rollers lie in the same vertical plane, the top and bottom rotating in the same direction. Warmed-up rubber is fed into the nip between the two top rollers; it follows around the middle roller, and is applied to fabric entering the machine between the two lower rollers. The speed of the centre roller can be varied in relation to the bottom: when it is even with the bottom roller, it coats the fabric; when it is higher, it frictions, or forces the rubber into the interstices of the fabric.

A three-roller calender treats only one side of fabric at one pass. To rubberize both sides, the alternatives are a second pass, the use of a second calender (preferably in train with the first) or a four-roller calender, which coats both sides in one operation. The four rollers may be in line vertically, or the top roller may be offset to form an inverted "L." A calender must be supported with an array of auxiliary equipment, including variable-speed motors; efficient means of feeding both rubber and fabric and collecting the treated material, spreading the fabric across the roller under uniform pressure; accurate gauging; and temperature control. The rollers are traditionally cored to permit the flow of cold water or steam, as required. In a more modern method of temperature control, the rollers are drilled longitudinally forming a series of bores that allow steam or water to circulate within 2.5 centimetres (one inch) of the surface.

Vulcanization. Reference already has been made to the principles of vulcanization and the chemicals used in the process. The simplest form of vulcanizer is the heater or autoclave, a vessel capable of retaining moderate pressures of steam and air (or inert gas). Simple shapes, and particularly thick sections of uncured rubber, may be exposed directly to steam, or molds with their cavities filled with uncured rubber may be inserted in the vessels.So-called bag-o-matic curing is the modern method of tire vulcanization, except for extremely large off-the-road tires. Bag-o-matic curing is entirely automatic and has reduced cure times for car tires from about two hours to 12 to 14 minutes. Heat is applied to the inside of the uncured tire through the medium of a one-quarter-inch gauge bladder, which has supplanted the less efficient air bag. Time, temperature, and pressure are controlled automatically.

Bag-o-matic curing

Smaller rubber products are normally formed in a mechanical press, usually the daylight press on which a number of individual molds can be mounted. The molds are held in a closed condition by a hydraulic ram operating in a vertical plane. Another technique employs a single press that accommodates one mold, possibly with several cavities, the two parts of the mold being held together by mechanical pressure. The operation of the single press is largely automatic.

Cold curing, or vulcanization at ambient temperature, is still practiced but only where it is unavoidable, as in repairing belts *in situ.* Special accelerators can perform at room temperatures, but controlled heat, time, and pressure are always preferable.

Infrared and high-energy radiation. These methods of vulcanization are of scientific rather than commercial interest. Use of internally generated heat to cure thick articles saves considerable time. Because of the poor conductivity of rubber, heat applied to the surface only penetrates slowly. The initial expense of equipment, plus operating costs, has limited interest to specialized high-temperature operations.

MAJOR APPLICATIONS

Tires and tire products. Automobile tires absorb 60 to 70 percent of the available rubber. As the outside diameter of tires has dropped progressively, widths of treads have increased, providing wider contact area with the ground and improved stability. The major strength of the wide-tread radial tires is built in transversely to the direction of motion; a series of inextensible narrow bands under the centre of the tread prevents its spreading. The conventional tire, in contrast, has cross plies. Increasing varieties of carcass materials are available: cotton, rayon, nylon, polyester, steel wire, and glass fibre, the last two for radial tires.

Tire-making remains a blend of craftsmanship and automation. Despite advances in mechanization, personal skill and experience are dominant in extruding, calendering, and in assembling all the components together in the process of tire building. The carcasses are constructed by laying ply upon ply of rubberized "fabric" on a cylindrical drum, followed by the other components, and ending with the tread and sidewalls (generally together as one unit). Adhesion in the uncured state relies upon the natural tack of the rubber.

Rainwear and shoes. Rainwear represented the first commercial use of rubber. Vulcanization overcame the obstacles encountered by Macintosh and the process of rubber coating fabric has continued into modern times. Doughs of rubber, naphtha, and other ingredients are fed between the roller transmitting the fabric and a "doctor" blade that spreads the rubber onto the fabric and controls thickness. The treated fabric then passes over a heated framework and is wound into rolls. If the heat is not sufficient to cure the rubber, an additional operation is performed.

Spreading has largely given way to the greater speed and accuracy of calendering, as has rubber to oil-soap treatment for shower-proofing and to plastics for more serious protection.

Shoes consume more rubber than any other commodity except tires. Styrene–butadiene types of rubber with high styrene content are favoured for soles because their wear characteristics render carbon black unnecessary. Attachment of soles to uppers by stitching has been largely replaced by either polymer adhesives or direct vulcanizing, a process requiring expensive equipment but providing rapid curing of two-and-one-half to three minutes. One-piece soles and heels made of rubber resins have been used successfully in children's footwear. Microcellular rubber, which feels soft and comfortable in wear, also has gone into soling.

Industrial uses. One highly important industrial use of rubber is in belting. The tendency for each piece of machinery to have its own electric motor has limited the use of driving belts, but the conveyance of coal and other minerals over thousands of yards has provided valuable business for the manufacturer. Strength is obtained by plying up any of the tire textiles, with polyester preferred where nonstretch is critical. Where fire is a hazard, the outer cover of belting is usually polyvinyl chloride. Cure may be either continuous or by sections, in long, specially constructed presses. The development of travelling sidewalks for passengers at airports and similar traffic centres opens up new and substantial possibilities.

Hoses of all shapes and sizes form an important sector of the rubber business. Small ones, such as garden hoses, are being replaced by polyvinyl chloride largely on a weight basis; otherwise an endless variety exists. The type of synthetic used depends on the material to be transmitted; some hoses must carry corrosive chemicals at high and low temperatures. Type of construction varies with the weight of the fluid to be transported and its pressure. Small all-rubber hoses may be molded to complicated shapes, but generally they are an extruded product. If reinforcement is required, it is primarily cord fabric plaited just ahead of the extruder. An outer cover, possibly of different composition, designed to resist abrasion, follows, and if further reinforcement is required, wire is wound spirally. Small hoses with strong walls that do not collapse under heat may be coiled and so vulcanized, but the larger types must be cured straight to prevent collapse.

Cables consume a substantial amount of rubber as insulation. Wire (normally copper) is drawn through an extruder and the insulation applied. Butyl rubber is preferred because of its resistance to heat and moisture; an outer cover of either butyl rubber or neoprene is added, but both are increasingly being replaced by polyvinyl chloride.

Rubber is useful for many more industrial or engineering purposes, generally to insulate against sound or vibration or both, as in building and bridge foundations. A wide range of rubber springs and flexible seals performs valuable service; rubber fenders are ideal to protect the sides of ships and docks because the rubber retains its resilience for a score or more years despite destructive workings of the sea, oxygen, ozone, and other agencies. (L.R.M./Ed.)

Plastics and resins

Plastics, in the modern meaning of the word, are synthetic materials that are capable of being formed into usable products by heating, milling, molding, and similar processes. The term is derived from the Greek *plastikos,* "to form." By some interpretations of the word, rubber and other natural products are said to be plastic, but the modern definition, especially in relation to industry, excludes natural rubber and also such other natural products as wood, leather, and metals. Resins are both natural and synthetic. Natural resins range from pitch and asafetida to frankincense, myrrh, and amber. Synthetic resins are the bases of an enormous range of products the growth of which shows no signs of abating.

The distinction between plastics and resins is at best arbitrary, since many of today's synthetic materials can properly be called both resins and plastics. Historically, it appears that the term resin was applied to those products primarily used as substitutes for the natural product in coating compositions, whereas the term plastic was applied to designate those compositions that involved a molding operation in their fabrication.

By custom, the fibre industry is considered to be outside the plastics industry, although the same raw materials—polyamides (nylon), cellulose, and cellulose acetate—are used by the two industries. For a detailed discussion of these substances and their use in the manufacture of fibres, see the major section *Man-made fibres.*

Composition and structure of plastics

The characteristic ability of most plastics to soften but not melt when heated, so that their shape can be changed under mechanical stress without losing cohesion, producing a rigid new form upon cooling, is a property resulting from their molecular structure. Plastics are composed of giant molecules, known as polymers, with immensely long chains of repeating units derived from short molecules; in the case of addition polymers the short molecules are known as monomers. The chains are built up by formation of chemical bonds. The assembly of polymeric chains is held together by a number of different types of forces: (a) London, Van der Waals, or dispersion forces, which are relatively weak; (b) for more polar materials, hydrogen bonding or dipole-dipole interactions can be involved and result, in general, in somewhat higher softening points; (c) the long polymer molecules are inevitably entangled with each other and considerable forces are required to disentangle them. When heated, these chains move apart sufficiently to permit them to slide over one another but retain their cohesiveness because of the forces still operating. By

contrast, the components in crystals of small molecules move apart sufficiently to melt to mobile liquids.

This ability to slide into new positions, which are retained on cooling, is the basis of plastics fabricating processes, such as molding or extrusion of sheets or other forms from heated granules, vacuum forming of sheets into desired forms, blowing of extruded tubes into bottles, and other processes. At higher temperatures maximum mobility is attained, allowing injection molding (in which the plastic is injected into a hollow casting and cooled) into an infinite variety of forms. These processes require only moderately high temperatures, usually in the range of 150–250° C (300–480° F), and are very versatile, enabling easy fabrication of a variety of forms. Plastics also have the advantages of toughness, lightness, and of lack of colour, giving rise to an enormous range of coloured products by the addition of dyes and pigments; some plastics are transparent.

Intermeshing of the long-chain molecules of the polymers of which plastics are composed is responsible for their mechanical toughness and ductility, in contrast to the brittleness of low molecular weight organic substances, whether glasslike or crystalline. The hardness and ductility of polymers may be modified by the choice of the constituents that make up the chains. The intensity of the molecular attraction between neighbouring chains and flexibility within the chains (*i.e.,* freedom or restriction of rotation about main chain bonds) are important factors.

All these polymers are known as thermoplastics. Some long-chain materials known as thermosets can also be fabricated; on heating or with chemical treatment during fabrication, the chains of thermosets link with one another to produce a network structure that prevents softening on reheating. Phenol–formaldehyde is a typical thermoset.

HISTORY OF THE PLASTICS INDUSTRY

The first plastic was Parkesine, later called Xylonite, exhibited by the English chemist and inventor Alexander Parkes in 1862. Essentially this substance was nitrocellulose softened by vegetable oils and a little camphor. It was John W. Hyatt in the United States who recognized in 1869 the vital plasticizing effect of camphor and named the product Celluloid. Though primarily regarded as a substitute for ivory and tortoiseshell, Celluloid, despite its inflammability, found substantial early uses as carriage and automobile windshields and motion picture film.

Celluloid

The first completely synthetic plastic was commercially produced from phenol and formaldehyde in 1910 by Leo Hendrik Baekeland, a Belgian-born American chemist. A dark thermosetting resin, later called Bakelite, this substance was used in the electrical industry. The first colourless resin, from urea–formaldehyde, though patented in 1918, was not marketed until 1928. Though available in any colour, it was moisture adsorbent (that is, moisture clung to its surface); this problem was overcome in 1939 by using melamine instead of urea.

Meanwhile, cellulose acetates received intensive study. At first, only the triacetate, soluble in toxic solvents, was produced and applied as a waterproof varnish to the fabric on airplane wings; at the end of World War I, the triacetate was manufactured in powdered form for later melting and molding.

During the same period, transparent acetate film was developed and used for packaging. With development of methylene chloride manufacture, triacetate film came into use as a photographic base in 1935.

Three principal factors contributed to the immense progress in plastics that took place after 1920. Of these, the work of the German chemist Hermann Staudinger is outstanding. In May 1922 he published a paper that demonstrated that rubber was composed of a chain of isoprene units, and the word macromolecule first appeared. His views aroused much opposition and this led him to an intensive study of styrene resin. By examination of the viscosities of its solutions, Staudinger showed that styrene resin was an assembly of macromolecules or giant polymers of different lengths that he called polystyrene. Staudinger's discovery of the theoretical background of polymerization coincided with two other factors: the growth of the chemical industry in general and a political impetus in Germany toward independence from many natural materials, leading to the industrial development of processes long known in the laboratory.

Vinyl chloride, discovered in 1835, is formed by the reaction of acetylene with hydrochloric acid. The polymer polyvinyl chloride (PVC) was first produced in 1912. There was a similar background to polymethyl methacrylate, starting with acrylates, which were polymerized in the laboratory in 1880 and came into production in 1927. A great deal of this work was done in Germany, including large-scale production of polystyrene.

Plastics research and manufacture was proceeding on a considerable scale in the United States as well. Outstanding was the study of polymers in the laboratory of E.I. du Pont de Nemours & Company from 1928 onward, which led to the "superpolyamide" now known as nylon.

The next great step forward was the British invention of polyethylene, which came into commercial production in 1939. Originally of unique value in radar, it is now used in enormous quantities in the form of film and in products made by injection-molding techniques. Further developments included polyurethanes in 1937, epoxy resins in 1943, polyethylene terephthalate in 1941, and polycarbonates in 1956.

Invention of polyethylene

Along with the work on polymers, there was a corresponding development of special machinery for plastics fabrication. When Celluloid was invented, the only equipment available for possible fabrication was that used in the rubber industry, such as the crude hydraulically operated "stuffing" machines for extrusion. The first patent for an extrusion machine employing an Archimedean screw was dated 1879; as thermoplastics were developed, more of these machines were used until one was designed especially for thermoplastics in 1935. Later, extrusion machines were produced for manufacture of plastic sheets and film and for blown bottles. The injection-molding machine was introduced in 1921. The first automatic machine was produced in 1937; later, an improved feed mechanism was added.

FORMATION AND USE OF PLASTIC MATERIALS

Polymerization. *Principles.* The manufacture of plastics depends upon the building of chains and networks either by condensation polymerization or by addition polymerization.

A condensation polymer is formed by a synthesis that involves the stepwise reaction of reactive molecules with one another, with the elimination of small molecules such as water. The reaction gradually slows down as polymers are built up.

An addition polymer forms chains by the linking of small identical units (which may be diversified to a degree) without elimination of small molecules.

The most important concept in condensation polymers is that of "functionality" (f)—*i.e.,* the number of reactive groups in each molecule participating in the chain buildup. Each molecule must have at least two reactive groups, of which hydroxyl (— OH), acidic endings (— COOH), and amine endings (— NH_2) are the simplest.

Hydroxyl is characteristic of alcohol endings, combining with an acid ending to give an ester, the polymer being known as a polyester. Examples are polyethylene terephthalate (by reaction of ethylene glycol containing hydroxyl groups at each end and terephthalic acid containing two acidic groups) and polycarbonate resins. (For a description of polymerization, see CHEMICAL COMPOUNDS.)

Alcohols are a particular class of oxygen-containing chemical compounds with a structure analogous to ethyl alcohol (C_2H_5OH). Amines are various compounds derived from ammonia by replacement of hydrogen by one or more hydrocarbon radicals (molecular groups that act as a unit). Esters are compounds formed by the reaction between an acid and an alcohol or phenol with the elimination of water.

Processes. Bulk addition polymerization of pure monomers is mainly confined to styrene and methyl methacrylate. The process is highly exothermic, or heat producing. The dissipation of heat (necessary to maintain

chain length) is achieved in the case of styrene by intensive stirring of the viscous, partially polymerized mixture, which is then passed down a tower through zones of increasing temperature. Alternatively, polymerization may be completed in containers or molds that are small enough to avoid an excessive temperature rise as a result of the heat released during polymerization.

Methyl methacrylate is also partially polymerized before being poured into molds consisting of between sheets of plate glass, to produce clear acrylic sheet.

Ethylene polymerization

Ethylene is polymerized in tubular reactors about 30 metres (100 feet) long and less than 25 millimetres (one inch) in diameter at pressures of 600–3,000 atmospheres (one atmosphere equals the pressure of air at sea level) to give 10–20 percent conversion to low-density polyethylene. Residual gas is recycled.

Polymerization of monomers in solution allows easy temperature control, but the molecular weight of polymers formed is reduced because of chain transfer reactions. (Such reactions involve growing polymer chains reacting with small solvent molecules to give inactive polymers and starting a new chain, rather than allowing the original chain to grow to the high molecular weight that would have been obtained in the absence of such transfer.) Solvent removal from such a solution may also be very difficult. The process can be applied advantageously to vinyl acetate and acrylic esters.

Suspension polymerization producing beads of plastic is extensively applied to styrene, methyl methacrylate, vinyl chloride, and vinyl acetate. The monomer, in which the initiator or catalyst must be soluble, is maintained in droplet form suspended in water by agitation in the presence of a stabilizer such as gelatin, each droplet of monomer undergoing bulk polymerization.

In emulsion polymerization the monomer is dispersed in water by means of a surface-active agent (a substance slightly soluble in water that reduces the surface tension of a liquid), its bulk aggregating into tiny particles held in suspension (micelles). The monomer enters the hydrocarbon part of the surface-active micelles and is polymerized there by a water-soluble catalyst. This process is particularly useful for the preparation of very high molecular weight polymers.

Exposure of certain substances to X-ray or ultraviolet radiation initiates chain reactions that can be used for manufacture of such thermoplastics as polyethylene and polyvinyl chloride.

Thermosetting resins. Resins that cannot be softened by heating include the phenolics, furan resins, aminoplastics, alkyds, allyls, epoxy resins, polyurethanes, some polyesters, and silicones.

Phenolics or phenol–aldehydes. The important commercial phenolic resin Bakelite is based on phenol and formaldehyde. The two processes in general use are the one-step process producing resol resins (the first stage in the formation of a phenolic resin) that are either liquid or brittle, soluble, fusible solids, from more than one molecule of formaldehyde per phenol molecule; and the two-step process, using an excess of phenol to produce novolacs, resins that have no reactive methylol groups and must be mixed with an aldehyde to undergo further reaction.

Resol resins thermoset on heating and are used for adhesives. Novolacs require a further source of formaldehyde in the form of hexamethylenetetramine to produce molding powders. Both resins are run out from the reaction vessel, after removal of water by distillation, and ground up, then compounded on heated rolls with fillers that vary from wood flour to mica; for strength and heat resistance fibrous asbestos is used as a filler (hexamethylenetetramine is also added at this stage in the case of the two-step resin). Final grinding produces the molding powders, which on further heat treatment will yield the typical thermoset resin.

Characteristics of phenolic moldings

Phenolic moldings are resistant to heat, chemicals, and moisture and are preferred for wet-dry applications as in washing machines. Their stability to heat and low heat conductivity suit them for use in appliance parts, and their electrical insulation qualities qualify them for electric fittings such as switches, plugs, and distributor caps; resistance to hydraulic fluids has led to their use in automotive parts. All these applications have been made more economical by the development of injection molding and extrusion methods not hitherto applied to thermosets. Complex phenols from, *e.g.*, cashew-nut shell liquid, are used in manufacture of brake linings.

Furan resins. Furfural is a five-membered ring compound (*i.e.*, the basic molecule has a ring shape and contains five atoms) of four carbon atoms and one oxygen atom, carrying the aldehyde group, — CHO; it reacts like formaldehyde with phenols in the presence of an acid catalyst to give a rigid polymer with high chemical resistance, used for coatings in industry. It can be prepared in semiliquid form with a low viscosity and remarkable penetrating power when applied to porous forms such as foundry sand cores or graphite blocks, being in this respect superior to other liquid resins.

Aminoplastics. Urea resins are made by the condensation in aqueous solution of formaldehyde and urea in the presence of ammonia as an alkaline catalyst, giving a colourless syrupy solution to which cellulose filler is added to yield a molding powder upon drying, which when heated in a mold gives a water-white (transparent) molding unless previously coloured by pigment.

The filler confers considerable strength, so that thin sections such as in cups and tumblers can be molded. Very large quantities of urea–formaldehyde resin are used in closures, kitchen and bathroom hardware details, and electric appliance housings and fittings.

Melamine behaves in the same way as urea, but the product is more moisture resistant, harder, and stronger, leading to wide use for plates and food containers. Melamine moldings are glossy and harder than any other plastic and retain a dust-free surface. Solutions of the thermoplastic forms of urea–formaldehyde resins are widely used as bonding agents for plywood and wood-fibre products.

Alkyds. Alkyds are polyesters, generally of phthalic acid (with two acid groups) and glycerol, a triol—*i.e.*, an alcohol with three hydroxyl groups. The solid resins are molded at high speed under low pressure, cured quickly, and are used where insulating properties, strength, and dimensional stability over a wide range of voltage, frequency, temperature, and humidity are required, as in vacuum-tube bases and automotive ignition parts and with glass-fibre reinforcement for switchgear and housings for portable tools.

DAP and DAIP resins

Polyesters of unsaturated alcohols. The resins known as DAP and DAIP, are crossliked allyl esters of phthalic and isophthalic acid, respectively. They are notable for maintaining rigidity and excellent electrical properties at temperatures up to 230° C (450° F), properties also manifested by allylic resin-impregnated glass cloth, used in aircraft and missile parts. Other advantages are good storage life and absence of gas evolution during polymerization. The resin allyl diglycol carbonate, optically clear and colourless, is used for making cast objects; fully cured castings are more heat and abrasion resistant than other cast resins.

Epoxy resins. Epoxy resins have outstanding mechanical and electrical properties, dimensional stability, resistance to heat and chemicals, and adhesion to other materials. They are used for casting, potting, encapsulation, protective coatings, and adhesives, and for reinforced moldings and laminates of the highest quality. Popular adhesives (epoxy glues) contain the resin components and the curing agent, usually an amine or an anhydride, in separate packages. The two are mixed just before use.

Polyurethanes. Formed by the reaction between diisocyanates and polyols (multihydroxy compounds), polyurethanes are among the most versatile of plastics, ranging from rigid to elastic forms. Their major use is for foams, with properties varying from good flexibility to high rigidity. Thermoplastic polyurethanes that can be extruded as sheet and film of extreme toughness can also be made.

Glass-fibre-reinforced plastics

Polyesters of unsaturated acids. Certain esters can be polymerized to an infusible resin and are used on a very large scale in glass-fibre-reinforced plastics. Unsaturated acid (usually maleic acid in the form of its anhydride) is

first polymerized to a relatively short polymer chain by condensation with a dihydric alcohol such as propylene glycol, the chain length being determined by the relative quantities of the two ingredients. The resulting syrupy condensation polymer is then diluted with a monomer such as styrene and an initiator for addition polymerization added. This mixture is quite stable at room temperature over a long period. Frequently, a silicone compound is added to promote adhesion to glass fibres, and wax to protect the surface from oxygen inhibition of polymerization. Glass-fibre materials are impregnated with the syrup and polymerization is brought about by raising the temperature. Alternatively, the polymerization can be carried out at room temperature by addition of a polymerization accelerator to the syrup immediately before impregnation. After an induction period, which can be controlled, polymerization takes place, with rapid increase in temperature, to give a glass-fibre-reinforced cross-linked polymer, which is effectively a thermoset type of plastic and very resistant to heat. The properties of the resin are frequently varied by replacing part of the unsaturated maleic anhydride by anhydrides of saturated acids such as sebacic acid.

Silicones. Silicon, unlike carbon, does not form double bonds or long silicon chains. It does, however, form long chains with oxygen such as in siloxanes with hydrocarbon groups attached to the silicon; these result in a wide range of oils, greases, and rubbers. Produced through a series of reactions involving replacement of certain atoms in the chain, silicon resins, or silicones, can be used for high- and low-pressure lamination, with glass-fibre reinforcement and with mineral or short glass-fibre fillers, or for molding powders. The outstanding characteristic of these products is high dielectric strength (that is, they are good insulators at high voltages) with low dissipation over a wide temperature and humidity range. Silicones are not distorted by heat up to 400° C (752° F). They are also physiologically inert and therefore valuable for prostheses (artificial body parts).

Thermoplastic resins. Resins that soften upon heating include cellulose derivatives, addition polymers, and condensation polymers.

Cellulose derivatives. Cellulose is a naturally recurring chain of units, each of which carries three hydroxyl groups, which can be esterified by acids. Cellulose nitrate is formed by treating cellulose with a mixture of strong sulfuric and nitric acids; after a purification process, mixing with alcohol and camphor yields a plastic mass that is rolled into sheets. These are pressed into a block, sliced to the required thickness, and then dried and polished by hot pressing against polished metal sheets. The sheets made by slicing still contain alcohol and can be cut up and repressed in various ways to give patterned material or, if cut into dice, can be extruded as plain sheet. Nitrocellulose cannot be injection molded because of its inflammability. Otherwise it is tough and moisture resistant and is used for eyeglass frames and table tennis balls.

Cellulose acetates

If cellulose is pretreated with acetic acid, followed by more acid and its anhydride, acetate esters are formed, giving cellulose triacetate. Complete acetylation would produce an acetic acid content of 62.5 percent; in practice about 61 percent is usual, giving a moisture-resistant, dimensionally stable material with a crystalline structure, soluble in a limited range of solvents, preferably methylene chloride with 10 percent methanol. It cannot be worked into a plastic with plasticizers, but plasticizers can be incorporated in a 15–16 percent solution from which film is cast as a photographic base. This is the type of solution referred to as "dope" in Figure 13.

Further treatment of the triacetate in solution in the presence of sulfuric acid splits off some acetic acid giving diacetate, soluble in acetone, and compatible with a range of plasticizers that can be incorporated in a rugged type of mixer without solvent to yield molding powders especially suited to injection molding.

Tool handles are a common use of cellulose acetate. It can also be treated by the slicing process described above, but its largest use is as packaging film, made either by extrusion or by a casting process, which gives a superior product.

If butyric or propionic acids are included in the esterifying mixture, acetate butyrates and acetate propionates are formed. These require less plasticizer and are tougher and more moisture resistant. An important use is for automobile steering wheels.

Ethyl cellulose is formed from cellulose and ethyl chloride. Very tough and requiring little plasticization, it is used for athletic helmets, flashlight cases, and fire-extinguisher components. It is also compatible with a very wide range of plasticizers and waxes and, when mixed with wax, can be used for protective coatings.

Low-density and high-density polyethylene

Addition polymers. Polyethylene is made in two principal qualities: low density, polymerized at pressures from 630 to 3,150 kilograms per square centimetre (9,000 to 45,000 pounds per square inch); and high density, made with special catalysts at 18 to 35 kilograms per square centimetre (250 to 500 pounds per square inch).

Low-density polyethylene has a 20° C (68° F) lower softening point than high-density polyethylene, which is more crystalline and consequently stiffer. The rigidity characteristics and surface of high-density polyethylene are comparable with polystyrene. It resembles nylon in feel, has a bursting strength three times that of low-density polyethylene, and withstands repeated exposure to 120° C (250° F) so that it can be sterilized.

Low-density polyethylene is principally used in the form of film for bags of all sizes, with thickness ranging from 0.00635 to 0.178 millimetre (0.00025 to 0.007 inch), for agricultural mulch, building-site protection, lamination, all kinds of packaging, and electrical-cable insulation.

Because of its stiffness, high-density polyethylene is used for large drums, gasoline tanks, bottle crates, domestic appliances that can be injection molded to large sizes, and squeeze bottles. It can also be produced in the form of foam.

Polyethylene can be chlorinated either in solution in carbon tetrachloride or in suspension, both in the presence of a catalyst. Below 55–60 percent chlorine it is more stable than polyvinyl chloride and is compatible with many polymers, especially polyvinyl chloride, to which it gives increased impact strength.

The low pressure process can be used to copolymerize polyethylene with propylene and butylene to increase resistance to stress cracking.

Copolymerization with vinyl acetate at high pressures increases flexibility, resistance to stress cracking, and also sealability, which is especially valuable in the food industry.

Ionomers are ethylene polymers in which carboxyl groups are located along the ethylene chain. These acid groups are converted into salts of sodium, potassium, magnesium, or zinc; this stiffens and toughens the plastic without destroying its thermoplastic character. As a result, unplasticized sheet has drawing characteristics superior to those of any other resin.

Ionomers are highly transparent, particularly resistant to organic solvents, and maintain toughness at low temperatures. They are used in composite films with polyethylene and nylon because of their puncture resistance; they have strong natural adhesion to aluminum. Their main applications are in food packaging and as injection moldings in the shoe industry.

Polypropylene

Polypropylene is made by the polymerization of propylene. Stronger and more rigid than polyethylene, polypropylene has a harder surface, softens 30 to 40° C (54 to 72° F) higher, and is the lightest common plastic. It is more prone to oxidation than polyethylene unless appropriate stabilizers and antioxidants are added.

Injection molding affords its greatest outlet over a very wide range of applications including beer crates, luggage, and automotive panels and trims. An outstanding use is for integral hinges, because of its fatigue resistance. (An integral hinge is part of an object as it is manufactured, in contrast to most hinges, which are fastened to the object later.)

Polypropylene is being used in increasing quantities for fibres and for film. The latter is extruded and develops its greatest strength when oriented by stretching biaxially—that is, along two axes at right angles. Film clarity is obtained by minimizing the size of crystals and by various

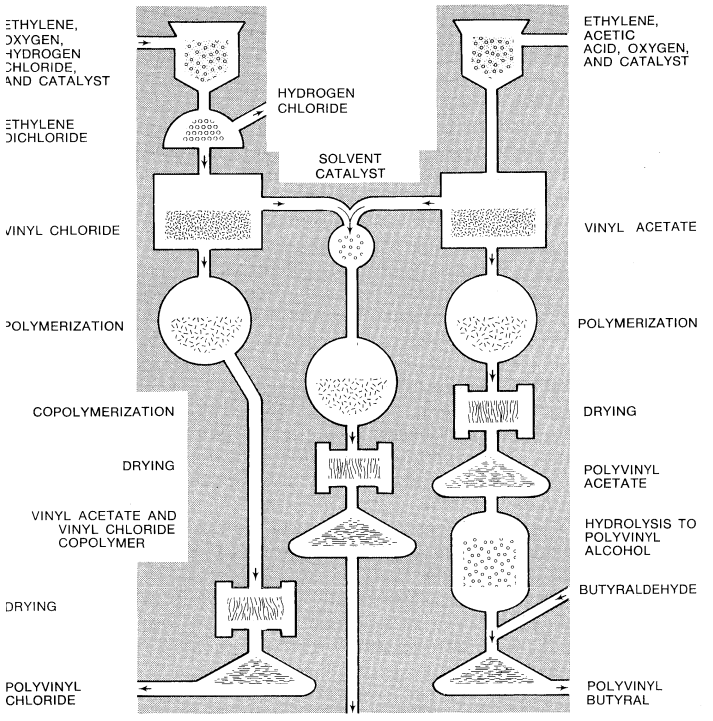


Figure 9: Principal steps in the production of vinyl resins.
Drawing by D. Meighan

additive agents. It is made in thicknesses down to 0.005 millimetre (0.0002 inch). Copolymers of propylene with minor amounts of ethylene are also used for making film.

Methylpentene polymer is the lightest plastic, has a high crystalline melting point of 240° C (464° F), retaining form up to 200° C (392° F), has a light transparency of 92 percent, and electrical properties similar to polytetrafluoroethylene (described later). Its impact strength is greater than that of polystyrene and polymethyl methacrylate. It is also resistant to alkalies, weak acids, and solvents (except chlorinated solvents) and is excellent for laboratory ware and medical equipment (because it can be sterilized) made by injection molding, and for food packaging and catering, since food can be cooked in methylpentene polymer containers in conventional ovens, providing a nonstick surface resistant to hot fat.

Vinyls

Vinyl chloride is made from ethylene and chlorine, though acetylene can also be used (see Figure 9). The polymer (called polyvinyl chloride) is mainly processed in a highly plasticized form, with varying degrees of flexibility, by calendering, extrusion, and molding, often from "dry blends," mixtures made below fusion temperature from polymer, plasticizer, and pigments. Plasticizers are chosen to maintain flexibility at low temperatures; the most commonly used plasticizer is dioctyl phthalate, but adipates may be used for special resistance to temperature change.

The range of applications of flexible polyvinyl chloride is enormous, covering clothing, flooring, wire insulation, home furnishings, piping, toys, and steel sheet covering. Extrusion, calendering, and molding of rigid polyvinyl chloride (*i.e.,* without plasticizer) is possible only with additions of stabilizers to prevent decomposition. Such stabilizers are mainly metallic (tin or cadmium) salts aided by epoxy resins, which also act as plasticizers, and lubricants such as stearates. Processing also may be aided by blending with such polymers as styrene acrylonitrile, polymethyl methacrylate, and chlorinated polyethylene.

Rigid polyvinyl chloride applications include bottle blowing, ducting in chemical plants, rainwater goods, piping for water and drainage mains, and light roofing. It has one of the highest combinations of stiffness and impact strength.

Plastisols and organosols

Plastisols and organosols are made by dispersing a very fine emulsion of polymerized resin in a plasticizer, usually in excess of the amount in flexible polyvinyl chloride (about 50 percent), together with stabilizers and gellants such as a calcium soap. Solvents and diluents are also used, and if they amount to more than 10 percent of the compound the product is known as an organosol.

The formulation can be difficult, as the flow characteristics can be varied greatly by choice by plasticizer and pigment (if used) and must be varied to suit the end use. Products are made by spreading, as on fabric, by dipping or by filling hot molds and pouring off excess material after gelling the outer layer at about 170° C (338° F). Plastisols can also be cast in molds; when maximum rigidity is required, diluents are used to permit lower plasticizer content. Plastisols can also be foamed.

Copolymers of vinyl chloride containing 5 to 40 percent vinyl acetate are made by the inclusion of vinyl acetate in the polymerization process; they have lower softening points and flow more easily than polyvinyl chloride at temperatures below decomposition points. They are thus suitable for making phonograph records, rigid clear sheeting, and molding powders. They are soluble in ketones, such as acetone, and certain esters, and they can be cast as film from solutions.

Copolymerization with vinylidene chloride, in quite small quantities, has the same effect on flow as vinyl-acetate and also increases the tensile strength of the plastic.

The chlorination of polyvinyl chloride, originally done in solution, is now carried out in a water–chloroform suspension. Extruded tube can be used in soil and waste systems and for central heating since its maximum service temperature is 100–105° C (212–221° F).

Vinylidene chloride can be made directly from ethylene and chlorine or by the further chlorination of vinyl chloride with subsequent removal of hydrochloric acid by alkali treatment. Copolymers of vinylidene chloride with minor proportions of vinyl chloride or acrylonitrile are used for making synthetic fibres, coating resins, and extruded thin films. The outstanding property of vinylidene chloride polymers is their low permeability for water vapour and gases. Coatings of the polymer are used on cellophane and other films to make them impervious to moisture.

Vinyl acetate is polymerized to a tough resin (polyvinyl acetate) by addition polymerization in solution to give an adhesive with a very high degree of tack (stickiness) and consequently of good adhesion; its main use, however, is as a basis of emulsion paints, in which case it is emulsion polymerized and not separated from the reaction mixture. Its use as a copolymer with ethylene and vinyl chloride has already been mentioned.

Polyvinyl alcohol is made by the saponification of polyvinyl acetate—*i.e.,* the removal of the acetic acid by alkali. In practice, a small amount of combined acetic acid is left in the molecule. The product is a water-soluble film-forming material used in dental molding and in the form of fibres for reinforcement (*e.g.,* of polyester), notably for injection-molded rail-tie retainers. If reacted with butyraldehyde, it gives polyvinyl butyral, an extremely tough but soft plastic used in sandwich-type safety glass.

Acrylics

Acrylic is a generic name applied to derivatives of acrylic acid, of which methyl methacrylate is the most important. The process of polymerization is directed toward control of chain length, which for cast resins is from 1,800 to 3,000 monomer units; for molding powders only about 500 units are required to facilitate flow into a mold. A small amount of plasticizer such as dibutyl phthalate may be added before bulk polymerization to assist in deep molding.

The outstanding property of polymethyl methacrylate is its light transmission of 92 percent and water-whiteness (transparency), a property maintained over years of exposure to weather. It is also resistant to ultraviolet radiation from fluorescent lamps and is readily fashioned by deep drawing (*e.g.,* for airplane canopies) and by injection molding (*e.g.,* for light fittings and automotive details, including signal lights).

A most successful use is for internally illuminated signs. Any colour can be introduced, and structural uses that take advantage of its transparency and toughness are numerous and varied. Its dimensional stability recommends it for many optical uses.

Sheets can be cast in cells made from two sheets of glass plates; the cells are filled with a viscous solution of par-

tially polymerized material. Other shapes are cast by the use of appropriate cell forms. Sheet is also made by an extrusion process, the emerging sheet being drawn through a calender with a highly polished chromium finish heated to 100° C (212° F). Acrylic dentures are made by stirring the powdered polymer into methyl methacrylate monomer containing a catalyst, and completing the polymerization in a mold formed from a dental impression.

Acrylics are also modified by copolymerization to improve impact strength, though they lose their extreme transparency and are usually pigmented, as for piano keys. An example is acrylic-modified polyvinyl chloride sheet, which is tougher than acrylonitrile–butadiene–styrene and polycarbonate and is suitable for corrosion-resistant parts, aircraft components, and materials-handling equipment.

Fluorocarbon resins

Polytetrafluoroethylene and fluorinated ethylene–propylene are the only resins composed wholly of fluorine and carbon. The monomers are polymerized in water either to a granular solid or a dispersion. In the polymers the fluorine atoms surround the carbon chain as a sheath, giving a chemically inert and relatively dense product with immensely strong carbon–fluorine bonds. Polytetrafluoroethylene cannot be molded in the ordinary way but must be preformed at a very high pressure, followed by a sintering (heating under pressure) process and, in certain cases, by further processing to obtain an accurate form. Fluorinated ethylene–propylene can, however, be injection molded and extruded as thin film. Both plastics have exceptional heat resistance and very low friction coefficients. This leads them to be used for such mechanical applications as seal rings, nonlubricated bearings, and packings.

For electrical applications, polytetrafluoroethylene and, to a slightly less degree, fluorinated ethylene–propylene resins are exceptional in that electrical and other properties are not adversely affected by long exposure to high temperatures. Polytetrafluoroethylene can be used in temperatures up to 260° C (500° F), and neither it nor fluorinated ethylene–propylene shows any embrittlement; they are also flexible at low temperatures and are standard insulation in aircraft and spacecraft. Their excellent electrical insulating properties lead to their use in very high frequency, ultrahigh-frequency, and other electronic applications. Polytetrafluoroethylene is also available in emulsion form to give adherent coatings to metals. Its complete inertness and low friction coefficient lead to its use in surgical prostheses, especially in joints.

Chlorotrifluoroethylene is a thermoplastic resin produced in varying formulations from combinations of fluorine-containing monomers. It has inherent flexibility, toughness, and clarity, differing in this last respect from the waxlike polytetrafluoroethylene. It is chemically resistant to all inorganic corrosive liquids and to most organic solvents, has zero moisture absorption, and a very wide range of service temperatures.

Its flexibility, combined with high insulating and thermal properties and facilitated by its extrusion properties, recommends it for wire and cable covering; its thermoplasticity, which can be augmented by plasticizer addition, or copolymerization with vinylidene fluoride, makes it an excellent material for injection molding of valve parts, closures for corrosive liquids, and for medical supplies.

Available with surfaces prepared for adhesion, polyvinyl fluoride transparent extruded film is unequalled for weatherability; it is used for external surfaces as a laminate. Polyvinylidene fluoride, a transparent injection-moldable plastic stronger than other fluoroplastics, is slightly less inert but particularly resistant to oxidizing agents found in the chemical industry.

Polystyrene

Styrene is polymerized to a clear transparent water-white resin called polystyrene on a very large scale by a mass-block or continuous-mass process or by suspension polymerization in a water medium. In the mass-block method, polymerization takes place in a system of cells alternately filled with monomer and a cooling medium; continuous polymerization takes place in a tower, with cooling zones and a heated base from which the polymer is extruded. Polystyrene is an ideal injection material, hard, brilliantly transparent, and a good insulator; but it softens in boiling water, is brittle, and burns readily.

In very thin moldings for food containers, the polymer becomes oriented by flow, which toughens it; extruded film can be vacuum formed to give containers suitable for fresh food because of their high gas permeability and good water-vapour transmission. Polystyrene is readily foamed by blowing agents, flame retardants being added to the polymer.

High-impact polystyrene is made by dissolving styrene–butadiene rubber in styrene and then polymerizing the styrene. Polymerization of the mixture may be carried out by bulk or suspension processes.

Acrylonitrile–butadiene–styrene polymers are essentially similar in physical structure but with acrylonitrile added in the polymerization to harden the polymer. Minute rubber particles diminish clarity but act as stress-relief centres, making acrylonitrile–butadiene–styrene in particular an outstanding polymer, especially for large shapes (luggage, car bodies, etc.). Special grades are also made for automotive parts (softening point above 100° C, or 212° F) and for chrome plating. It is regarded as a good engineering plastic, can be foamed, injection molded, blown, and alloyed with other plastics.

Formaldehyde polymers

Formaldehyde polymerizes readily to form paraformaldehyde. Under controlled conditions, a stable polymer of high molecular weight may be obtained. Polyformaldehyde is one of the strongest and stiffest thermoplastics, having a tensile strength of 700 kilograms per square centimetre (10,000 pounds per square inch). It can be injection molded and extruded. Exceptional fatigue endurance has led to its employment in automotive safety steering columns. Its impact strength and elastic recovery make it useful for snap-fit assemblies and springs, and its resistance to moisture and high abrasion and frictional resistance are advantageous in engineering uses generally. It also has good electrical properties over a wide temperature range.

Made mainly from trioxane (by condensation of formaldehyde) and ethylene oxide, acetal copolymers have properties resembling those of polyformaldehyde but in addition are notable for their stiffness at high temperatures, making them suitable for carburetor and automobile parts and for plumbing items. When glass-fibre-filled, they have exceptional resistance to creep (slow deformation over time when placed under stress).

A crystalline polymer, chlorinated polyether has a very low volume change from a melt to a solid state, permitting injection molding to very close tolerances. It has very high corrosion resistance to all reagents except oxidizing agents and is a valuable engineering plastic for valves, threaded socket connections, and precision machined gears, with stability up to at least 120° C (248° F). It can also be coated on metal parts by a number of coating techniques to give corrosion resistance.

Condensation polymers. The formation of polyamides has been briefly described earlier. By far the greatest production is of nylon-6,6, made by heating the amine and acid in an autoclave (an airtight vessel for carrying out chemical reactions under pressure) from which it is extruded on a chilled drum and cut into chips. It has a high melting point, strength, stiffness, toughness, and chemical resistance. After careful drying it can be extruded and injection molded.

Nylon plastics

These properties class it as an engineering plastic (a plastic used as a substitute for metal) and lead to a very wide range of applications such as gear wheels, oil seals, bearings, and parts for textile machinery. Extruded film is also used for boil-in-bag packaging. A large number of different molding powders is made by variation of molecular weight and copolymerization of various amide-forming constituents.

In addition, a number of different nylons are made by varying the molecular length of the diamines and the dibasic acids: nylon-6,10 using sebacic acid (10 carbon atoms); nylon-11, from an acid derived from castor oil; and nylon-12, made by a multistage process from butadiene. Terephthalic acid with methyl hexamethylenediamine gives a transparent polymer instead of the normal waxy product. The primary object of these variations is to decrease the moisture absorption.

Nylon-6 differs from these nylons in that it polymerizes

directly, without elimination of water, from caprolactam. The polymer can be cast in place and is specially suited to irregular forms and thick sections because the low viscosity of the monomer prevents voids and polymerization is independent of part size, producing items ranging from 227-kilogram (500-pound) blocks to fuel tanks. The polymer can be extruded as easily sealable film and used for bake-in-bag bread and roast-in-bag meat. It is also used for packaging as a laminate with polyethylene.

Other variations are the use of amines with a ring structure, particularly the aromatic nylons, which give polymers with softening points above 300° C (572° F). A paper is made from chopped fibres from the phenylenediamine nylon that has electrical properties similar to mica.

Polyethylene terephthalate

Polyethylene terephthalate is a polyester made from glycol and dimethyl terephthalate by the process known as ester interchange. During the reaction methanol is produced instead of the water displaced during ordinary esterification. It is extruded as a transparent film in amorphous form and is immediately chilled. If it is stretched in two directions at right angles (biaxially stretched) at a temperature above 80° C (176° F), it is still mainly amorphous and can be used for shrink wrapping foodstuffs (a shrink wrapping shrinks after it is put around the object) and medical accessories by momentarily reheating it above the stretch temperature.

If, however, the film, biaxially stretched to three or four times its original dimensions, is held at a higher temperature, it crystallizes and develops very high tensile strength (1,760 kilograms per square centimetre, or 25,000 pounds per square inch) and dimensional stability.

Its main application is in food packaging, where it is used for meat packaging and for heat-and-serve packaging of convenience foods. For these applications it is frequently laminated with polyethylene. Its great strength and high dielectric strength make it ideal for insulation in electric motors, printed circuits, and cable insulation. Its dimensional stability and strength lead to its use for magnetic tape and increasingly for photographic film base. A special variety of the polymer has been developed that will mold to give the hardest plastic known.

Polycarbonates

Polycarbonates are long-chain linear polyesters of carbonic acid and dihydric phenols, such as bisphenol A. This product is a faintly amber transparent plastic of good toughness, tensile strength, ductility, and dimensional stability maintained over a temperature range of 16° C to 150° C (60° F to 300° F).

It can be injection molded, blow molded, and extruded and is an ideal engineering plastic with good electrical insulating properties, finding applications in portable tool housings, safety helmets, computer parts, and vandal-proof windows and light globes. A number of special grades are made, including a special type to meet specific flammability requirements. Glass-reinforced polycarbonate can carry continuous loads up to 280 kilograms per square centimetre (4,000 pounds per square inch) at boiling-water temperature.

Polyimides

Polyimides are made by a two-step reaction of dianhydride (tetrabasic acid with two molecules of water removed from each molecule of tetrabasic acid by chemical reaction) with an aromatic diamine giving first a polyamic acid that is converted by heat or catalysts to high molecular weight linear polyimides, available as film, laminating resins, adhesives, and molding compounds.

The polyamic acid is soluble; films can be cast from it and it can be used for laminates and adhesives. Polyimides are the best organic materials available for high-temperature insulators, capacitors, and bearings, and have special applications in space technology. Exposure at 275° C (527° F) for one year causes only a 1 percent elongation. Polyimides can also be produced as foamable powders for aerospace structural applications.

If one amino group in *o*-phenylenediamine is converted to an amide group by formic acid, an intermediate is formed that on loss of water becomes benzimidazole. This reaction can be conducted with a wide range of reactants, producing resins (polybenzimidazoles) that can be used as high-temperature adhesives for laminates in the aerospace industry and compounded with tiny bubbles of silica (microballoons) and alumina fibres to provide insulation for re-entry vessels, fuel cells, and nuclear reactors. If the fillers are phenolic microballoons and carbon fibres, very high temperatures develop a carbonaceous char with excellent structural and ablative properties.

Aromatic polysulfones are transparent linear polymers made from the sodium salt of bisphenol A and *p, p'*-dichlorodiphenylsulfone. Light amber in colour and exceedingly resistant to oxidation, these plastics have extraordinary dimensional stability, a stress of 105 kilograms per square centimetre (1,500 pounds per square inch) for 100 hours at 150° C (300° F), producing barely visible deformation. These are first-class engineering plastics, used for underhood automotive parts, housings for power tools, and for dishwashers. Other phenols and sulfones are used for special high-heat-resistant materials.

Polyphenylene oxide is a linear polymer formed by oxidation of 2,6-xylenol by oxygen in the presence of a catalyst. It has very accurate dimensional tolerances at high temperatures and high chemical resistance. It is suitable for clock parts and surgical instruments and is in general a very good engineering plastic because of its stability, heat resistance, and low moisture take-up. Many variations in phenols are possible.

Though the starting materials for polyhydroxyethers, (phenoxy resins) are the same as for epoxy resins, phenoxy resins are of high molecular weight. An excess of epichlorohydrin is avoided to ensure thermal stability. Their properties can be varied by the use of different phenols, but the polymer from bisphenol A is commonly used. For coatings the hydroxyl group is usually esterified, and with drying oils thermosetting coatings can be made; phenoxy coatings are excellent primers because of high adhesion to metals and overall compatibility with topcoats of other plastics.

As molding powders they give tough, lightweight, clear moldings, easily printed and with good resistance to water and other liquids making them specially suitable for blown bottles. They have high abrasion and chemical resistance and can be hot gas welded. A special use is for jet-fuel tanks.

Steam at 950° C (1,740° F) converts *p*-xylene to an active intermediate that can be polymerized on a cooled surface to form poly-*p*-xylylene. It is used only for very thin coatings as in miniaturized capacitors and for encapsulating such materials as lithium.

Coumarone–indene resins

Coumarone–indene resins, readily fusible, low-molecular-weight polymers, are not strictly plastics. They are made by polymerization of coumarone (benzofuran) and indene obtained from crude petroleum or from gas tar and coal tar resulting from the high-temperature carbonization of coal. The coumarone content is often less than 10 percent, and the polymer may contain styrene and methyl styrene, softening points varying with the composition up to 150° C (300° F). The catalysts are cationic (positively charged), and batch operation is traditional. Coatings of the resin are water and alkali resistant and are particularly useful vehicles in aluminum paint. They are also used in conjunction with chlorinated rubber and modified phenolics. They also greatly assist in the blending of fillers in rubber for floor coverings.

Other types of resins. *Oil-soluble or modified resins.* Alkyds used as coatings are polymers in which the spare hydroxyl, resulting from the esterification of a triol (usually glycerol) by a dibasic acid, is esterified by a long-chain (fatty) acid, which may contain double bonds. Since oils such as linseed are already glycerol esters, reaction is brought about by splitting off one or two acid chains—*i.e.,* reconstituting glycerol hydroxyl groups and re-esterifying these with more triol and a dibasic acid, usually phthalic acid, giving a complex polymer in which the acid chains polymerize among themselves and may retain double bonds which oxidize in air. As such, these are good drying oils.

Physical properties, including solubility, depend on composition, but their outstanding characteristic is compatibility with other resins.

Simple phenol– or cresol–aldehyde resins are insoluble in drying oils, but if resols are heated with an excess of an unsaturated organic acid such as rosin, they become

soluble with removal of water and can be dispersed in ester gum (rosin glyceryl ester). By cooking with unsaturated oils such as tung and linseed, the rosin-modified resols disperse with some interesterification to produce air-drying varnishes.

The oil may also be incorporated by reacting phenol and formaldehyde in its presence or by reacting the phenol with the oil, with subsequent condensation with formaldehyde. Substituted phenols are sometimes used because they are more readily soluble in vegetable oils.

Plastics from other natural products. Casein plastic is made by reacting casein from milk with formaldehyde. It is produced in a variety of opaque colours, including pearl; it softens in hot water but takes a very high polish. Its use is greatly diminishing, but buttons and buckles are still made, partly because short runs, advantageous in fashion trades, are practicable.

Uses of lignin

Lignin is a mixture of complex thermoplastic phenolic polymers present in wood and straw. Its chief use is for hard fibreboards, made by compressing high-lignin-content material at 700 kilograms per square centimetre (10,000 pounds per square inch) at 250° C (480° F), with or without previous treatment with phenol and formaldehyde. The lignin content is increased by disintegrating the cellulosic material and hydrolyzing part of it by treatment with steam and water at high pressure, the remaining fibre acting as reinforcement.

Synthetic resins for special applications. These applications include resin adhesives, paper, and textiles.

Apart from nitrocellulose and polyvinyl acetate resin, adhesives are thermosetting.

Resins for bonding

Phenol–formaldehyde and urea–formaldehyde resins are used chiefly for bonding wood, especially plywood. Phenol–formaldehyde joints are resistant to boiling water and by addition of resorcinol acquire greater strength, which, for metal to metal joints, achieves a maximum with the addition of polyvinyl formal. Epoxy resins also adhere well to metal.

Urea–formaldehyde resins are cheap, show no glue line, and are used extensively in furniture. Polyurethane joints formed in place with diisocyanates are chiefly used to cement rubbers to metals or textiles. Polysulfones, polyimides, and polybenzimidazoles are used as high-temperature adhesives.

By the addition of small amounts of amino-resins to paper pulp, wet strength, bursting strength, rub resistance, and dry tensile strength are greatly improved. Urea–formaldehyde resin is also used to produce wallboard. Paper is rendered waterproof by coating with coumarone–indene resin and still more effectively by lamination with polythene film at very high speeds.

A number of polymers are used for external finishes, either in organic solvent solution or as emulsion, the most important being urea- and melamine–formaldehyde resins, and complicated derivatives thereof, acrylics and alkyds.

Amino resins are used for washfast, glazed, shrink-resistant chintz and for embossed effects. Acrylics stiffen without destroying flexibility and improve the abrasion resistance of both cotton and rayon, while polyethylene is applied in film form for permanent glazing. Butadiene–acrylonitrile improves the wear of cotton goods, and polyvinyl acetate is used extensively on linings that are often calendered to give a silk effect. The most interesting use of alkyds is that of a thermosetting castor-oil-modified glyptal that stiffens, is a fixing agent for pigments, and preserves mechanical finishes.

Plastics are also applied for crease resistance, especially of rayon and cotton. Urea–formaldehyde and melamine derivatives are chiefly used in solution, catalyzed by ammonium salts of strong acids; a variation is the inclusion of octadecyloxymethylpyridinium chloride, which confers water repellence, and of urea–formaldehyde resin with dichloropropanol, which produces an epoxy finish and is used in drip-dry materials.

PLASTICS MANUFACTURE AND FABRICATION

Natural raw materials. *Cellulose.* Originally the only source of cellulose for nitration, acetylation, and other processing was cotton linters, the short fibres left after ginning, which removes the long staple fibres. After cleaning and bleaching the linters are ready for acetylation and other processing.

Undegraded cellulose is a crystalline material with chains built up of about 3,500 repeating units of glucose anhydride, but this number falls to about 500 on acetylation, the fall being limited by process control as far as possible since molecular weight and physical properties are closely connected. Linters are now being replaced by cellulose derived from wood pulp by alkali treatment.

Coal and petroleum products. The products of the carbonization of coal are tar and light oils (or benzol), in proportions and quantities dependent upon the type of coal and the process employed. The amount of coal carbonized to make marsh gas (methane gas) is diminishing rapidly because of natural gas and the use of petroleum products, so that coke ovens are the largest source of by-products, with substantial additions from smokeless fuel plants.

Tar is distilled, and processing of the fractions yields phenol, cresols, and naphthalene, the first two used in formaldehyde resins and the latter for phthalic anhydride by oxidation.

Distillation of the light oils after alkali washing yields benzene, toluene, and xylene, and a residue of "heavy" naphtha containing coumarone and indene, which are separated by polymerization; coal tar is the only source of these polymers.

Coal by-products thus contain most of the raw materials for synthetic plastics, especially if acetylene from calcium carbide is included, but the vast quantities now required for plastics can only be provided by recourse to the petrochemical industry.

Raw materials from oil

Ethylene and propylene are produced on an enormous scale by cracking naphtha in tubes heated by steam to 750–900° C (1,380–1,650° F) or by direct heat transfer. Ethylene is also formed by catalytic cracking at 700° C (1,290° F) of a special fraction that also gives a high content of aromatic hydrocarbons, benzene, toluene, and xylenes, as well as a residue from which napthalene is recovered.

Aromatics are also obtained by the platforming process in which feedstock is passed over platinum-coated silica–alumina, but naphthalene cannot be obtained directly from this product because most of it is present as an alkyl derivative and must be dealkylated by a special process that may also be applied to toluene to produce benzene.

Xylenes, which are also present, are separated by distillation and low-temperature crystallization, the *para*-xylene being required for oxidation to terephthalic acid; vapour-phase catalysis of the other xylenes produces more *para*-xylene.

Ethylene is now used to make vinyl chloride by way of ethylene dichloride; it combines with benzene to make styrene via ethylbenzene and yields glycol via ethylene oxide. Propylene is converted to glycerol through oxidation to acrolein; further treatment with hydrogen peroxide and chlorination of glycerol gives epichlorohydrin for epoxy resins. Ammoxidation, in which propylene reacts with air and ammonia, yields acrylonitrile. Acrolein is also an intermediate in the route to acrylic acid. The oxidation of isobutylene yields methyl methacrylates.

There are a number of ways of converting benzene to phenol—by sulfonation followed by fusion with sodium hydroxide, or chlorination and hydrolysis; and by the catalytic reaction of propylene with benzene, giving cumene, which is then oxidized with air to give cumene hydroperoxide, this product in turn being converted by strong acids to phenol and acetone. The mixture of phenol and acetone can be condensed directly to bisphenol A, widely used in epoxy resins and polycarbonates. Oxidation of benzene by air yields maleic acid, and treatment with methanol vapour gives cresols.

Methane is converted to acetylene, which reacts with acetic acid to form vinyl acetate; vinyl acetate can also be formed directly from ethylene and acetic acid in a reaction catalyzed by copper and palladium salts. Another petroleum product is cyclohexane, from which both adipic acid and hexamethylenediamine are made for polymerization to nylon-6,6.

Chemicals and additives. *Plasticizers.* Plasticizers, usually high-boiling, low-volatility esters, are incorporated with polymers by heating and mixing to increase flexibility by inserting themselves between the polymer chains and thus reducing the attractions between chains. They are essential in cellulose plastics and include triphenyl phosphate, a common ingredient to reduce flammability. The esters are usually diethyl and dimethoxy glycol phthalates but in polyvinyl chloride, which accounts for the largest proportion of plasticizer used, phthalate esters of much higher molecular weight are employed because the plasticizer proportion is higher and low volatility is imperative. It is also important that flexibility be maintained at low temperatures. The compound di-2-ethylhexyl phthalate (DOP) is much used, and for maximum low-temperature flexibility the adipate as well as the azelate are employed. These three plasticizers have the advantage of being safe in vinyl food-wrapping film. Of the phosphates, only octyl diphenyl phosphate may be used for this purpose.

Increasing flexibility

Some materials are classed as extenders rather than plasticizers and cannot be used alone, nor are esters universally acceptable; nylon, for example, requires a glycol or an aromatic sulfonamide.

Another class of plasticizer, either a complete polyester made from a glycol, or an epoxy compound, usually of soya oil, is known as a "permanent" plasticizer because of its nonvolatility; the epoxy types are used with other plasticizers in vinyl compounds and also function as stabilizers.

Stabilizers and antioxidants. Polymers are subject, in varying degrees, to degradation by heat, light, and oxidation and can be protected by substances that only arrest but do not entirely prevent these processes.

Polyvinyl chloride in particular requires stabilization, especially with the development of rigid products, since the manipulation temperatures are high enough to initiate decomposition, but polyolefins and polystyrene are also improved by suitable additives. In polyvinyl chloride the main problem is emission of hydrochloric acid; this can be reduced by adding lead salts or soaps such as lead stearate. Though these are inexpensive, they cause opacity and are toxic; they are widely used for insulation compounds, however. Where a degree of transparency is required with freedom from stain, barium–cadmium salts, with or without zinc, are employed; they have the advantage that they can help to eliminate "plate out," the deposition of constituents on the metal of the processing plant.

With silicate and asbestos fillers, epoxy resins and phosphites are used, but for rigid transparent polyvinyl chloride, processed at high temperature, a different class of stabilizer, that of alkyl compounds of tin, is used.

Dibutyltin derivatives

After 1950 a series of dibutyltin derivatives, such as the dilaurate and maleate, were recognized as excellent though expensive stabilizers for high-clarity polyvinyl chloride, a property becoming more and more important with the increasing use of blown polyvinyl chloride bottles. These stabilizers are now being replaced by dioctyltin derivatives containing maleate groups or mercaptoacetates, which are far less toxic than the dibutyltin compounds.

For plastisols, barium, cadmium, and zinc compounds, together with organic additives, are preferred. Nontoxic stabilizers also include calcium–zinc compounds with phosphates and epoxies. Formulations are widely varied according to processing conditions, resin characteristics, and end uses.

Catalyst residues and metal traces from the processing plant promote degradation, which is diminished by antioxidants and ultraviolet absorbers. Organic phosphites are used to deal with metallic impurities and traces of heavy metals and also to decompose any trace of peroxide.

Antioxidants fall into five classes: phenol derivatives, amines, esters, organic phosphites, and miscellaneous—the choice depending on the polymer to be protected and the end use.

Low-density polyethylene is an important example of a polymer that must be protected by antioxidants. It is characterized by side chains, and each junction is vulnerable to oxidation, which is dealt with by phenols, alone or in conjunction with sulfur-containing esters or phosphites. Oxidation caused by light calls for ultraviolet absorbers, the most satisfactory of which is carbon black, the antioxidants in this case being phenols and naphthols containing sulfur.

Because of its structure, polypropylene is extremely susceptible to oxidation, a reaction that becomes important because the substance is handled at high temperatures. Antioxidants similar to those used for low-density polyethylene are effective, especially phenols and sulfur compounds, as distearyl thiodipropionate, used together.

Polystyrene is relatively stable to oxidation but sensitive to light degradation. High-impact polystyrene, however, because of its rubber content, requires the use of phenols. Most other polymers are less subject to oxidative degradation resulting from exposure to oxygen, but phosphite esters are often used, especially with polyurethanes, which are likely to discolour.

As already indicated, ultraviolet light combines with oxidation to promote degradation, and combinations of absorbers and antioxidants are common, the ultraviolet absorber being increasingly important as plastics are used in outdoor applications.

Classes of absorbers

The two classes of absorbers are screening agents and excited-state quenchers. Excited-state quenchers stabilize the polymer by removing energy from the molecule that has already absorbed ultraviolet light. Screening agents are the most widely used, absorbing most of the incident ultraviolet light and converting it to heat. The most effective of these are the benzophenones, especially their ethers and benzotriazoles. Ultraviolet absorbers also work by converting shortwaves into harmless longer light waves through fluorescence. Protection by fluorescent dyes is particularly applicable to films. Of the excited-state quenchers, nickel compounds are the most effective, especially organo-nickel complexes.

Organic peroxides. Organic peroxides and hydroperoxides are derivatives of hydrogen peroxide and split readily into the free radicals (groups of atoms that act as a unit in a chemical reaction) required in addition polymerization, over a temperature range determined by composition. Room temperature splitting can be brought about by activators, such as cobaltous ion or tertiary amine. Industry offers about 45 different peroxides.

Colorants. Colorants are inorganic or mineral pigments, organic pigments (either toners, insoluble organic salts, or metal complexes of dyes), or are lakes, dyes deposited on alumina.

Uniform dispersion and standardization of colour are usually accomplished by predispersion in a resin at high temperatures or, especially in the case of epoxy resins, polyesters, and urethane systems by dispersion in the plasticizer, monomer, or solvent.

Dyes must be used for transparent plastics, either dissolved in solvent or in such fine dispersions, as in the case of basic dyes, that no particles are visible.

Flame retardants. Where nonflammability is especially important, all polymers except rigid polyvinyl chloride and fluorocarbons require flame-retardant additives, even though some, such as nylons or polycarbonates, are self-extinguishing. The most important flame retardants are bromine, chlorine, antimony, and phosphorus compounds. Antimony compounds can be used only with opaque materials, however.

Additives vary with the polymer; various phosphates are used with plasticized polyvinyl chloride, but with polyurethanes the phosphates must be brominated, and the same applies to polystyrene foam, where highly brominated phosphates or hydrocarbons are incorporated in the beads as supplied for molding. An alternative method is to build bromine-containing units into the polymer chains, as can be done in the case of polyesters, polyurethane foam, and epoxy resins.

Fillers and reinforcements. These fall into five main categories: mineral and synthetic inorganic powders, carbon blacks, cellulosics, metal powders, and microspheres.

Reinforcements include cotton and asbestos flocks; glass fibres, chopped or in the form of rovings, mats, or monofilaments; carbon fibres; and mineral whiskers.

A great variety of fillers especially suitable for different polymers and end uses is available—*e.g.,* wood flour, first

used for phenolics, which is cheap and reduces mold shrinkage; asbestos flock (tiny fibres of asbestos), which greatly increases impact strength; various fillers that improve polyurethanes; and aluminum, magnesium, and titanium oxides, which give added stiffness to composites.

Fibre reinforcement

Cotton flock is essential in urea–formaldehyde resin moldings, and many synthetic fibres are used with different polymers; the great bulk of reinforcement, however, is glass fibre of less than 25 microns (10 millionths of an inch) in diameter and available as glass cloth; the cloth can be chopped for incorporation in moldings or used as mats or as filament for a thin surface mat or for winding. Carbon fibres made by heating acrylonitrile fibres have enormous tensile strength, and plastics bonded with them can replace metals. Sapphire fibres and single-crystal whiskers also provide powerful reinforcement.

Basic processing techniques. The chief plastics-processing techniques are extrusion, blow molding, film and sheet extrusion, calendering, thermoset molding, injection molding, rotational molding, thermoforming, laminating, casting, film casting, and foaming.

Extrusion. A major technique of the plastics industry, extrusion consists essentially of the melting and compression of plastic granules by the rotation of a screw conveyor in a long barrel to which heat or cold can be applied. The screw drives the plastic through a nozzle (see Figure 10).

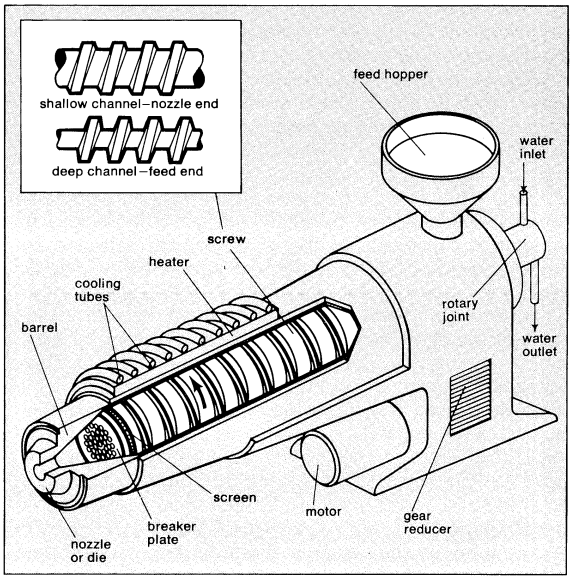


Figure 10: Extrusion-molding equipment. The screw has a shallow channel at the nozzle end and a deep channel at the feed end, as shown in inset.

The technique is used on a large scale for the continuous production from thermoplastic materials of pipe, sections of various forms, sheeting and films, wire covering for insulation, plastic feed for blow-molding plants (see *Blow molding*), and preplasticizing systems for injection molding. Screws may vary considerably in design, but the most widely used is the constant pitch screw divided into feed, transition, and metering zones. The feed zone preheats the plastic granules in a deep screw channel of constant cross section and feeds them to the transition zone, which is of decreasing depth so as to compress and generally homogenize the melting plastic. This in turn discharges into the metering zone, which is of a constant depth less than that of the feed zone; it serves to complete the melting and mixing operation and to generate maximum pressure for pumping the molten plastic through the shaping nozzle. Additional homogenization may be obtained by including a mixing device; *e.g.*, a fluted cylindrical section in the screw that shears the molten plastic intensively as it passes through fine channels between the flutes and the barrel wall.

The pressure of the plastic melt depends on the resistance of the nozzle but may be regulated by means of a throttle placed near the screw discharge. This may be a simple breaker plate with a screen pack, which also helps to homogenize the melt and to filter extraneous matter, or a streamlined adjustable valve that can be regulated while extrusion is in progress. Devices for automatic changing of screen packs without stopping extrusion are available.

Modern extruders may have a longer and more complex screw than the basic one described above. Such a screw is typically in two stages, each of which has a carefully selected combination of zones. Between the stages a decompression of the melt makes possible the incorporation of a vent hole in the barrel wall at this point for the removal of volatile materials and trapped air from the melt. Vacuum may be applied to this vent to increase the effect, so as to give a product of enhanced quality at high production rates.

Variations in screw design

The screws vary in design to give optimum results for different plastics. An important specification is the ratio of length to diameter, generally about 25 to 1 although 35 to 1 or more is not uncommon for two-stage screws. Special alloys usually are used both for barrel and screw, the latter being bored for heating or cooling. Direct current motor drives are preferred for accuracy of control; the heating or cooling zones on the barrel, which may be built in, are controlled precisely, usually by electronic methods.

Normal nozzle diameters range up to 20 centimetres (eight inches) but can be larger. For extruding pipe, a device called a mandrel, centred in the nozzle, determines the wall thickness.

Blow molding. Extrusion is used to push a molten tube (called a parison) into a bottle-shaped mold. Compressed air then forces the parison against the cold walls of the mold, hence blow molding. Molds are generally side fed, with wall thickness controlled by a tapered mandrel (core) or variable-orifice die. Continuous extrusion is possible by use of multiple blow molds.

Film and sheet extrusion. Thin polyethylene film is made in large quantities by extruding a tube through an annular nozzle behind which is a device through which air is introduced into the emerging tube. The air expands the tube to two or three times its original diameter. The plastic is then pressed between rollers over which the film passes as a double layer; subsequent slitting gives two flat films.

Thick-sheet extrusion

Thicker sheet is extruded through special nozzles (see Figure 11). Acrylonitrile–butadiene–styrene sheet six millimetres (0.24 inch) thick is made in this way. Plastics that melt, such as nylon and polyethylene terephthalate, are deposited on a roll, cooled, and stretched before reeling. Controlled stretching at elevated temperatures in one or more directions can lead to improved clarity and strength (oriented films).

Calendering. Soft flexible polyvinyl chloride sheet, plain or embossed, is made in large quantities by feeding hot premix between rollers individually driven and heated to about 150–180° C (300–355° F), with an output ranging up to 4,500 kilograms (10,000 pounds) per hour. When desired, the sheet can be embossed, laminated to fabric, or laminated to another plastic sheet (to make tiles); this may be done immediately as the sheet emerges from the rollers.

Thermoset molding. Thermosetting resins are usually molded in hydraulic presses, using steel molds cored for heating or cooling. Pressures up to 210 kilograms per square centimetre (3,000 pounds per square inch) are required for cold powder or pellets, but these feeds are usually preheated by hot air or by infrared or radio frequency

From M. Smith, *Manufacture of Plastics*; Chapman & Hall

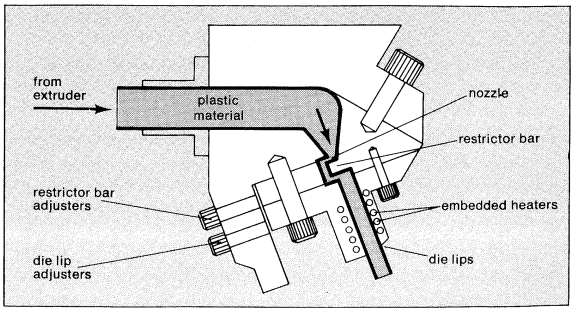


Figure 11: Apparatus for forming thick sheets by extruding plastic through a special nozzle.

dielectric heating and pressures are lower. Preheating reduces curing time. Radio-frequency heating is particularly suitable for semi-automated molding.

Injection molding. Injection molding differs from compression molding in that the plastic material is rendered fluid in a separate chamber outside the mold, then is forced into the mold by external pressure. The mold is cooled and split, the two halves being locked by pressure during molding and opening automatically when the mold cavity has been filled by a shot of the liquefied plastic. The cooled article is then ejected from the mold and the process repeated.

In its simplest form (see Figure 12A), the machine consists of a heated barrel with a hydraulically operated ram, an opening in the rear of the barrel being fitted with a hopper into which measured amounts of plastic pellets are fed. Falling in front of the plunger, the pellets are forced under high pressure into the heating section of the barrel, pushing out an equivalent volume of molten plastic into the mold. A torpedo-shaped distributor is fixed between the end of the ram and the nozzle feeding the mold.

'he ecipro- ating crew feed

Of the various methods used to speed preplasticization, the most common involves a reciprocating screw feed that builds up a shot of molten plastic behind the nozzle, moving backward against the injection ram pressure until an electrical device stops rotation and the screw operates as a ram to force the plastic into the mold (Figure 12B).

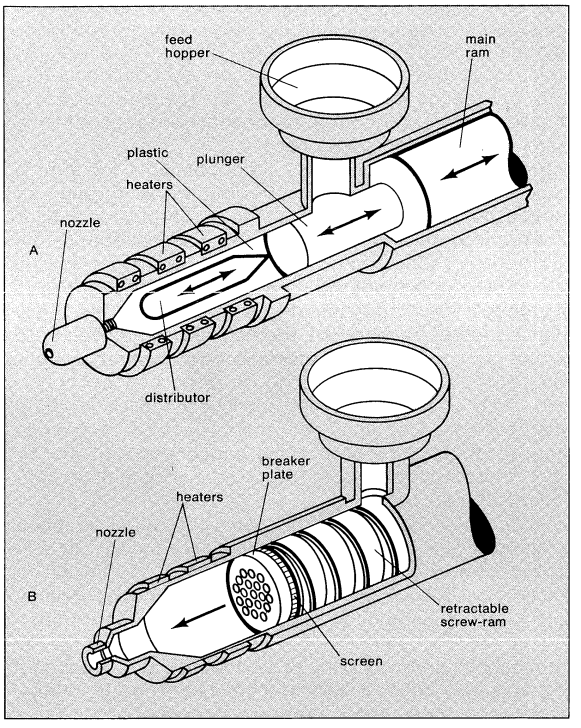


Figure 12: Thermoplastics injection machine with (A) hydraulically operated ram and (B) retractable screw-ram (see text).

The two halves of the mold are clamped at a pressure that resists that exerted by the plastic entering the mold. This is usually achieved by a hydraulic ram or by a hydraulically operated toggle mechanism; the nozzle may be fitted with a device to prevent premature leakage of plastic at the nozzle and permit high pressure buildup before the shot.

Rotational molding. Plastics structures with heavy walls and complex shapes can be produced by rotational molding, in which a measured amount of plastic in powder or liquid form is placed in the cold mold, which is then rotated about both its vertical and its horizontal axis to distribute a layer of plastic evenly over the whole of the interior surface. While rotating, the mold is passed into a heated chamber where fusing or curing takes place; quick cooling solidifies the plastic, producing an accurate replica of the internal mold surface.

Heating is usually by hot air, up to 450° C (840° F); rotation speeds are infinitely variable up to about 12 revolutions per minute on the major axis and 40 revolutions per minute on the minor; the molds are inexpensive, usually made of cast aluminum. For extremely large cavities, sheet metal can be used, especially for prototypes. The ideal mold material has a high heat conductivity; beryllium copper is sometimes used. Polyethylene, polypropylene, and polyvinyl chloride are the plastics most widely used. Large commercial containers, special carrying boxes, and automotive headrests and armrests are made by rotational molding.

Thermoforming. A thermoplastic sheet, heated to its softening point, can be forced against the contours of a mold by air pressure, forming a desired shape after cooling. Many methods differing in detail are employed, but five principal techniques exist: simple vacuum forming, drape forming, and three modifications to permit deep drawing with maintenance of even thickness of walls and bottoms of the shapes. In simple vacuum forming, the plastic sheet is clamped over an airtight box, from which air is exhausted, rapidly drawing the heated plastic down on a "former" within the box, thus reproducing its shape. The sheet is heated above the box by electric heaters. In drape forming, the electric heater plate above the sheet is removed and a male former is driven up into the clamped soft sheet. Vacuum is employed to pull the sheet tight to the former.

Deep drawing

Both processes are modified to permit deep drawing. One technique involves plug-assisted drawing, in which a plug is forced down into the sheet before the application of vacuum. In drape forming, a controlled bubble is first blown upward in the sheet into which the former then enters. Another technique resembles plug-assisted forming except that the sheet is forced into the female former by compressed air.

Laminating. At first, only solutions of phenol–formaldehyde and urea–formaldehyde resins were used to impregnate paper, cotton fabric, absorbent forms of asbestos, and other fabrics. Later, melamine–formaldehyde resins were added. Material from a roll of fabric was passed through a bath and then a drying chamber; cut up sheets of the dried material were stacked in a press, then heated and cured at a pressure of more than 70 kilograms per square centimetre (1,000 pounds per square inch); hence they were known as high-pressure laminates. Today, the development of polyester resins and epoxy resins that cure without giving off any by-products permits the use of low pressures, sufficient only to hold the reinforced layers in sheet form; these are known as low-pressure, or contact, laminates.

A variation of the sheeting method involves addition of the resin or its components to paper pulp, which is then sheeted according to standard paper methods. The most important of these are polyester resin laminates of glass-cloth and epoxy-resin laminates with carbon fibres. "Prepregs" and "postforming" laminates are impregnated uncured sheets that can be molded in metal molds.

Where tubes are required, the impregnated material is wrapped on suitable metal cores, revolved between heated rolls, removed, and baked to complete polymerization.

Casting. Polyesters, polyurethanes, silicones, and epoxy resins can be obtained in liquid form without solvents and can be formed by mixing in a catalyst and pouring the liquid into an inexpensive mold. Under suitable conditions, the mixture hardens into a solid shape. Electrical potting (sealing an electric component by placing it in a container and filling the container with the plastic compound) and industrial patterns are typical cast products. Specimens such as coins or butterflies can be embedded in attractive form in polymethyl methacrylate by immersing them in partially polymerized monomer and completing the polymerization very slowly.

Typical cast products

Film casting. Film casting produces continuous rolls of thin film by spreading a polymer solution on a moving surface passing through a heated tunnel (see Figure 13). The solidified film is stripped as it returns to its casting point, and residual solvent is removed by passage through heated multiroller cabinets, in the course of which, if necessary, coatings can be applied. The moving surface may be a wheel (5.5 metres, or 18 feet, in diameter) with

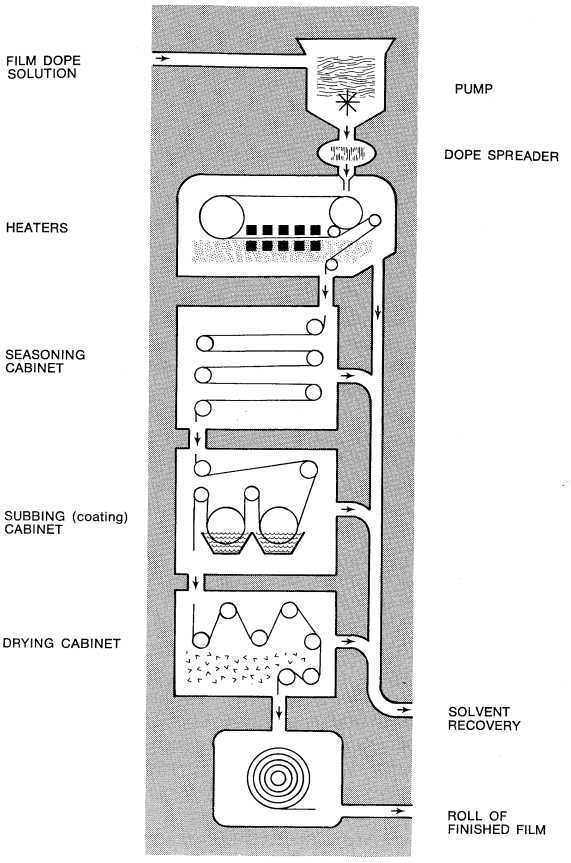


Figure 13: Basic operation of a film-casting machine.
Drawing by D. Meighan

a polished chrome surface or a copper band, up to 60 metres (200 feet) long with a special polyvinyl alcohol surface. The process is applied almost entirely to cellulose triacetate and diacetate, the former for the production of photographic film base and the latter for packaging and diazo reproduction films.

Foamed plastics production. Foamed or expanded plastics are low-density materials made by forming gas bubbles in the plastic while in a liquid or molten state. A variety of processes and types of products are available, depending upon the type of polymer and whether the structure is closed cell or open cell (*i.e.,* with the bubble cavities intercommunicating).

In some cases (*e.g.,* polyurethane foams), bubbles are produced in the low molecular weight mixture of reactants undergoing polymerization, partly arising from gases produced in the condensation process but usually supplemented by a blowing agent (*e.g.,* a volatile fluorocarbon liquid). Flexible polyurethane foams are of the open-cell type, like a sponge, and indeed are used for such purposes, as well as for heat insulation, upholstery, and other uses. Rigid or flexible thermoplastic materials (polyvinyl chloride, polyethylene, polystyrene) may be foamed in one of two general ways:

1. By dispersing a blowing agent in the material and heating it above its softening temperature. The blowing agent forms a large number of small bubbles in the material, and may be a volatile liquid such as pentane used in expanded polystyrene, or an unstable chemical such as azodicarbonamide, releasing much gas when heated.
2. By dissolving a gas such as nitrogen or carbon dioxide in the heated plastic at a high pressure. When the pressure is released, a large number of small bubbles are formed and grow in the soft plastic. This process may be carried out in autoclaves or continuously in extrusion machines, the plastic expanding as it leaves the extrusion nozzle.

Most of the foamed materials manufactured by the above two processes are produced in simple forms from which sheets or simple shapes are cut. In recent years so-called structural-foam processes have appeared, by which molded articles are produced from foamed plastics materials. Rigid plastic foams of relatively high density (*i.e.,* 0.5 gram per cubic centimetre) have a useful combination of lightness and rigidity, particularly if the molding process results in the formation of a relatively solid skin of plastic over a low-density expanded core.

Fabrication and finishing. Techniques of fabrication and finishing include mechanical machining, laser machining, roll-leaf hot stamping, printing, welding, radiation processing, vacuum metallization, electroplating, painting and decorating, mold decoration, and inserts.

Mechanical machining. Most plastic articles do not require machining except for the removal of sprues (small projections) on injection molded parts. Machining properties of plastic are entirely different from those of metal. Plastic is more flexible, necessitating maximum support to prevent bending. Buildup of heat in the workpiece is rapid, causing it to soften. Furthermore, the plastic may deform slightly after machining, making it difficult to do highly accurate work. To diminish heat, tools are honed and polished. Lubrication is supplied, and the speed of cutting is adjusted for different plastics. Drills have wide polished flutes, and oversized taps are used to counter deformation after tapping.

Machining properties of plastics

Plastic sheet, especially acrylonitrile–butadiene–styrene, can be cold-formed under high pressures between metal dies. Good lubrication and polished drawings are required—these drawings control the drawing in of the sheet during forming. Attractive articles may be formed from previously decorated flat sheet. Plastics can also be forged like metals if the blank is heated nearly to the crystalline melting temperature. Forged parts can have 10 times the impact strength of the blank.

Laser machining. A laser beam removes material by vaporizing it. Holes can be drilled in less than a thousandth of a second; thus perforations can be made on a moving plastic web, using a controlled pulse rate from the laser. The holes range in size from two to 50 mils (one thousand mils equal one inch or 25 millimetres) in diameter and are slightly tapered.

Aerosol nozzles in nylon and acetal are satisfactorily bored with lasers. Film can be slit with a continuous beam, without forming any dust; the beam can also be used for heat sealing.

Roll-leaf hot stamping. This process derives from the art of tooling gold leaf on leather. It is applied mainly to decoration or labelling of thermoplastic articles and consists of the application of a complex layer of which a typical composition includes a carrier film, usually polyglycol terephthalate, 12–25 microns (0.0005–0.001 inch) thick; a release coating on the carrier, based on a waxy material and about one micron thick; a lacquer coating that protects the decorative coating that is to follow (the lacquer is metallized with aluminum or gold to a thickness of a few microns, having been previously printed, or is coated with a pigment without a metallic layer); and, a final heat- and pressure-sensitive adhesive layer to give maximum adhesion to the plastic to be printed. The adhesive-coated side of the film is pressed by a heated die to the object to be decorated, and the adhesive bonded to the plastic under pressure. After a short cooling period to allow the adhesive to solidify, the carrier film is stripped off, that part of the resin coating carrying the metal coating or pigment that corresponds to the die form adhering to the object.

Decorating thermoplastic articles

Polyester film as a carrier has the advantage of withstanding high temperatures, which is important because the die may be at a temperature as high as 130° C (270° F), though the usual temperature is about 110° C (230° F). Metal dies are usually steel, though softer metals can be employed for short runs. There is, however, a growing use of silicone rubber, the latest development being a metal–rubber sandwich, the rubber layer being about 1.6 millimetres (0.63 inch) thick, faced with 0.8 millimetre (0.32 inch) of copper, which can be photoengraved. Such a die gives flexibility sufficient to cope with surface irregularities on the plastic faces.

Printing on plastics. There are two principal factors involved in printing on plastics: the choice of ink and the type of printing process. Ink choice is influenced by the

printing method. The printing method in turn is determined by the shape and the plastic used, usually flexible material such as polyethylene and polypropylene, thin polyvinyl chloride, nylon, or high-density polyethylene. For these materials flexography, which involves printing with a rubber printing plate, gravure, using engraved metal rollers, and lithography are widely employed, chiefly for printing on the plastic web. Dry offset printing or stamping is employed for molded bottles, jars, tubes, and tubs. For bold decoration, the silk-screen process, a form of stencil printing, is employed. Another process under rapid development is electrostatic printing, in which ink powder is expelled from an image screen by a difference in electrostatic potential. Ink formulation is specific to the material and process. Alcohol-soluble polyamides are used for polyolefins; oxidizing alkyd resins and oxidizing thermosetting systems are used for polyvinyl chloride, polystyrene, nylon and also polyolefins. When printed sheets such as cut film are stacked as they come off an offset press, they are interleaved with paper panels (until the ink is dry) to prevent ink smudging.

Heat, high-frequency, and ultrasonic welding. Polyvinyl chloride sheet, in its various forms, can be welded by a hot jet of nitrogen or carbon dioxide and a rod of lower melting polyvinyl chloride. Overlap joints in tank linings of polyvinyl chloride, polyvinylidene chloride, and polyisobutylene can be welded with a type of soldering iron, followed by a small pressure wheel. Polymethyl methacrylate with a hot jet welds through contact with the monomer released by heat decomposition.

Electronic heat sealing

Polyvinyl chloride sheet is also particularly suitable for electronic heat sealing, as is any plastic that will absorb radio-frequency energy. A heavy-duty steel press with brass dies is employed. A radio frequency generator is required, with power output ranging from three to 25 kilowatts. Electronic techniques are used to control positioning, sealing, and the length of time the press stays closed (dwell time). Very strong joints are made, and the process has been much used for applique decoration.

Ultrasonic (above the range of human hearing) vibrations, at about 20,000 hertz (cycles per second) are transmitted into the plastic assembly, generating heat at the joint surfaces where melting takes place. The harder the plastic, the more effective are ultrasonic welding techniques. Ultrasonic welding can be used for spot welding; staking, in which a plastic stud is inserted into a hole in metal; and for molding in inserts by means of a hole slightly smaller than the insert, heat being generated at the point of contact between metal and plastic in a second or less.

Radiation processing. Two plastics industrial processes employ high-energy electrons, the first being the curing of certain types of resin, which become thermosets (causing them to take on a rigid form) at room temperature and the second, the cross-linking of thermoplastics, causing the molecules to form a network structure, thus making the plastic infusible. Radiation sources are either electron accelerators or radioactive cobalt. In the case of accelerators, the material is passed in front of a window of the accelerator, where it is briefly exposed to radiation; the cure, or cross-linking, occurs rapidly without any catalyst. Thermoplastics, such as polyethylene, which have been cross-linked in this way, if stretched while hot, can be frozen in the new dimension; subsequent heating causes rapid and strong shrinking. Thus, for example, a four-core electrical cable can be capped with a four-tubed glove of treated polyethylene and shrunk down throughout by heat to make tight joints at all points. The process can be applied to polyethylene and other wire insulations by exposure to an electron beam, thus providing increased thermal stability and improved stress-cracking properties. Similar exposure of film gives increased clarity when shrunk in place, with better resistance to chemical attack but little reduction in permeability to water vapour.

Vacuum metallizing. Vacuum metallizing, the earliest method of metal coating of plastics, is accomplished by evaporating a metal, usually aluminum, onto plastic surfaces within a high vacuum chamber, up to 2.1 metres (seven feet) in diameter and containing hangers so arranged that the workpieces can be rotated. Rotation is necessary because the metal particles travel in straight lines. Alongside the work are tungsten heater filaments on which pieces of aluminum are suspended. The tungsten filaments are electronically heated to melt the aluminum and the temperature then raised to evaporate it while the workpieces are rotated. For vacuum metallizing, very high vacuum is required; to achieve this, most plastics are pretreated to remove moisture and any volatile substances. The time required for this process can be reduced by lacquering and baking the article, then applying a coat of lacquer after metallizing to protect the metal surface and provide colour if necessary.

Continuous coating

Continuous coating from rolls of plastic of many types is accomplished by use of special crucible sources and winding mechanisms. Metallizing is used largely for decoration; metallized polyester film is used for electrical capacitors and, finely slit, for metallized yarn.

Electroplating. The use of plastics such as acrylonitrile–butadiene–styrene, polypropylene, and polysulfone for automotive parts has been enhanced by electroplating with metal. The process depends on etching, which is easiest with acrylonitrile–butadiene–styrene because the dispersed rubber particles can be removed by oxidizing. Other plastics depend on strong acid attack. Immersion in a tin salt solution, followed by palladium salt, results in the deposition of catalytic palladium that adheres to the etched holes. Further immersion in copper or nickel salt solutions causes a layer of metal to be deposited on the plastic through reaction with the palladium, thus providing a conductive surface on the plastic for electroplating with copper, nickel, and chromium. The copper deposition may be thick enough to serve for printed electronic circuits. A number of intermediate steps are required and there are many proprietary variations. It is important that the moldings be free from strain and the visible surfaces as perfect as possible. Areas up to 2,260 square centimetres (350 square inches) can be plated.

Finishing. Paint can be used to obtain multicoloured or pearly finishes, to reduce gloss, and to reduce the typical plastic appearance of an object. It can also give light-coloured finishes to dark resin moldings and objects made from reclaimed scrap. Paint is usually sprayed but can be roller coated, especially for raised letters.

Printed foils can be located, either mechanically or by static electricity, in a mold; thermosets are decorated with printed melamine-impregnated paper, as for dinnerware, the foil being inserted after partial curing by interruption of the molding process. For thermoplastics, film, preferably of the plastic being molded, is inserted printed-side inward.

Inserts, placed in the female half of a thermoset mold, can be molded in. For subsequent insertion, self-tapping and thread-forming inserts can be used. Locking devices such as serrated rings are used to secure bushings (metal linings for a hole). The cheapest inserts are driven into place, using projections that fit in grooves (splines) or expansion devices instead of threads.

INDUSTRIAL PLASTICS

Structural foams. *Rigid foams.* Rigid polyurethane foams in sandwich forms have wide applications as building components because of the stiffness imparted by the thick foam centre for a given weight. They are also the best insulants known today and so have wide application in refrigeration and in buildings, where they are applied in fitted slab form or are foamed into cavities at the building site. They can also be applied by spraying about six millimetres thickness with each pass of the spray gun. The ability to spray a foaming mixture through a single nozzle is a great advantage in application.

Sandwich forms

A very important use of rigid foam is for furniture parts to reproduce wood structures; these can be injection molded. Since the polyurethane remains in the mold a rather long time to allow foaming and setting, molds can be mounted on a rotatable table, coming around in turn for a shot of foamable material. Rigid polyurethane foam can be screwed and nailed with a retention about equal to white pine lumber.

A major advance in the manufacture of sandwich struc-

tures is a new method of injection molding invented in Britain, in which a large machine is used to produce moldings up to 1.2 metres (four feet) square. A double valve provides differential feeds to the mold, the first supplying a skin of selected plastic, the second a foamable core material that forces the first polymer to the surface of the mold; the process resembles the blowing up of a balloon. Moldings of great strength and any desired surface are obtained.

Flexible foams. Flexible foams, usually polyether urethane, are made in slab form up to 2.4 metres (eight feet) in width and as much as 1.5 metres (five feet) high; these are then cut to required shapes or sizes or are molded. The molded foams may be hot molded. This involves filling heated aluminum castings and gives a product having high resistance to compression, as for automotive seats; or they may be cold molded, a process used particularly for semiflexible foams with high load-bearing properties. Used almost exclusively by the automobile industry for crash pads, armrests, and dashboard covers, the process involves machine mixing the ingredients and pouring them into aluminum molds lined with vinyl or acrylonitrile–butadiene–styrene skins, which become the cover material for the part.

Polyester flexible foams with a special silicone surface active agent are used for textile interliners, with densities up to about 0.11 gram per cubic centimetre (seven pounds per cubic foot); these can be made down to 30 mils thick and used for lamination.

Polystyrene foams

Polystyrene foams are made in a wide range of densities, from expandable beads, either by extrusion through slot-shaped openings to 40 times the original volume to form boards directly or by foaming in steam chests to form large billets. Using small beads in stainless steel molds, cups can be molded with thin sections.

Thin sheet for embossed packaging can also be made by the tube extrusion technique. Though packaging is a major use for forms made in closed molds, the largest use is for building panels; they can be plastered directly. Acrylonitrile–butadiene–styrene can be expanded from pellets and is particularly suitable for wood-grain effects and for the production of heavy sections.

Expanded vinyls can be made from plastisols for flooring or textile linings by calendering with a blowing agent and laminating to a fabric base, and by injection molding for insulation and such articles as shoe soles. An improved material is now obtained from cross-linked polyvinyl chloride and competes with polyester in glass reinforced plastic. Syntactic polybenzimidazole owes its expanded properties to the incorporation of extremely small hollow spheres (microballoons) of glass or polyfluoride resins. It is stable up to 300° C (570° F); its heat resistance and light weight adapt it particularly for use in space vehicles.

Sheets and films. *Vinyls.* Plasticized polyvinyl chloride, made by a calendering process, can be sewn, heat-sealed, or electronically sealed; it is used for apparel, shower curtains, protective clothing, and the like. Made in any colour, transparent, translucent, or opaque, polyvinyl chloride can be coated with pressure-sensitive adhesives and printed with decorative patterns. Thicker sheet is coloured and embossed for women's handbags, luggage, seat covers, and upholstery.

Thin film is also used for packaging, especially of meat and fruits; if biaxially stretched, it forms a shrink film that retracts up to 60 percent. Another important use is as a laminate for printed paper.

Flooring tiles, largely made of polyvinyl chloride, are built up by lamination and decorated either by printing or by rolling in colour chips; the commonest tile is vinyl asbestos, pressed into sheets on calenders and then embossed and cut into tiles.

Rigid polyvinyl chloride sheet has high dimensional stability and flame retardancy and is often used in corrugated form for building construction. Partitions, rain gutters, downspouts, industrial tank linings, vacuum-formed signs, and lighting panels are other uses.

Styrene film is widely used for semirigid food containers, especially packaging, being heat formable if held on all four sides to prevent shrinkage.

Acrylonitrile–butadiene–styrene possesses a wide range of properties, notably scuff resistance, plus high impact strength at lower temperatures, making it suitable for high-quality luggage, refrigerator linings, and food and detergent containers because of its chemical resistance and freedom from stress cracking. Because it is easily formed by heat, thick acrylonitrile–butadiene–styrene sheet is used for sports-car bodies and automobile doors.

Uses of nylon films

Nylon films are mostly from nylon-6,6 and are ideal for food packaging because of strength, impermeability to oils and greases, and high melting point. As such films are steam sterilizable, they find many uses in hospitals. They are frequently laminated. Acrylic films have high resistance to ultraviolet light and external exposure; their prime use is in surfacing laminates.

Fluoroplastics. Fluorinated ethylene–propylene and chlorotrifluoroethylene are the principal fluoroplastics. Fluorinated ethylene–propylene resists extreme temperatures from that of liquid hydrogen to 275° C (525° F) and is extremely inert in the presence of chemicals. Since it can be bonded to metal with special adhesives, fluorinated ethylene–propylene is used in printed circuitry, to insulate flat cable used in the computer industry, in space vehicles, and for heat-sealed industrial and biological packaging.

Both fluorinated ethylene–propylene and chlorotrifluoroethylene are noncombustible and have very low moisture vapour transmission rates; the fluorohalocarbons also have excellent thermoforming characteristics. They are used in packaging of critical hardware, in aircraft insulation, encapsulation, and in other electrical objects.

Polyurethane film and sheet are available in thicknesses from ½ mil to 100 mils (2.5 millimetres), the outstanding characteristic being toughness and resistance to tearing. Tensile strength and elongation are both very high, leading to uses for awkward packaging and for inflatable sheet used to stabilize ship cargoes.

Cellulose acetate. Cast film up to 10 mils thick is used for rigid packaging, especially for "blister" packaging, designed to protect small articles by inserting them in preformed shapes.

Acetate film can be printed or drawn upon and hence is used widely for visual aids and in engineering and architectural drafting.

Cellulose triacetate is more moisture resistant than ordinary diacetate and, until recently, plasticized with high-boiling plasticizers, has been pre-eminent as a photographic film base, being cast, for example, to make rolls of motion picture film base 910 metres (3,000 feet) long and more than one metre (40 inches) wide. After emulsion coating, these are slit into standard motion picture film widths for edge perforation and addition of the sound track. Cellulose triacetate is similarly used for amateur film bases at 3.5 mil thickness and for X-ray films at 7.8 mil. In the field of colour-separation films, however, it is being replaced by polyethylene terephthalate, which has greater dimensional stability. Cellulose acetate film has a wide range of other uses, the result of its great tensile strength (1,760 kilograms per square centimetre, or 25,000 pounds per square inch), its high dielectric strength (withstands high voltages), and resistance to heat. Thus it is extensively used for electrical insulation of motors and capacitors; for packaging, which accounts for three-quarters of its utilization; and for magnetic recording tapes. Polycarbonate sheet is heat-, odour-, and stain-resistant and practically unbreakable.

Uses of polyimide film

Polyimide film has no melting point and so can be used at high temperatures; it maintains its flexibility at 4° above absolute zero. Its toughness, flame resistance, and radiation resistance make a valuable combination for electrical uses, especially in space technology.

Laminates, reinforced plastics, and composites. High-pressure laminates with phenolic resin binders are available in many grades, from strong paper laminates with varying quantities of resin to fabric-reinforcements of cotton, asbestos, nylon fibre, and glass-fibre fabric. All are rigid, can be built up to required thicknesses, can be readily machined into required dimensions, and in many cases can be punched. They are chiefly characterized by excellent electrical properties and resistance to humidity, combined with high mechanical strength, all these properties

depending on choice of resin content and reinforcement employed. Cotton-fabric-reinforced laminates having very high impact strength are used in a wide variety of mechanical applications, especially gears.

Special uses of composites

Composites, which incorporate a third layer in addition to the resin and reinforcement, have special applications. Very large quantities of copper-clad laminates are used for printed circuits, and there is an increasing use of plasticized polyvinyl chloride laminated to steel or, in some cases, to aluminum. The polyvinyl chloride protects the metal and also carries pigmentation of desired colours.

Decorative laminates are widely used on a base of phenolic paperboard by laminating the surface with a printed sheet of wood grain or specific designs and finished with a urea- or melamine–formaldehyde transparent protective surface layer. Melamine, silicone resins, and epoxides are used instead of phenolic resins in applications requiring electrical properties under wet conditions or flexural (bending) strength at high temperatures. Where reinforced shapes other than sheets are required, postforming laminates or prepregs are used. Postforming laminates are sheet assemblies in which the phenol–formaldehyde retains a degree of thermoplasticity, so that it can be molded into shapes such as kitchen trays and covers for electrical apparatus, the molded depth depending on the extensibility of the reinforcing fabric. Prepregs are glass-fibre mats impregnated with resin, fillers, colour, and catalyst, the resin being polyester, epoxides, phenolics, or silicones, partially cured, the final curing process being effected in the mold. Prepregs are particularly adapted to mass-production techniques.

Low-pressure (contact) laminates

Contact laminates are formed by placing reinforcing materials with polyester or epoxide resins, in viscous liquid form, on a mold. The major plastics material used is the polyester type previously described; the reinforcing material is glass fibre coated with a silicone or silyl compound to promote adhesion to the plastic. The fibres are used either as rovings (slightly twisted strands or rolls of textile fibres), mats, yarns, or woven fabrics, or, for reinforcing molding, as powders in the form of short strands cut in lengths up to five centimetres (two inches).

The desirable quality of glass fibres is their tensile strength, amounting to 17,600 kilograms per square centimetre (250,000 pounds per square inch) in commercial form. Because of this great strength, glass-reinforced plastic can be used for such structures as small ships, footbridges, etc. A glossy finish can be given to the surface of these laminates by finishing them with a thin mat of staple filaments.

Because of its strength and lightness, glass-reinforced plastic has been much used for transport applications, and there is a small production of all glass-reinforced-plastic car bodies, made by a hot-press molding process. Automobile manufacturers are beginning to turn attention to compression-molded polyester glass composites for large area moldings, including body panels.

A different method of producing glass-reinforced plastic is filament winding, in which glass-fibre filament is passed through a resin bath and wound on a former to produce pipe and tanks of great strength. Pipes of large diameter can be made in this way.

Many other fibres are in use or under development; polyvinyl alcohol fibres, for example, adhere better than glass to plastic, giving higher impact strength and lighter weight; but the greatest interest lies in materials that offer greater stiffness and strength. Boron fibres are about six times as stiff as glass, but no stronger, whereas carbon fibres made from acrylonitrile filament combine increased stiffness with very high tensile strength ranging from 300,000 to 400,000 pounds per square inch. Composites of carbon fibre and resin can have the same tensile strength as steel but weigh only a quarter as much. Wear rate and friction can be much reduced, and resistance to high temperatures can be greatly increased. For further information on composites, see below *Miscellaneous inorganic synthetic fibres.*

Miscellaneous industrial plastics. Extruded forms, made on an extruder with a special die, though commonly produced as rods or tubes, may also be manufactured in geometrical shapes to suit given tasks (curtain runners, for example), as interlocking strips for panels and window frames, and as a special interlocking strip that is wound into a spiral to form polyvinyl chloride pipes of large diameter finished by continuously welding the spiral joint.

A special application of extruded film, generally polypropylene, is the use of tapes, which are woven into simulated jute products for sacks and carpet backing. Special treatment of the tapes leads to manufacture of carpet yarns.

Engineering plastics

Engineering plastics, materials that can substitute for metals, require tensile strengths well above 350 kilograms per square centimetre (5,000 pounds per square inch) and flexural strengths (resistance to bending or flexing) at least 15,000 pounds per square inch. Plastics with these properties as a rule have softening points well above 100° C (212° F), so that their stability is unaffected by any ambient temperatures to which they are likely to be exposed; also, they can be sterilized. The earliest use of plastics in engineering was as gears cut from laminated phenolics and used for heavy duty with water lubrication. In current use are thermoplastics such as acrylonitrile–butadiene–styrene, polycarbonates, and nylon, admirably adapted for light mechanisms such as clocks and cameras as well as heavy parts such as fuel tanks and many smaller automotive parts. The use of chopped-glass-fibre filling, which can increase tensile and flexural strengths by about 490 kilograms per square centimetre (7,000 pounds per square inch), has increased the scope for such uses.

(E.G.C./Ed.)

Man-made fibres

Man-made fibres consist of two broad groups, based upon the origin of the fibre-forming substance. The first group, of which rayon and acetate are examples, are produced by modifying natural fibre-forming materials such as cellulose. The second group, frequently called synthetics and including such fibres as nylon and polyester, are produced from synthetic chemicals. Like other textile fibres, man-made types are long and narrow; they are composed of long, threadlike molecules, which are made up of hundreds or thousands of atoms strung together in a chain.

HISTORY AND DEVELOPMENT

The English physicist Robert Hooke suggested in the 17th century the possibility of extruding artificial silk by a mechanical imitation of the silkworm. In the 19th century, an English weaver named Louis Schwabe successfully produced filaments from molten glass by forcing the liquid through nozzles ending in fine holes; on contact with air the glass hardened into a fibre.

Cellulose as a raw material

This first man-made fibre was not suitable for textiles. The natural cellulose of wood was subsequently considered as a potential raw material for the production of textile fibres. Wood cellulose is associated with various gummy substances, however, making it unsuitable for textile use. Experiments were undertaken with the object of dissolving the cellulose, separating the resulting liquid from its impurities, and then extruding and hardening the liquid to form a fibre. In 1846 a German chemist discovered that cellulose could be converted into another substance, nitrocellulose, by treatment with nitric acid. Unlike its parent material, nitrocellulose dissolves readily in solvent substances such as a mixture of ether and alcohol. The molecules retain their long, threadlike shape, and small groups of atoms become attached to their sides, making the new substance more soluble. A patent was issued in 1855 for the manufacture of nitrocellulose fibres, but despite their strength and flexibility, their extreme flammability made these fibres unacceptable for textile manufacture.

In 1883 Sir Joseph Wilson Swan, a British scientist working on the problem of producing a filament for an electric light, patented a process for squeezing a nitrocellulose solution through holes to form filaments, then treating the filaments with chemicals to change the flammable nitrocellulose back into harmless cellulose. Swan exhibited articles made from his filament, but he did not attempt to develop its potential for textile manufacture. In France, Count Hi-

laire de Chardonnet produced a fibre that he called artificial silk. Like Swan, he employed a nitrocellulose solution squeezed through tiny holes, hardened in warm air, and then converted back into cellulose by chemical treatment. Chardonnet exhibited materials woven from this artificial fibre at the Paris Exposition in 1889 and secured financial backing to build a factory at Besançon, in the east of France. Manufacture of Chardonnet silk, later known as rayon, the first commercially produced man-made fibre, began in 1891.

Development of rayon

Although Chardonnet's process was simple and involved a minimum of waste, it was slow, expensive, and potentially dangerous. In 1890 another French chemist had patented a process that made use of the earlier discovery that cellulose could be dissolved in cuprammonium liquor, a solution of copper salts and ammonia, and after extrusion be regenerated in a coagulating bath, forming continuous filaments. The cuprammonium process replaced the Chardonnet process, but was expensive and fell into disuse for some time until development of an improved spinning method, permitting manufacture of very fine filaments, revived the popularity of cuprammonium rayon. In 1892 another method of producing a regenerated cellulose fibre was devised, involving the chemical treatment of cellulose to form a crumblike material called cellulose xanthate, which could be dissolved to form a viscous liquid known as viscose. Extrusion of the fine jets of viscose into an acid coagulating bath resulted in the regeneration of the cellulose in the form of filaments. Commercial production began in 1905, and the industry developed rapidly, with viscose becoming a major man-made fibre.

A second artificial fibre made from cellulose was developed after World War I. Two Swiss brothers, Henry and Camille Dreyfus, established factories in Great Britain and the United States to produce a cellulose acetate solution used as varnish for the fabric wings of aircraft. At the end of the war, with no further demand for aircraft varnish, the Dreyfus brothers turned to fibre production, and in 1921 they began commercial manufacture of the product called Celanese. This process involves conversion of the cellulose to a derivative, cellulose acetate, which is then dissolved in a solvent such as acetone, forming a solution that is extruded to form a filament.

Although cellulose remains the most important source of man-made fibres derived from natural substances, other natural materials capable of forming fibres are also used commercially. Proteins, for example, exist in nonfibrous forms that can be manipulated to make fibres. Fibres produced from protein materials include casein, from milk; zein, from maize (corn); and arachin, from peanuts.

The protein is dissolved and the solution is extruded in the form of fine jets that can be hardened into filaments. None of these has achieved the success for textile applications of fibres derived from cellulose.

Alginic acid, extracted from seaweed, is a chemical relative of cellulose that is also used as a raw material to produce fibres. Because they are easily dissolved, alginate fibres have specialized applications in the manufacture of very sheer wools and laces, but are of minor importance as textiles fibres.

Development of synthetic polymers

In the period following World War I, chemists carried out extensive research on methods of forming the chainlike molecular structures called polymers by linking up atoms or groups of atoms to create long molecules. Various synthetic polymers were made, including plastic and rubber types. Finding that many of these polymers could be dissolved in solvents, researchers extruded the polymer solutions through spinneret holes to determine whether filaments could be formed in the same way as rayon.

An experimental fibre had been produced by German chemists from the organic chemical compound polyvinyl chloride as early as 1913, but it had little potential for textile use. In 1928 fibres were spun in Germany from a copolymer, a chainlike molecular structure formed of two or more simple chemical compounds, in this case vinyl chloride and vinyl acetate. Possibly the first synthetic textile fibre was commercially produced in 1936 from chlorinated polyvinyl chloride but was of limited value for textile applications.

The real beginning of the synthetic fibre industry began with research on polyesters and polyamides, both chemically produced substances having their atoms assembled in long molecular chains, under the direction of Wallace H. Carothers at E.I. du Pont de Nemours & Company, in the United States. This work led to the development of nylon fibre in 1935.

Increasing research in polymer chemistry after World War II resulted in the development of many new classes of commercially important synthetic polymers for use as textile fibres, including polyamides, polyesters, polyacrylonitriles, polyvinyl chloride, polyvinyl alcohol, polyolefins, and polyurethanes.

Outside the textile field new fibres, as previously noted, have been produced to fill the technological needs of modern industry. The space industry employs carbon, boron, and metallic fibres in the construction of composite materials having high strength and heat resistance. Glass, aluminum silicate, and other inorganic fibres have varied applications in industry, ranging from the filtration of corrosive materials to insulation of equipment against heat and sound.

Industrial uses of new fibres

FUNDAMENTALS OF PRODUCTION

The manufacture of a synthetic or semisynthetic fibre requires a raw material having the requisite long threadlike molecules. Although the raw material for semisynthetic fibres already exists in nature, in synthetic fibres the substance is built up from simpler chemicals.

Spinning. A substance that is to be converted into fibre must first be converted to a liquid or semiliquid state either by being dissolved in a solvent, or by being heated until it melts. This process frees the long molecules from close entanglement with each other, allowing them to move independently.

The resulting liquid is extruded through small holes, or spinnerets, emerging as fine jets of liquid that are hardened, forming solid rods having all the superficial characteristics of a very long fibre, or filament, such as silk.

The extrusion of liquid fibre-forming material, followed by hardening to form filaments, is called spinning, and is similar to the spinning process of silkworms and spiders, producing continuous filaments.

Several spinning techniques are used in practice, including wet spinning, dry spinning, or melt spinning.

Wet spinning. In wet spinning the solution of fibre-forming material is extruded into a coagulating bath that causes the jets to harden as a result of chemical or physical change.

Viscose, for example, is wet spun. The solution of cellulose xanthate is extruded into an aqueous solution of acids and salts, in which the cellulose, being insoluble in water, is regenerated to form solid filaments.

Dry spinning. In dry spinning the fibre-forming substance is dissolved in a solvent before the solution is extruded. As the jets of solution emerge from the spinneret, a stream of hot air causes the solvent to evaporate from the spinning solution, leaving solid filaments. Acetate is dry spun by extruding acetone solutions of cellulose acetate into hot air.

Melt spinning. In melt spinning the fibre-forming material is melted and extruded through spinnerets, and the jets harden into solid filaments as they cool on emerging from the spinneret. Nylon is a melt-spun fibre.

Stretching and orientation. The extrusion process produces some orientation of the long molecules forming the filament. Orientation is especially pronounced on the outer surface of the filament, where the molecules have been influenced by the edges of the spinneret hole, and the surface of an extruded filament is usually more highly oriented than the inner portion. This surface alignment, known as the skin effect, influences the dyeing properties of the fibre.

Skin effect

Orientation of the long molecules is completed by stretching the filament, pulling the long molecules into alignment along the longitudinal axis of the fibre, and causing them to lie alongside one another and develop cohesion. The degree of orientation depends upon the amount of stretch to which the filament is subjected, and by controlling the

stretching, or drawing, it is possible to control the tensile strength to a high degree.

The production of a synthetic fibre allows control over the chemical nature of the fibre-forming substance, producing a fibre with well-defined chemical properties and characteristic behaviour. Control of the chemical structure also allows control of the shape and physical behaviour of the long threadlike molecules produced.

Slender, uniform molecules may be packed alongside one another more efficiently than irregular molecules having awkward knobs and angles that destroy their uniformity, just as a bundle of bamboo canes will pack together more tightly than a bundle of twigs.

In producing a synthetic fibre, it is usual to design the long-chain molecules in a form allowing them to pack together with reasonable efficiency. Large groups of atoms attached to the sides of long molecules prevent the close packing that contributes to fibre strength.

Crystalline and amorphous regions. Wherever the threadlike molecules are able to pack closely together in a fibre, there is a tendency toward an ordered arrangement of the atoms with respect to one another. These tight-packed bundles of thread molecules are regions of crystallinity, possessing the regular and precise arrangement of atoms characteristic of all crystals.

Between these regions of crystallinity are regions in which the molecules have not been able to align themselves so precisely. These are the amorphous, or noncrystalline, regions of the fibre. In considering fibre structure the long threadlike molecules may be regarded as passing through regions of ordered crystalline arrangement that are embedded in amorphous material.

During the stretching operation the long molecules slide over one another as they are pulled into alignment in the direction of the fibre's longitudinal axis. As drawing continues, more and more of the molecules are brought to a state where they can pack alongside one another into crystalline regions; in these regions, the molecules are able to hold tightly together as a result of their cohesive forces and will resist further movement with respect to one another.

Resistance of nylon to stretching

When nylon is drawn after spinning, the filament may stretch to as much as five times its original length, and will then resist further stretching. At this point the molecules are aligned as effectively as possible into crystalline regions and are holding tightly together. The filament is then able to withstand great force without further stretching; if the load becomes too great the filament will rupture.

Effects of stretching and orientation. The degree of alignment of fibre molecules affects the properties of a fibre in several ways. The more closely the molecules pack together, the greater is the tenacity, or breaking strength, of the fibre. This increase in tenacity is accompanied by a decrease in the amount of elongation the fibre can sustain before reaching its breaking point; the molecules are not able to slide over one another as they could before alignment took place. Because the closely packed molecules no longer have great freedom of movement, a high degree of orientation also tends to increase fibre stiffness or rigidity.

Water is unable to penetrate between the molecules in a crystalline region of the fibre as readily as it penetrates the amorphous regions; therefore, increased alignment tends to lower the moisture absorption of the fibre. Increased resistance to water penetration affects the dyeing properties in highly oriented fibres; the molecules of dyestuff cannot migrate from the dyebath into the spaces between the fibre molecules. Increased resistance to penetration by foreign molecules improves the general chemical stability of a fibre; highly oriented fibres are more resistant to chemical attack.

Fibres change in appearance as they are drawn. In the undrawn state, nylon is usually dull and opaque; as the filaments are drawn and orientation increases, the filaments acquire the transparency and lustre characteristic of drawn nylon.

Filament modification. Manipulation and modification of the spinning process can produce various changes in the physical structure and form of a filament, which in turn affect the behaviour in use and the "hand" of fabrics made from them. Hand, or handle, is a general term for the characteristics perceived by the sense of touch when a fabric is held in the hand, such as drapability, softness, elasticity, coolness or warmth, stiffness, roughness, and resilience.

Cross section. The cross-sectional view of a filament may be varied by the shape of the spinneret holes through which the material is extruded. Such modification is increasingly important because of the many changes it can produce in the characteristics of both yarns and fabrics. Because circular cross-section filaments have poorer covering power, or ability to occupy space, than the lobed cross sections typical of the normal viscose filament, many synthetic fibres are now produced with dog-bone and trilobal cross sections, providing greater density when made up into fabric.

Straw-shaped filaments are sometimes produced; flat filaments are formed by extrusion through slit orifices instead of circular openings. Although these filaments have improved covering power, they tend to have a harsh feel.

Special effects may be produced in fabrics by the use of filaments in which the diameter is varied continuously between thick and thin.

Bubble-filled filaments. Yarn that has relatively high bulk for its weight produces fabrics having good covering power. The covering power of the filaments used to make up a yarn may be increased by spinning methods that cause bubbles of air or other gas to be trapped inside the filament. This effect may be achieved by spinning a solution that has been previously agitated to entrap air bubbles. A disadvantage of this technique is the tendency for the air bubbles to cause breaks in the filament during the spinning process. Breakage may be overcome by generating the gas bubbles after spinning. This effect may be achieved by adding a small proportion of sodium carbonate to viscose, so that bubbles of carbon dioxide are formed inside the filament by reaction with the acid of the coagulating bath.

Bubble-filled viscose filament was produced in the United States during World War II for use as a kapok substitute in life jackets, pontoons, insulated clothing, and similar items. In the production process, air injected in the filament during extrusion produced a continuous filament containing discrete bubbles about three to six millimetres long.

Spun-dyed filament and staple. Control of the spinning process enables the manufacturer to mix finely dispersed pigments with the solution or melted substance before extrusion. When spinning is completed, the pigments are locked inside the filaments, thus providing spun-dyed fibres that have good colour retention when exposed to light and laundering conditions. White titanium dioxide is sometimes added by this method to dull the natural sheen of man-made fibres.

Crimp. Man-made fibres cut to predetermined lengths in preparation for spinning into yarn are called staple, or staple fibre. The spinning qualities of staple are usually enhanced if the fibre has a waviness, or crimp, similar to that of wool. This crimp may be introduced mechanically by passing the filament between gear-like rollers, or chemically by controlling the coagulation of the filament to create a fibre having an asymmetrical cross section.

Chemical crimp

Chemical crimp has mainly developed from experimental work carried out in Japan, where much of the viscose staple produced is now crimped. The crimp is introduced by spinning viscose into a coagulating bath containing less acid and more salt than usual, followed by carefully controlled stretching. The resulting crimped filament is cut into staple and dried.

Filaments produced in this way have an asymmetrical cross section, with one side thick-skinned and almost smooth, and the other side thin-skinned and highly serrated. When wet, such fibres swell to a greater extent on the thin-skinned side than on the thick-skinned side, causing a tendency to curl.

A similar effect may be introduced into rayon by using the bicomponent technique originally developed for use in the production of certain synthetic fibres. This process involves the extrusion of twin filaments through orifices

set side by side in such a way that the two filaments join as they coagulate. The composite filament is composed of two viscose solutions, each having different swelling properties. Moisture encourages the filament to curl as one side swells more than the other.

The amount of crimp given to a fibre depends greatly upon the denier, or fibre size (the denier is the weight in grams of 9,000 metres of fibre). Too much crimp may cause neps, or tangles, during processing; too little crimp may cause low cohesion during processing.

Surface-modified fibre. The nature of the fibre surface influences the processing properties of the fibre and affects its behaviour in use. The striated surface of a regular viscose fibre, with its typical lobed cross section, influences the spinnability of viscose staple and affects the appearance and feel of viscose yarns. Because particles of dirt tend to cling to such indentations, fibres of this type are often more difficult to clean than similar fibres having nonserrated surfaces.

The scaly surface of wool contributes to the ability of this fibre to resist soil and wear, and attempts have been made to create similar surfaces on man-made fibres. One use for such fibres would be to provide improved blends with wool. Surface modified fibres have been produced by means of finishing treatments, and also by extruding the fibre through vibrating spinnerets.

Physical forms of man-made fibres. Man-made fibres may be produced in almost any desired length. When a number of filaments produced from a spinneret are twisted together, they form continuous filament yarns. Continuous filaments may be combined to form a thick rope, called a tow, containing thousands of individual filaments. A tow may be cut or broken into short lengths of staple fibre, similar in length to such natural fibres as wool or cotton. Staple fibre may be combed and spun into yarns by techniques similar to those used in spinning natural fibres. The yarns produced in this way are called staple, or spun, yarns.

Continuous filament yarns. Continuous filament yarns, consisting of unbroken filaments, tend to be smooth and compact and are used for satins, taffetas, failles, and similar fabrics.

For many applications, continuous filament yarns are less satisfactory than the looser, fuller staple yarns. Techniques have been developed to modify the texture of continuous filament yarns, however, to give them some of the characteristics of spun yarns. Yarns modified in this way are known as textured yarns. Various texturizing processes may be employed to introduce crimps, loops, or curls, giving these yarns a fuller, more open form; to introduce a high degree of elasticity, forming stretch yarns; or to increase the bulk of a yarn without necessarily introducing stretch characteristics.

Textured yarns

Spun, or staple, yarns. Spun, or staple, yarns consist of short fibres held together by the twist given to an attenuated (reduced in size) strand of fibres. The short fibres lie at various angles with respect to the long axis of the yarn, with the degree of uniformity depending on the combing and other treatments given to the fibre strands before they are twisted together. The surface of a spun yarn is rougher to the touch because of the protruding fibre ends, and spun yarns generally have a fuller hand and are warmer than continuous filament yarns. Spun, or staple, yarns are popular for sports shirts, suitings, sheets, blankets, upholstery materials, and other fabrics.

Monofilaments. Monofilaments, used for fishing lines and nets, consist of fibre-forming polymers spun into filaments sufficiently thick to be used singly.

NATURAL POLYMER FIBRES

As explained above, polymers are large molecules made by linking together many smaller, unattached molecules, or monomers. Natural polymer fibres are produced by altering substances of natural origin to raw materials that can be made into fibres suitable for textiles. Such products include rayon and acetate, made from cellulose; casein and zein fibres, from protein materials; and alginate fibre, from a seaweed substance.

Rayons. The production of rayon requires the conversion of cellulose to a soluble compound. The original Chardonnet process converted the cellulose to the highly flammable nitrocellulose. The development of safer methods led to the growth of the rayon industry.

Viscose rayon. One of the new methods, developed in 1892 by three British chemists, dissolved the cellulose to form a viscous liquid called viscose. A method of producing textile filaments from viscose was developed, as well as an aging process to improve spinning qualities, a multiple hole spinning jet, and a spinning box that coils the newly formed filaments by centrifugal force (see Figure 14).

The raw materials for viscose rayon may be cotton linters, the short fibres adhering to the cotton seed, or wood pulp derived from such timber as northern spruce, western hemlock, eucalyptus, or southern slash pine. The pulps of these soft woods, containing about 94 percent cellulose, are especially suited to fibre manufacture.

Raw materials for viscose rayon

Drawing by Don Meighan based on J.G. Cook, *Handbook of Textile Fibres* (1968); Courtaulds, Limited

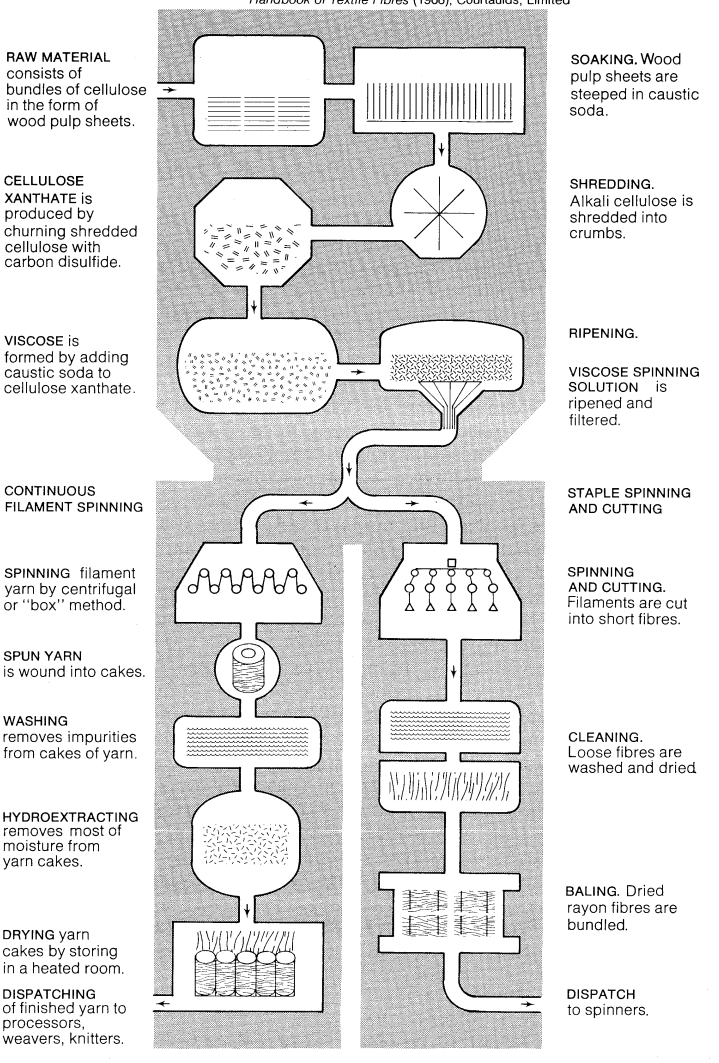


Figure 14: Basic steps in the manufacturing processes of viscose rayons.

The cellulose is purified and formed into thin sheets that are steeped in caustic soda and then shredded, forming powdery crumbs. During the aging period that follows, the caustic soda reacts with the cellulose, forming alkali cellulose. The aged crumbs of alkali cellulose are mixed with carbon disulfide in a revolving drum, forming cellulose xanthate, and are then treated with a dilute solution of caustic soda, forming sodium cellulose xanthate, a thick orange-brown solution.

The result is the viscose spinning solution, which is allowed to ripen for several days at carefully controlled temperatures, during which time it undergoes repeated filterings and increases in viscosity. After a final filtering stage, the solution is forced through the tiny holes bored

in the metal cap forming the spinneret. Emerging from the spinneret hole, the jet of viscose enters a coagulating bath of acids and salts, in which it is reconverted to cellulose, which coagulates to form a solid filament (see lower left of illustration). The filament may be manipulated and modified during the spinning process to produce variations of the cross-sectional shape, to increase covering power, to allow solution dyeing, or to introduce crimp or is cut into short fibres (see lower right of illustration).

The use of high-strength rayon in tires, conveyor belting, transmission belting, hosepipes, and similar products developed rapidly during World War II. The high-tenacity yarns employed for such uses are produced by application of a high degree of stretch during manufacture, when the individual filaments are in a pseudoplastic state. The stretching is influenced by selection and control of the chemicals used in both the spinning solution and spinning bath.

Increased understanding and control of spinning and processing techniques have led to the development of new types of viscose. The high-wet-strength, or polynosic, fibres are high-tenacity viscose rayons in which modification of the molecular structure has resulted in a fibre having high strength, especially when wet, and good dimensional stability and firmness.

In man-made fibres, the importance of rayon is similar to that of cotton among the natural fibres. Viscose rayon, produced in greater quantity than any other man-made fibre, is cheap and has a wide range of applications. It is a cellulose fibre, similar to cotton in its cellulosic structure, but it provides a range of yarns and fabrics with their own characteristic properties.

Viscose rayon conducts heat more readily than silk, has a cooler feel against the skin, and is highly absorbent, enhancing its value as a clothing material. The loss of strength occurring in most rayons when wet is a serious shortcoming, but modern resin finishes are overcoming this problem, and properly finished rayon garments have high-dimensional stability when wet.

Rayon staple

The introduction of rayon staple has allowed the blending of rayon with other natural and synthetic staple fibres, contributing moisture absorption and other cellulosic characteristics to blends of stronger and less absorbent fibres, among them most of the synthetics. Staple lengths suit particular blends and systems; a two-inch staple can be handled on cotton machinery and a four- to six-inch staple may be used with wool.

Rayon in its many forms is highly versatile and is used in every branch of the textile industry, including men's, women's, and children's outerwear and underwear; furnishings and carpets; household textiles; and medical fabrics. Crimped rayons are widely used in the manufacture of tufted carpets and rugs, tufted chenilles, curtains, upholstery, and nonwoven fabrics for surgical use. Spun-dyed rayon provides exceptional stability in textiles subjected to light, as in curtains and car upholstery.

Rayon and acrylic fibre blends are popular for knitted goods. Jersey knit fabrics, plain knits without a distinct rib, have long been made of cotton, characterized by low shrinkage; good hand, or feel; and good cover, or density. The introduction of foam-backed fabrics and laminates, in which two layers of fabric are joined together, has allowed other textiles to offer the advantages of cotton, and the use of spun-dyed rayon has increased in this field.

Flat filament viscose

Flat filament viscose is used in materials for which increased lustre and a firmer hand are required. A potentially important use of flat filament staple is in blending with wood pulp in paper production.

Cuprammonium rayon. Cuprammonium rayon, also known by the trademark name Bemberg, was developed in 1890, six years after the introduction of Chardonnet's process, but was not commercially successful until 1919. In the cuprammonium process, cellulose is dissolved in a solution of copper salts and ammonia, called cuprammonium liquor, and regenerated cellulose fibres are produced by extrusion of this solution into a coagulating bath.

Cotton linters or wood pulp are the raw material for cuprammonium rayon. Cotton linters, yielding very pure cellulose, are preferred for high-quality products, although the use of the less costly wood pulp has increased.

After purifying, the cellulose material is mixed into the cuprammonium liquor at low temperature, and stabilizing agents and caustic soda are added. The resulting solution is filtered by passing through a succession of nickel filter screens, and then is deaerated.

The filtered spinning solution is pumped to a nickel spinneret and extruded through the very fine holes. The jets of solution emerging from the spinneret holes flow into a glass funnel, meeting a stream of pure water that flows down through the funnel, and are coagulated to form plastic filaments. As the filaments are carried along by the stream of water, they are stretched continuously.

Manufacturers of cuprammonium yarns, or cupro, have been particularly successful in producing novelty yarns, such as slub yarn, having alternating thick and thin sections, and nub yarns in which lumps are intentionally introduced. These yarns may be made by extruding the spinning solution through spinnerets having two sets of orifices. The filaments from one set of orifices collect on a flat surface, forming bundles that adhere together, and are carried away at intervals to join the filaments extruded from the other sets of orifices, forming a composite yarn with the bundles creating slubs at intervals.

Desirable characteristics of cuprammonium rayon

Cuprammonium rayon is generally more expensive than other man-made cellulosic yarns, but its extra fineness and strength, attractive hand, subdued lustre, and good draping properties justify the extra cost for the manufacture of high-quality goods. This rayon, similar to other cellulosics in behaviour during laundering and dry cleaning, should be treated in the same way as viscose.

Cuprammonium rayon is made into satins and such sheer fabrics as chiffons, nets, and ninons, and is popular for underwear, dress fabrics, and linings. Fine denier slub yarns are used in dresswear, sportswear, curtains, and fine drapery fabrics; heavier denier slub yarns are used in suitings, upholstery fabrics, bedspreads, draperies, and other fabrics.

Yarn-dyed fabrics are popular for high-quality silklike linings, and for dress and upholstery fabrics. Reel-spun yarns produced in skeins ready for yarn-dyeing in the untwisted state are especially suited for such applications. The dyed yarn remains untwisted to form the crosswise weft of the fabric, and is twisted for the lengthwise warp.

Saponified cellulose ester. In the early manufacture of cellulose acetate, the fibre was often treated with alkali as a prelude to dyeing, partially converting the material back to cellulose. This treatment, called saponification, was eliminated from the dyeing process with the development of new methods for dyeing cellulose acetate. The process was revived, however, for application in the stretching operation. Saponification with a caustic soda solution during stretching deacetylizes the substance, producing a strong fibre, extremely fine in diameter, that is no longer cellulose acetate but has been converted to a highly oriented filament of regenerated cellulose.

Saponified rayon yarns are used for applications in which a high ratio of strength to volume and excellent dimensional stability are advantageous; *e.g.,* in parachute ropes and fabrics, tire cords, belting, hoses, and balloon fabrics. The high degree of fineness possible in filaments made by this technique has enabled the yarns made from them to replace natural silk in such applications as electrical insulation materials; *e.g.,* in hearing-aid equipment. Coated fabrics provide light, strong tarpaulins and other protective fabrics.

Cellulose ester fibres. Viscose and cuprammonium are regenerated cellulose, resembling cotton or flax in their chemical structure. There is another important form of cellulosic man-made fibre in which treatment with acetic acid changes the wood or cotton cellulose to an entirely different substance, rendering it soluble and spinnable, and the filament remains in its changed chemical form after spinning. This fibre is made from a chemical derivative of cellulose, cellulose acetate.

Secondary cellulose acetate

Fibres spun from secondary cellulose acetate, in which the cellulose is not completely acetylized, or converted by acetic acid to cellulose acetate, were formerly described as acetate rayon. In modern terminology, rayon is used only in describing fibres consisting of regenerated cellu-

lose, such as viscose and cuprammonium rayons. The secondary cellulose acetate fibres are now known simply as acetate, and fibres spun from primary cellulose acetate, or completely acetylized cellulose, are called triacetate.

Acetate fibres. In the production of acetate fibres, the spinning solution consists of secondary acetate blended from a large number of batches to assure a high degree of uniformity. The blended acetate is dissolved in a solution of the solvent acetone and a small proportion of water, up to 5 percent of the weight of the acetone, and pigments may be added at this stage. Finely divided carbon black may be added in making black fibre; titanium dioxide is added to produce a dull fibre when a high degree of lustre is not desired.

The treated spinning solution, containing 20–30 percent cellulose acetate, known as dope, is filtered and deaerated. After a final filtering the dope is pumped through very fine holes in the spinneret, then emerges into a spinning tube, an enclosed vessel through which hot air flows at a temperature of about 100° C (212° F). As the jets of spinning solution meet the hot air stream, the acetone is evaporated, leaving solid filaments of cellulose acetate. More than 90 percent of the acetone in the jets is evaporated during the fraction of a second that the jet is moving downward through the spinning tube.

The newly formed filament of cellulose acetate is stretched slightly while still plastic, aligning the long molecules and developing the strength of the filament. Acetate filament yarns produced in this way are ready for immediate textile use without any of the washing or purification treatment necessary for wet-spun fibres.

The natural attractiveness of cellulose acetate, combined with its useful practical properties, has created a demand for acetate fibre that has continuously increased since the 1920s. Modern acetate fabrics can be treated in much the same way as natural fibres and rayons. They can be dyed and finished, washed and dry-cleaned, and will withstand all the conditions met in ordinary commercial and domestic use.

Because acetate is fundamentally different in its chemical structure from the rayons and natural fibres, certain differences in behaviour must be considered in the handling of acetate materials. Acetate is thermoplastic, softening at high temperatures, and will deform under pressure at about 205° C (400° F).

The tendency for acetate fibres to soften with heating is useful in the processing of acetate goods, so that fabrics can be embossed with patterns that are impressed on the warmed material.

The relatively low-moisture absorption of acetate fibres makes acetate less subject to damage by staining from many substances. Fruit juices, ink, food, and other water-soluble stains are easily sponged or washed out. Acetate fabrics dry rapidly, and are thus particularly suitable for bathing suits, rainwear, and umbrellas. Because cellulose acetate does not conduct heat readily, acetate garments are cool in summer and warm in winter.

Desirable characteristics of acetate

Acetate has little natural colour, so that the dyer can produce acetate fabrics in a range of shades varying from delicate tints to deep, heavy colours. The richness and variety of the possible shades, allied with the softness and pliability of the acetate fibre, have helped to make acetate into a beauty fibre. Acetate garments drape well and have an attractive hand. They are soft and never harsh, retain their shape if treated with reasonable care, and do not easily wrinkle.

Triacetate fibres. Complete acetylation of cellulose yields primary cellulose acetate, or cellulose triacetate. This product is soluble in chloroform and methylene chloride, but unlike the secondary cellulose acetate, is insoluble in acetone.

In 1950 Courtaulds Ltd., London, began the commercial development of cellulose triacetate fibres. Later, Courtaulds and British Celanese Ltd., Coventry, Warwickshire, linked their efforts to produce a triacetate fibre, under the trademark Tricel, with British Celanese Ltd. as the producer. In 1954 Celanese Corporation, U.S., began production of a triacetate fibre under the trade name Arnel, and triacetate fibres are now produced in several countries.

Experience gained in the production of secondary cellulose acetate was applied in the commercial development of cellulose triacetate.

Cellulose triacetate from many batches is blended and dissolved in a 20 percent solution of alcohol and methylene chloride. The solution is then filtered, deaerated, and pumped to the spinneret, emerging into a vertical spinning tube where it meets a stream of hot air. The methylene chloride evaporates to leave solid filaments of cellulose triacetate. An antistatic and lubricating substance is applied; the fibres are collected in the same way as secondary acetate filaments.

The chemical relationship between triacetate and acetate is a close one, but it is only in their tensile properties that the two fibres bear any real resemblance to each other. Triacetate has greater heat resistance than acetate. In many respects, triacetate behaves more like a true synthetic fibre than a semisynthetic fibre, possessing the thermoplastic properties and low moisture absorption associated with synthetic fibres.

Desirable characteristics of triacetate

Triacetate fibres retain some 70 percent of their strength when wet, and thus fabrics made from triacetate are easily washed and quickly dried. The heat-setting characteristics of triacetate are of great practical value, because heat setting renders fabrics free from shrinkage, imparts excellent dimensional stability, and permanently sets the fabric in the desired shape. Heat treatment applied after dyeing increases the fastness of the dyestuffs to washing and light.

Permanent effects may be produced in triacetate fabrics by heat treatment. Permanent pleats may be set in woven fabrics, and permanent embossing may be set in both knitted and woven materials. The permanency of pleats in blended fabrics depends upon the amount of triacetate in the blend, and two-component blends with cellulosic fibres require a minimum of 67 percent triacetate for permanent pleating to take effect.

The high melting point of triacetate provides a wide margin of safety in high-temperature treatment used in clothing manufacture and laundering.

Because triacetate fabrics have little tendency to shrink even before heat treatment, tighter constructions are required in triacetate and triacetate-blended fabrics than in equivalent fabrics of acetate; therefore, these fabrics do not tighten up during dyeing and finishing.

Protein fibres. Fibres manufactured from the regenerated protein substances obtained from plant and animal sources are sometimes known by the generic name of azlon. Sources for these materials include milk, peanuts, corn (maize), soybeans, egg albumen, chicken feathers, and gelatin.

Casein fibres. About 1935 Antonio Ferretti, an Italian chemist, produced pliable casein fibres having many of the properties associated with wool. Casein fibres have since been made under various names in a number of countries—*e.g.,* Lanital in Belgium and France; Fibrolane in Great Britain; and Merinova, an improved form of the original Lanital, in Italy.

Casein is obtained by treating skimmed milk with acid, causing it to coagulate and form a curd that is washed, dried, and ground to a fine powder. It is then dissolved in caustic soda. After the solution has ripened to suitable viscosity, it is filtered and deaerated.

The spinning solution is wet-spun by extrusion through spinnerets into a coagulating bath containing such chemical ingredients as sulfuric acid, formaldehyde, and glucose with water. The jets of solution coagulate into filaments in a manner similar to the coagulation of viscose filaments, and are stretched to some degree during coagulation.

Bunches of filaments are collected together into a tow as they leave the coagulating bath and are then steeped in a formaldehyde solution to harden. Further stretching may be applied at this stage. After treatment, the tow is washed and dried, crimped mechanically, and cut into staple.

Uses of casein fibre

The low strength of casein fibre, and its sensitivity to water, have restricted its use, but it has found a number of applications in certain textile fields, especially as blend fibres for mixture with wool, cotton, rayon, acetate, nylon, and other synthetic staple fibres. In blends with cotton and rayon staple, casein brings warmth, resilience, and a full,

soft hand. It allows production of good whites, and prolonged wear trials have shown that the whiteness of fabrics made from these yarns is preserved throughout the life of the garment, irrespective of the number of launderings.

Blends containing one part casein to two parts of either rayon staple or cotton are particularly satisfactory. Twill and float weaves bring out maximum suppleness in the fabric. Crease resistance is good and may be strengthened where required (*e.g.*, in suitings) by application of a crease-resistant finish. Shrink-resistant finishes are also used where necessary to improve the shape, stability, and resistance to shrinkage on washing.

Because casein filaments can be spun to very fine diameter, they may be blended with the finest qualities of wool. The proportion of casein used in blends with wool depends upon the effect required; one-third casein is generally satisfactory. Because casein may increase the shrinkage in finishing in some constructions, allowance for shrinkage in the setting of the cloth is necessary.

Alginate fibres. In 1883 an English chemist discovered alginic acid, a substance in common brown seaweed that serves a purpose similar to that of cellulose in land plants. During World War II, coarse monofilaments of chromium alginate were first produced at the University of Leeds. They were green in colour and were originally manufactured for use in camouflage netting.

In the production of alginate fibres seaweed is dried and milled to a fine powder. The alginic acid in the seaweed is converted to sodium alginate, and a sodium alginate solution is wet spun into a coagulating bath containing dilute sulfuric acid and sodium sulfate, to produce alginic acid fibre, or a bath containing calcium chloride and dilute hydrochloric acid to produce calcium alginate fibre. The jets emerging from the spinneret are coagulated into filaments of calcium alginate that are brought together, washed, oiled, dried, and wound.

The nonflammability of the alginate fibres has made them valuable for several applications in which fire hazards might exist, as in theatre curtains. Although nonflammability is desirable for certain types of clothing, the low wet strength of the fibres is a disadvantage, and a satisfactory fibre that can survive washing and also provide wearing comfort has been difficult to devise.

The solubility of calcium alginate fibres in alkali (*e.g.*, sodium carbonate solution), while a drawback to most applications, has been an advantage in some. In the production of loosely spun wool yarns the alginate fibres are dissolved away after knitting, leaving a fluffy lightweight fabric that could not have been made by normal methods.

Special applications of alginate fibres

Calcium alginate yarn is useful in hosiery manufacture. Socks are linked together by a few courses of alginate yarn, allowing continuous production, and are separated by cutting the yarn and dissolving away the remains. This technique enables perfect welts to be obtained in socks of all types.

Calcium–sodium alginate yarn has medical application, providing styptic elastic dressings and dressings that are hemostatic, nontoxic, and absorbable in the bloodstream. It is also used in dental surgery to plug cavities.

Other protein fibres. Other fibres have been made from protein regenerated from such natural sources as peanuts (groundnuts), corn (maize), soybeans, and collagen (insoluble fibrous animal protein). These fibres have been excessively soft and tended to lose strength when wet, and few have been successful.

Research into new uses for products of developing countries resulted in a fibre made from the protein of peanuts, a staple crop in many hot, humid regions. Experimental production of the fibre, called Ardil for the town of Ardeer in Scotland where it was first produced, began in 1946 and was suspended in 1957.

A fibre called Vicara, made from zein, the protein material of maize, was manufactured in the United States between 1948 and 1957. The woollike fibres were low in strength and were mainly used in blends.

The Ford Motor company of the U.S. pioneered in attempts to make a useful fibre from the protein of soybeans, manufacturing a fibre for use in automobile upholstery between 1939 and 1942. Production cost was high and fibre strength was poor, and the process was taken over by another company, which discontinued production after a short time.

Marena, a protein fibre manufactured in Germany from the collagen obtained from leather and hide waste, was used mainly for the bristles of brushes, and was not produced after 1959.

Continued research is aimed toward finding a cheap, waste-protein source having textile-fibre potential.

Natural rubber fibres. A technique for production of rubber filaments was perfected in the 1930s. Rubber latex, a suspension of rubber globules in water, is mixed with vulcanizing agents and other materials and extruded in the usual way. The jets of latex emerge from the spinneret into a coagulating bath, forming filaments of rubber that are subsequently vulcanized. This new technique made possible the production of very slender round filaments, in virtually unlimited lengths.

In the 1950s, rubber filaments achieved prominence because of improvements in thread quality that allowed production of white filaments that will not discolour unduly in use and are resistant to degradation by oxygen, light, and other agents. At the same time the introduction of high-speed warp knitting allowed a rapid expansion in the production of elastic fabrics with two-way stretch, thus increasing the demand for lightweight body-support garments.

Applications of natural rubber fibres

Rubber threads are used in knitted and woven fabrics, either in the bare form or covered with spirally wound textile yarns. A wide variety of fabrics has been designed employing rubber threads to improve garment fit or provide support. Use of the appropriate type of rubber yarn and associated textiles allows a wide range of fabric hand, from relatively stiff batistes and nets to soft and supple swimwear and underwear fabrics, all providing exceptional comfort.

SYNTHETIC FIBRES

The synthetic fibres are made entirely from complex organic chemicals such as polyamides, polyesters, polyvinyls, polyolefins, and polyurethanes.

Polyamide fibres. Polyamides are polymers containing recurring amide groups as integral parts of the main polymer chains. Synthetic polyamide fibres form nylon, a major textile fibre. Synthetic polyamides are produced by a condensation reaction between small molecules, in which the linkage of the molecules occurs through the formation of amide groups. They may be produced by the interaction of a diamine, a compound containing two amino groups, and a dibasic acid—*e.g.*, hexamethylenediamine and adipic acid—or by the self-condensation of an amino acid, or a derivative.

Fibres have been spun experimentally from thousands of polyamides produced by one of these condensation reactions, but only a very few have attained commercial importance in textile manufacture. Nylon was the first commercially successful synthetic textile fibre.

The term nylon, coined by du Pont, was defined as a generic term for any long-chain synthetic polyamide, which has recurring amide groups as an integral part of the main polymer chain, and which is capable of being formed into a filament in which the structural elements are oriented in the direction of the axis.

Nylon nomenclature

A method of nomenclature has been devised that retains the term nylon, but distinguishes between the different forms of polyamide. The number of carbon atoms in the constituents of the nylon are indicated by appropriate figures, with the diamine being considered first in polyamides made by condensing diamine and dibasic acid. The original nylon, made from hexamethylenediamine and adipic acid, is nylon-6,6, with the diamine and dibasic acid both containing six carbon atoms. The nylon made from hexamethylenediamine and sebacic acid is nylon-6,10.

When the polyamide is made by self-condensation of a single constituent (*e.g.*, an amino acid), the nature of the polyamide is indicated by a single figure representing the number of carbon atoms in the molecule of the original constituent. Nylon-6 is the fibre made by self-condensation of the amide caprolactam.

Nylon-6,6. Nylon-6,6 fibre is spun from polyhexamethylene adipamide, a polyamide made by condensation of hexamethylenediamine and adipic acid.

Unlike rayon and acetate, nylons are melt spun. The polymer is heated until it melts, and the molten material is forced through holes in spinnerets. As the jets of molten nylon emerge, they are cooled and solidified by contact with a stream of cold air, forming solid filaments. The filaments are drawn to produce a lustrous and strong fibre. Nylon-6,6 fibres offer a range of properties that make it one of the most successful of all synthetic textile fibres.

Desirable mechanical properties of nylon

The outstanding mechanical properties of nylon include high strength-to-weight ratio, high breaking elongation, excellent recovery from deformation, high abrasion resistance, and high flex resistance. These properties are affected only slightly by moisture, and fabrics remain strong and stable when wet.

The nylon fibre, yarn, or fabric may be heat-set readily in dry heat or steam. Nylon fabric that has been heat-set holds its shape during processing in boiling water and has wrinkle recovery characteristics commensurate with the severity of the setting operation. The thermoplastic nature of nylon-6,6 has allowed successful development of a wide variety of permanent bulking and textured effects on continuous filament yarn.

These nylons are widely used in home furnishing manufacture for carpets and upholstery. Useful characteristics of nylon for carpet manufacture include appearance retention, because of abrasion resistance; texture retention; recovery from crushing; durability; and dye fastness. In upholstery, the combination of strength, dyeability, abrasion resistance, high light-fastness, and lustre provided by nylon fabrics is valued for beauty, performance, and practicality. In wearing apparel fabrics nylon-6,6 offers protection, aesthetic appeal, ease of care, durability, appearance retention, and comfort.

Nylon lends itself to most types of fabric construction, including such apparel fabrics as surah, plain and brushed tricot, simplex (reversible knit), various linings, circular and fully fashioned hosiery, taffeta, crepe, satin, reinforced twill, ski-wear fabric, velvet fleece, brocade, matelassé, circular and full-fashioned knits, tape and ribbon, lace, sheer, organza, and seersucker. These nylon fabrics are used for lingerie, swimwear, women's outerwear, men's outerwear, stretch sportswear, hunting apparel, uniforms, hosiery, sweaters, dresses, socks, and gloves.

Industrial applications of nylon-6,6

The high strength, elasticity, abrasion resistance, and other characteristics of nylon-6,6 contribute to its suitability for a wide variety of industrial fabrics, including air springs, belting, filter fabrics, fish net, twine, hose, wash nets, press covers, ironer covers, paddings, parachutes, webbing, sewing thread, cordage coated fabrics, body armour or ballistic cloth, screening, felts, reinforced plastics, ropes, blend paper, papermakers' felts, tires, and the gray fabric used to back cloth to be roller printed.

The high strength and abrasion resistance of nylon-6,6, coupled with such properties as resistance to moisture and heat, and good resistance to fatigue and rupture, have made this nylon popular for tire cords.

Nylon-6. Nylon-6 fibre is spun from polycaproamide, a polyamide made by the polymerization, or linking together, of 6-aminocaproic acid molecules.

It is melt-spun by a process similar to that used for nylon-6,6. Nylon-6 polymer melts at 215–217° C (419–423° F); *i.e.,* about 35° C (95° F) lower than polyhexamethylene adipamide, or nylon-6,6 polymer. Molten polycaproamide is comparatively stable and may be held under an inert atmosphere at 250° C (482° F) for 16–24 hours without deterioration. Polycaproamide is much less sensitive to spinning conditions than nylon-6,6, and the spinning of polycaproamide directly from the polymerization vessel may be carried out more readily.

Nylon-6 filaments are drawn by passing over two sets of rollers, the second set moving with about four times the surface speed of the first set.

Nylon-6 offers a range of properties generally similar to those of nylon-6,6, except for certain differences, such as those resulting from the lower melting point of nylon-6.

Nylon-11. Nylon-11 fibre is spun from polyundecanamide, made by the self-condensation of 11-aminoundecanoic acid. This fibre was developed in France, with the cooperation of an Italian firm, and is now produced in several countries—most notably, Brazil, India, and the U.S.S.R., as well as France.

The compound used to produce nylon-11 may be made from castor oil, from ethylene and carbon tetrachloride, or from dodecane. Polymerization is carried out in three stages, with the monomer, the material in the form of small, unattached molecules, fed as an aqueous suspension into the reaction vessel.

Molten nylon-11 polymer is highly stable at the temperature used during melt spinning (about 215° C, or 419° F), and may be stored for long periods without deterioration. Oxidation is prevented by maintaining the molten polymer under an inert atmosphere.

The spinning process is similar to that used for nylon-6, and the filaments are drawn to a degree depending on the type of fibre required.

Distinctive properties of nylon-11

Nylon-11 resembles nylon-6 and 6,6 in many important properties, but has certain features affecting its potential applications. The melting point of nylon-11 (189° C, or 372° F) is on the low side for general textile use, and care must be taken in such elevated-temperature treatments as ironing.

The initial modulus, or stiffness, of nylon-11 is higher than those of the other nylons, making it useful in applications such as brush bristles; it also makes for easier processing. Nylon-11 yarns do not stretch as easily as nylon-6 or 6,6 yarns when subjected to such physical processing as winding.

The high initial modulus of nylon-11 makes it useful for the automobile tire cord market, because tires reinforced with nylon-11 are not subject to formation of flat spots when the car is parked in cold weather to the extent that nylon-6 and nylon-6,6 reinforced tires are.

The low moisture absorption of nylon-11, enabling it to retain its excellent insulation properties at high humidities, is a useful characteristic in electrical applications.

Having a specific gravity of only 1.04, nylon-11 is a light fibre, with greater covering power than other polyamides. It is used in a variety of textile applications, such as tricot knitted lingerie and underwear, hosiery, and woven fabrics, and serves in the same areas as nylon-6,6 and nylon-6, combining their ease-of-care characteristics with durability and comfort.

Heat-setting processes are used in the same way as with other thermoplastic fibres, and permanent pleats and creases may be obtained effectively.

Nylon-6,10. Nylon-6,10 fibre is spun from polyhexamethylene sebacate, made by the condensation of hexamethylenediamine and sebacic acid.

Nylon-6,10 fibres are similar to those of nylon-6 and 6,6 in many respects, but moisture absorption is lower than that of either, and the fibres are unusually resilient.

Other polyamide fibres. Since the early 1960s a number of new types of polyamide fibre have come on the market. Some were designed to serve in comparatively narrow specialized fields, and others extended the range of textile applications and brought improvements in the characteristics of nylon fabric generally.

Adjustment of the segment of the polymer chain separating the amide groups in a polyamide allows significant modification of the properties of the polyamide. This technique has been used in the production of polyamides now serving in specialized fields for which the normal nylon fibres are unsuitable. High-temperature resisting polyamides (*e.g.,* Nomex [HT-1 nylon]) and polyamides suitable for use as electric insulation (*e.g.,* nylon-11) are examples.

Polyester fibres. Polyesters are polymers made by a condensation reaction taking place between small molecules, in which the linkage of the molecules occurs through the formation of ester groups. They are usually made from esters of ethylene glycol and terephthalic acid.

In 1941 British researchers developed a synthetic fibre from polyethylene terephthalate by condensing ethylene glycol with terephthalic acid. These fibres are manufactured in many countries, in both their original and in modified forms. Other polyesters have been produced and

spun into fibres, some achieving commercial importance.

Polyethylene terephthalate fibres. Polyethylene terephthalate (PET) fibres are made by the condensation of terephthalic acid, or a derivative such as dimethyl terephthalate, with ethylene glycol. The spinning procedure is like that used for polyamide fibres.

The molten polymer, pumped through holes in a spinneret, emerges in the form of filaments that solidify and are wound into packages of undrawn yarn. This yarn is stretched to about five times its original length on draw-twist machines, at elevated temperature. In the production of high-tenacity yarns, the filaments are drawn to a higher degree than in the manufacture of regular tenacity yarn.

Desirable properties of PET polyester fibres

The desirable properties of PET polyester fibres, both filament and staple, include low moisture retention, high dry and wet strengths, high initial modulus, high resistance to and recovery from bending, low creep, ability to be heat-set, high abrasion resistance, good electrical insulation properties, good resistance to exposure to elevated temperatures, good resistance to most common chemicals, and good resistance to common solvents. The end uses for PET polyester fibres have developed largely around these useful characteristics, making PET one of the most versatile of all modern synthetic fibres.

Poly-1,4-cyclohexylenedimethylene terephthalate (PCDT) fibres. In 1958 the United States firm Eastman Chemical Products Inc., introduced a new type of polyester fibre under the trademark Kodel. This fibre is spun from the polymer made by condensing terephthalic acid with 1,4-cyclohexanedimethanol and is therefore of fundamentally different constitution from the polyethylene terephthalate polyesters that form the bulk of commercial polyester fibres. The polymer (PCDT) from which Kodel-211 is spun is made by condensing terephthalic acid with 1,4-cyclohexanedimethanol.

As the polymer is melt-spun, the filaments solidify as they meet cold air. They are then drawn to 4½ to 5 times their original length at a temperature of about 120° C (250° F).

Fibres of PCDT generally resemble PET fibres, but certain differences exert a significant effect on their uses. Fibres of PCDT have lower tenacity and elongation than the PET fibres but have superior recovery from stretch. They serve in applications where resilience and bounce are of greater importance than high tenacity, as in carpets, rugs, and knitwear.

Desirable characteristics of PCDT polyester fibres

The lower tenacity of these fibres contributes to their improved resistance to pilling, and the fabrics will shed pills (the small tangles of fibres resulting from friction) more readily than fabrics made from the stronger polyester fibres.

The moisture absorption of PCDT fibres is similar to that of PET fibre, producing similar behaviour. Mechanical properties are not affected by moisture and electrical properties are excellent, but the accumulation of electrostatic charges may cause difficulties.

Fibres of PCDT have a lower specific gravity than PET fibres, thus giving increased covering power and providing lightweight fabrics of great warmth and comfort. The high melting point of PCDT fibre allows a high safe-ironing temperature (about 218° C, or 424° F).

Fabrics containing PCDT fibre may be heat-set at a relatively low temperature—*e.g.*, about 160° C (320° F)—an especially useful quality in the finishing of blended fabrics containing wool.

Polyvinyl derivatives. *Polyacrylonitrile fibres.* During the period 1955–60, polyacrylonitrile fibre factories were established in Germany, Japan, Italy, Belgium, Canada, France, and The Netherlands, and later in many other countries.

These polyacrylonitrile fibres, known by the generic name of acrylic, are spun from polymers consisting of at least 85 percent by weight of acrylonitrile units, produced from ethylene oxide and hydrocyanic acid. The polymerization of acrylonitrile and its comonomer is commonly carried out by stirring the monomers with water in the presence of a catalyst and surfactants (detergents or other surface-active agents that reduce surface tension).

Dry spinning process for acrylic fibres

In dry spinning, the polymer is dissolved in an organic solvent such as dimethylformamide to form a solution containing 25–40 percent of polymer. This is degassed, filtered, and heated almost to the boiling point, then extruded through spinnerets.

The fine jets of solution emerge into a vertical tube or spinning cell, through which flows air or other gas at high temperature. As the jets fall through the tube, the solvent evaporates, leaving solid filaments of polymer. These filaments are brought together at the base of the spinning cell and stretched while hot to several times their original length.

In wet spinning, the polymer is dissolved in a suitable solvent and the solution degassed and filtered. It is then pumped through spinnerets into a coagulating bath containing a liquid in which the solvent is soluble, but the polymer is insoluble, and the jets of polymer solution coagulate into fine filaments. These fibres may be produced in a variety of cross-sectional shapes, each affecting the nature of the fabrics produced from them.

The stress–strain relationships of different acrylic fibres vary over a wide range, depending upon the chemical and physical structures of the specific fibre. The tenacities of acrylic fibres are generally between those of standard rayon and nylon. Acrylics are sufficiently strong for all the normal apparel applications but are rarely considered for those high-strength applications in which nylon and polyester fibres serve. Acrylic fibres are usually made into staple, although some producers market continuous filament yarns, and high-tenacity acrylics are also in production.

Desirable characteristics of acrylics

Good tenacity and extension are combined with excellent elastic recovery and high initial modulus, mechanical properties contributing to the production of fabrics of good dimensional stability, which are unaffected in conditions that would tend to bring about deterioration in many other fibres. Acrylics are not significantly affected by exposure to sunlight, moisture, micro-organisms, chemicals, or solvents. Fabrics constructed from these fibres have an attractive hand characteristic. Acrylic fibres have low specific gravity, slightly higher than that of nylon, and provide lightweight fabrics that bulk well. The water absorbency of acrylic fibres is generally low, but is higher than that of polyester fibres. Acrylics rarely accumulate static electricity to the extent that polyester fibres do, and successful static elimination can be achieved with suitable antistatic finishes. The small water absorption contributes significantly to their dyeability.

Acrylic fibres tend to be heat-sensitive, and heat-setting is necessary for good stability. At very high temperatures, the fibres decompose rather than melt, but such temperatures are unlikely to be encountered in normal textile applications. Fabrics made from 100 percent acrylic fibre or from blends with 50 percent or more acrylic fibre may be durably pleated.

Acrylic goods are only moderately flammable; they have outstanding resistance to the effects of sunlight, micro-organisms, insects, and aging; and show excellent outdoor weathering resistance. They also display good resistance to all the chemicals likely to be encountered in normal textile use, including bleaches, dilute acids and alkalies, and dry-cleaning solvents.

Mod-acrylics

The modacrylic polyacrylonitrile fibres include modacrylic fibres spun from polymers consisting of less than 85 percent by weight of acrylonitrile units, but do not include those copolymers in which acrylonitrile is not the major component. Fibres in this category are spun from an extensive range of copolymers of acrylonitrile, in which the nature and proportion of the second, and possibly additional components, may vary within wide limits. The modacrylic fibres that can be produced within this broad category provide wide variations in properties, depending on their composition and method of manufacture.

Details of the chemical structure of individual modacrylic fibres are not always available. The second component of the polymer, however, is commonly chosen from vinyl chloride, vinylidene chloride, or vinylidene dicyanide.

Polyvinyl chloride (PVC) fibres. Fibres of PVC, spun from polymers or copolymers of vinyl chloride, are manufactured in a variety of forms and modifications. They have achieved a limited success in textile manufacture, but

their low softening point, often as low as 70° C (160° F), restricts their possible applications. Fibres of PVC, however, have characteristics that are highly suited to certain applications: they do not burn, and they have resistance to many chemicals. Their tendency to shrink at relatively low temperatures is useful in the production of high bulk yarns.

Vinyl chloride is produced in large quantity for the production of polyvinyl chloride plastics, and fibre manufacture uses only a small part of the output. It is polymerized typically as an aqueous emulsion in autoclaves (pressure vessels), under pressure of 45–50 atmospheres and a temperature of about 65° C (150° F). The polymer forms a suspension in the water and is recovered by spray drying. The fibres may be spun by dry-, wet-, or melt-spinning processes.

Polyvinylidene chloride (saran) fibres. In 1940 the Dow Chemical Company, of the U.S., introduced a new synthetic fibre consisting of a copolymer of vinylidene chloride and vinyl chloride that was given the generic name saran. Production of saran fibres spread to other countries, and it is now established as a specialty textile fibre.

In its production vinylidene chloride and the vinyl chloride or other comonomer are polymerized in an aqueous emulsion in the presence of a catalyst. The copolymer is melt spun through spinnerets at about 180° C (360° F), and the filaments are immediately cooled with water before being drawn to develop satisfactory tenacity. Pigments may be incorporated in the molten polymer before spinning, and titanium dioxide is used to dull the filament.

Desirable characteristics of saran

Saran is a flexible fibre, with a soft warm hand, its smooth rounded surface contributing to soiling resistance and easy dirt removal. It is unusually tough and durable, with excellent flex resistance, and fabrics woven from this fibre have excellent resistance to abrasion and hard wear.

The moisture absorption of saran is very low, and its mechanical properties are unaffected by moisture. Saran fabrics are dimensionally stable, washing and drying easily and quickly, and their low moisture absorption contributes to stain resistance; ink, food, drink, and many other stains may be removed with soap and water.

Because of saran's low resistance to heat, coupled with its negligible moisture absorption, its use in garment fabrics has been restricted. The fibre's nonflammability is an asset, and saran fabrics are popular for draperies and other application, where fire hazards are present. Saran has excellent resistance to sunlight, aging, and general weathering conditions, and is completely resistant to attack by insects and microorganisms.

Polyvinyl alcohol fibres. Polyvinyl alcohol fibres have been commercially produced in Japan since 1950, and later in South Korea, North Korea, China, France, Germany, Poland, and the U.S.S.R.

Polyvinyl alcohol is made by the hydrolysis of polyvinyl acetate, and fibres are commonly produced by wet spinning, or by a process combining features of dry and melt spinning. The fibres produced by the regular wet-spinning process are heat treated at about 240° C (450° F), followed by acetalization, usually with an aldehyde such as formaldehyde. Benzaldehyde may be used instead of formaldehyde to provide a fibre of high resilience. Aldehydes containing active groupings, such as amino groups, may be used to confer special dye affinity on the fibre.

Polyvinyl alcohol fibres have the flexibility associated with a flattened-tube type of cross section, and fabrics made from these fibres are soft and warm, with excellent hand, so that they feel comfortable when worn next to the skin. The fabrics are durable and hard wearing, with high tensile and bursting strength, and excellent impact and abrasion resistance.

The elastic recovery of polyvinyl alcohol fibres is on the low side, suggesting that dimensional stability and wrinkle resistance may not be high. Mechanical properties are not unduly affected by water, and the ability to absorb moisture contributes to the comfort of garments worn next to the skin. Fabrics made from these fibres will burn only with difficulty. The fibres are resistant to insects, microorganisms, and other influences encountered in outdoor applications. Their high resistance to acids, alkalies, and many other chemicals is an important factor in the industrial applications of polyvinyl alcohol fibres. Polyvinyl alcohol fibres are potentially very cheap, and they appear to be best suited in the field of hard-wearing, low-cost fabrics and garments.

Polytetrafluoroethylene (PTFE) fibres. In 1954 PTFE fibre was introduced by du Pont under the trademark Teflon, and such fibres are now produced by other manufacturers and in countries other than the United States, including the U.S.S.R., where it is called Ftorlon and Polifen.

Tetrafluoroethylene is purified and polymerized under heat and pressure in stainless steel autoclaves in the presence of a peroxide catalyst. After polymerization is completed, the dispersion is extruded through a spinneret, the jets emerging into an aqueous coagulating bath of a dilute solution of hydrochloric acid. The dispersion is coagulated, and the particles of PTFE hold together as weak filaments in which the particles remain entirely separate. The filaments are then heated rapidly to approximately 385° C (725° F) and maintained at that temperature for several seconds, causing the polymer particles to be fused into a coherent filament. This product is cooled quickly and drawn at room temperature to three or four times its original length.

PTFE fibres are unique in that they combine the resistance to chemicals, solvents, and elevated temperatures associated with inorganic fibres derived from mineral sources, with the flexibility and toughness typical of many organic fibres derived from animal and vegetable sources.

Because PTFE fibres are extremely costly, they are not suitable for general textile use. In addition, they have an unpleasant, greasy hand, no moisture absorption, high density, and low modulus. The outlet for these fibres is in end-uses where performance is of greater significance than initial high cost, such as filter fabrics and felts, gaskets, electrical tapes, and low-friction backings.

Fluorinated ethylene-propylene copolymer (FEP) fibres. Fibres spun from FEP are similar to PTFE fibres, except that they are thermoplastic, melting at about 290° C (550° F).

Polyvinylidene dinitrile fibres. Vinylidene dinitrile was synthesized in the United States in 1947 by the B.F. Goodrich Company. Polymers and copolymers were made that led to the development of a new type of synthetic fibre produced under the trade name Darlan, later changed to Darvan. The manufacture of Darvan was discontinued in the United States in mid-1961.

Polyolefin fibres. Polyolefin fibres are spun from polymers or copolymers of such olefin hydrocarbons as ethylene and propylene. Two types of fibre, polyethylene and polypropylene, dominate the field. Polyethylene fibres are limited to certain specialized applications; polypropylene fibres are of increasing importance, showing prospects of becoming major textile fibres in the late 20th century.

Polyethylene fibres. In the high-pressure–high-temperature process, ethylene is polymerized by heating at temperatures in the region of 150–200° C (300–400° F), with pressures of 1,000–2,000 atmospheres. The reaction is promoted by traces of oxygen or other catalysts. In the low-pressure–low-temperature process polymerization takes place at much lower pressures and at temperatures below 100° C (212° F). A variety of catalyst systems is used.

Polyethylene is extruded into monofilaments of round, flat, or other cross sections by extrusion techniques similar to those used with such other thermoplastic polymers as nylon and saran.

The commercial development of polyethylene fibres in the general textile field has been hampered by low softening point, high shrinkage, low stiffness, poor creep characteristics, and inability to take dyes. Such shortcomings restrict its use to specialized nonapparel applications.

Industrial uses of polyethylene fibres

Low-density polyethylene fibres are used for various industrial applications, including ropes and cordage, filtration fabrics, and protective clothing. They are used in the form of resilient shrinkable yarns for the production of three-dimensional fabrics. High-density polyethylene fibres are used in aerial tow targets, high-altitude balloons, and in equipment designed for Arctic conditions, in which the retention of strength and flexibility at low temperatures

is invaluable. The strength, lightness, and water- and rot-resistance of polyethylene fibres have established them in the marine cordage field for the production of ropes and nets that float, do not rot, and do not absorb water. Furniture fabrics, car upholstery fabrics, curtains, protective clothing, tarpaulins, and filter fabrics are other applications making use of the special properties of polyethylene fibres.

Polypropylene fibres. The specific conditions under which polymerization of propylene is carried out are proprietary. Catalysts are used, such as an organometallic compound of aluminum in the presence of titanium tetrachloride, and the reaction is carried out under pressure at about 100° C (212° F).

Polypropylene fibres are made by extrusion of molten polymer, followed by drawing to orient the molecules and crystals in the filaments. Polymer of high viscosity is spun to obtain optimum fibre properties. The filaments are cooled in air as they emerge from the spinneret and are then collected on bobbins.

Polypropylene monofilaments have applications similar to those of polyethylene, but offer higher strength, increased toughness, resilience, abrasion resistance, and creep resistance, and a higher melting point. These properties are allied to their water-resistance and chemical inertness. Polypropylene monofilaments have become established in a number of fields where they compete successfully with other fibres for such uses as ropes and cordage, fishing nets and twines, filter fabrics, and protective clothing.

The introduction of fine-denier multifilament yarns and staple fibre extended the range of potential applications for polypropylene fibre to the textile and apparel fields, bringing them into competition with established natural and synthetic fibres in applications where the shortcomings of polypropylene fibres often place them at a disadvantage. Despite the attractive characteristics of polypropylene fibres, less desirable properties, such as dyeing difficulties and relatively low melting points, have tended to limit their application.

Polyurethane fibres. Polyurethane fibres are spun from polymers made by a reaction taking place between small molecules, in which the linkage of the molecules occurs through the formation of urethane groups. Linear polyurethanes may be made by reaction of a glycol with a diisocyanate.

In recent years, polyurethanes have achieved increasing importance in textiles, becoming the basis of a novel type of elastomeric fibre known generically as spandex. Spandex fibres are spun from segmented polyurethanes made by a series of chemical stages, beginning with production of a low-molecular-weight polymer, or prepolymer, followed by a reaction of the prepolymer with diisocyanate and the coupling of isocyanate-terminated prepolymer to form segmented polyurethane.

Spandex

The technique used in spinning spandex fibres depends upon the type of polymer spun. Some segmented polyurethanes are essentially linear molecules, soluble in solvents. Other segmented polyurethanes may be insoluble branched or cross-linked structures.

Soluble polyurethanes are dissolved in an appropriate solvent, and the solutions may be extruded through spinnerets into a coagulating bath (wet spinning), or into an atmosphere that removes the solvent (dry spinning). The techniques are essentially the same as those used for spinning hard synthetic fibres, with allowance being made for the elasticity of spandex fibres.

When the molecule of polyurethane is allowed to grow into a three-dimensional structure, it is insoluble and cannot be spun by either wet or dry processes. Instead, a chemical spinning process may be used. The isocyanate-terminated prepolymer is spun at a stage when it forms a viscous melt, the jets emerging into an environment containing a chain extender that diffuses into the fibre and reacts to link the prepolymer molecules into their final form, thus producing the branched or crosslinked polyurethane in fibrous form.

Because of their segmented structure, spandex fibres may be made stronger than rubber fibres. To achieve a given stretch, spandex fibres require a force twice as high, over the entire range of elongation, as the force required by rubber. In contrast to hard fibres, such as nylon, spandex fibres have a very low modulus of elasticity. Spandex fibre has about twice the effective power of rubber at the important elongations at which additional stretch of 50 to 100 percent is available.

The durability of spandex fibre in garments, in which resistance to flexure, abrasion, and needle-cutting are necessary, is high. Spandex-fibre fabrics show high resistance to flex and low growth compared to their rubber-fibre counterparts. The durability of spandex fibres allows production of very fine, or high denier, yarns that have opened up new applications for elastic fibres. In addition, spandex fibres were the first elastomeric fibres that could be readily dyed.

INORGANIC SYNTHETIC FIBRES

Fibres produced from inorganic substances include glass, aluminum silicates, metallics, and various other types having special industrial applications.

Glass fibres. Glass fibres are spun from glasses of many different compositions depending on end use. Silica sand (silica) and limestone (calcium carbonate) are the basic ingredients to which are added varying amounts of other materials such as soda ash (sodium carbonate), potash (potassium carbonate), aluminum hydroxide or alumina (aluminum oxide), magnesia (magnesium oxide), or boric oxide.

The ingredients are charged into a furnace, where they are fused at high temperature to form molten glass. Filaments may be spun directly from this melt, or the glass may be formed into marbles of about 16-millimetre (0.73-inch) diameter that are passed to the spinning machines, also called fiberizing units.

Continuous filament strands are produced by allowing molten glass to flow through perforated tips on the baseplate of a platinum melter, or bushing, drawing away a stream of glass from each tip to form a fibre, and attaching the fibres to a high-speed winding device. The bushings may be mounted on glass-melting tanks, or may form the basis of independent fiberizing units, being fed with glass by the remelting of glass marbles.

As the 50 or more filaments are drawn from the orifices, they are brought together into a strand, and a lubricating size, or finish, is applied to facilitate subsequent processing. Strands intended for use in textile weaving may be converted through conventional textile processes into suitable yarns. There are various methods of producing glass staple fibre; the most important include the centrifugal, jet, and rod-drawing processes.

In the centrifugal process, molten glass is thrown out of holes in the base of a metal spinner rotating at high speed. The resulting fibres are bonded into a web for use in heat and sound insulation. This technique is not generally used for production of textile-grade glass fibres.

Centrifugal process for glass fibres

In the jet process, staple fibre is produced from molten glass that flows under gravity through holes in the base of platinum bushings, as in the production of continuous filament. The streams of glass are drawn into fibres by the action of high-speed streams of turbulent gas or steam, and the turbulence breaks the solidified fibres into staples of 200 to 380 millimetres (8 to 15 inches) that are collected into a web on a revolving vacuum drum. The web may then be guided from the drum and formed into a glass staple sliver, a continuous strand of loosely assembled fibres that is processed on conventional textile machines for weaving into glass-fibre staple fabrics.

The rod-drawing process is a modern development of the traditional process for making glass filaments. In a typical modern machine, 125 glass rods, usually about four millimetres (0.2 inch) in diameter, are mounted vertically and adjacent to one another in a kind of spinning frame. The rods are kept moving slowly downward, and are simultaneously melted at a temperature of about 1,200° C (2,200° F), either by individual movable and adjustable gas burners or by electric heating coils in a fireclay chamber. Drops of glass fall away from the ends of the rods, drawing glass filaments after them.

Glass is a heavy fibre, with a specific gravity of the order of aluminum, and has good transparency. Its moisture ab-

sorbency is negligible, and the fibre shows no swelling or shrinkage. This latter characteristic is useful in applications in which dimensional stability in the presence of water is desirable, and in electrical applications, but it is detrimental in apparel, where absorption of moisture is useful.

Water affects the tensile properties of glass fibre, bringing about a reduction in tenacity after a very short time, caused by the leaching out of some of the soluble materials from the glass, including alkali silicates; glass with a low alkali content is most resistant to this deterioration.

The high coefficient of friction of glass against glass, taken in conjunction with the high flexibility of fine glass filaments, results in poor abrasion resistance. The movement of filament against filament during use brings about fracture of the filaments, creating a hairy fabric. Suitable lubricants and finishes can minimize such effects, and metal coatings may be applied to glass fibres to reduce surface friction.

Glass fibre has the highest strength-to-weight ratio of any fibre, and one of the lowest elongations. These characteristics are useful in applications requiring high dimensional stability, as in the reinforcement of plastics. The low elongation is not useful in apparel fabrics, in which resiliency is an important factor. Glass fibres exhibit almost perfect elasticity, returning instantly and completely to their original dimensions upon release from strain. This characteristic, however, is exhibited over only a very small range.

Resistance of glass fibre to heat

Glass fibre has excellent resistance to the effects of heat over a wide temperature range, with fabrics showing an increase in strength up to about 205° C (401° F), after which strength and flexibility begin to fall. At 370° C (700° F), glass filaments retain 50 percent of their original strength; at 538° C (1,000° F) they retain about 25 percent. The effect of temperature depends greatly upon the composition of the glass.

Glass is completely nonflammable and is used where resistance to the spread of flame is of overriding importance. The high heat conductivity of glass is a useful attribute in electrical-insulation applications, where close-woven glass fabrics dissipate heat rapidly. Despite this quality, glass fibre is useful as a heat insulation material and is employed in the form of mats and waddings in which entrapped air provides the insulation.

Glass has high resistance to most chemicals but is disintegrated by alkalies. Glass fibres may be attacked by certain strong mineral acids and by phosphoric acid, but show complete resistance to biological degradation.

Aluminum silicate fibres. Fibres are spun from aluminum silicate, with or without the addition of minor amounts of other materials. Aluminum silicate fibres were developed originally for use in jet engines to provide high-temperature-resistant, lightweight, strong, fibrous material of low thermal conductivity. These fibres are also used for such products as high-temperature filters, packings, gaskets, and for thermal, electrical, and acoustical insulation.

In the production of aluminum silicate a mixture of equal parts of alumina and silica with a small proportion of sodium borate added to produce short staple, or zirconia to produce long staple, is fused in an electric furnace. The molten aluminum silicate is poured from the furnace in a small stream. As it falls, the stream disintegrates into tiny droplets or particles that are attenuated into fine fibres as they are blown through the air and collect on a mesh screen, forming a mat of fibre.

Aluminum silicate fibres are used largely for insulation purposes, filling a gap that exists from about 400° C (about 750° F), where glass and mineral wool are often inadequate, to about 1,260° C (2,300° F). Aluminum silicate fibres can be used for short exposures at higher temperatures, with a slight loss of resilience and some shrinkage. The devitrification taking place at these temperatures does not affect the insulating properties of the fibres. Extremely low thermal conductivity is an outstanding feature of aluminum silicate fibres.

Aluminum silicate provides a lightweight, high-temperature heat barrier with excellent thermal shock resistance, flame resistance, resilience, chemical stability, and electrical properties. Because the fibres of aluminum silicate interlock, there is no brittle structure to develop stresses during sudden heating or chilling.

Aluminum silicate fibres are used for a variety of high-temperature applications, including thermal, acoustical, and electrical insulation; filters, packings, and gaskets; and containment and conveyance of molten nonferrous metals.

Metallic fibres. Fibres produced from metals may be used alone or in conjunction with other substances. Those used in modern textile manufacture are of two main types: (1) metallic single-component, or SC, fibres and (2) metallic multicomponent, or MC, fibres.

Metallic single-component fibres. Late in the 1950s, the development of high-speed flight and space travel created demands for flexible materials capable of withstanding unique environmental conditions for use in the manufacture of protective garments, antenna membranes, parachutes, and other space-program structures. It was determined that the requirements for such applications could be met by the production of fibres of adequate flexibility, made from high-temperature metals or alloys; *e.g.*, of the stainless steel type.

Applications of SC fibres in space technology

Methods of producing metal filaments of the required fineness were developed in the 1960s, and metal fibres have since become available in quantities and at a cost entitling them to consideration as genuine textile fibres.

Metallic SC filaments are produced by drawing either single filaments or multifilament strands. Single-filament drawing produces excellent quality metallic fibres, but multifilament drawing offers the best possibility of reducing cost and increasing availability. One method of drawing multifilament yarns involves enclosing a bundle of two-mil (0.002-inch, or 0.05-millimetre) wires in a sheath of a dissimilar alloy. The sheathed bundle is drawn to the required extent, and the sheath is removed with nitric acid, leaving the multifilament yarn ready for sizing, warping, and weaving.

Metals possess a number of general characteristics that can be of practical use in textile applications. The potential high yield and fracture strengths of most metals allow production of fabrics exceeding in strength those that can be produced from existing organic or glass fibres. The absence of a significant viscous behaviour characteristic in metal fibres ensures dimensional stability in certain structures that would be difficult to achieve with conventional textile materials.

The most significant difference between organic fibres and metal fibres is in the rigidity or stiffness of the material. Most organic textiles have a stiffness less than one-twentieth that of most metals.

This high stiffness previously excluded metals from use as textile materials because it placed severe limitations on the processability and end uses of the fibres. Adequate flexibility has been achieved, however, with the advent of modern ultrafine metal fibres.

The higher weight of most metals, in comparison with organic materials, can be a serious disadvantage in the use of metal fibres for many applications. On the other hand, the resistance of metals to chemicals at room and elevated temperatures covers ranges generally exceeding those of organic fibres, thus allowing metal fibres to be employed in many military and industrial end uses in which specific chemical resistance is a requirement. In some applications, however, metals may suffer corrosion by chemical action.

Many metals can tolerate and endure temperatures higher than those that organic fibres will generally withstand. Most metals are good conductors of heat, and metal fibres may function as heat-sinks when used in adequate proportion in blends. The high electrical conductivity of most metal fibres provides an effective method of eliminating static electric charges.

Metallic multicomponent fibres. Multicomponent fibres are made from metal in association with other materials, commonly plastics. Modern metallic fibres of the multicomponent type are based largely on aluminum, which provides sparkle and glitter at a fraction of the cost of the early types of decorative fibre based on such substances as gold.

The aluminum in these fibres is in the form of a narrow

ribbon-filament of either metal foil, or a plastic film that has been vacuum-plated with vaporized aluminum and coated with a plastic film. In these composite structures, the metal is protected from corrosive influences of its environment and from mechanical damage. Multicomponent metallic fibres have achieved popularity as decorative fibres and are an important part of the modern textile industry.

Metallic filament "sandviches"

Modern metallic filaments are made by sandwiching a thin layer of metal, usually aluminum, between thin sheets of the appropriate types of plastic film. The lamination may be carried out in several ways. In the production of the acetate butyrate aluminum-foil metallic fibre, for example, a thermoplastic adhesive, either coloured or clear, is applied to both sides of a sheet of aluminum 0.01-millimetre (0.00045-inch) in thickness. The coated foil is then heated to about 90° C (200° F) and passed through a set of rollers, together with two sheets of cellulose acetate butyrate, in such a way that the aluminum foil becomes the centre of the "sandwich." The laminated material is then slit into narrow ribbonlike filaments of the desired width; *e.g.,* 3.2 to 0.20 millimetres.

In some types of metallic filament, the centre of the sandwich consists of a plastic (*e.g.,* polyester) film that has been vacuum-coated with aluminum, and this may then be protected by the application of layers of lacquer or plastic film as above.

Many types of coloured metallic filament are produced, including those with the metallic glitter of the aluminum shining through the coloured adhesive and those having coloured outer plastic films. Gold-coloured filaments are made by using an orange-yellow dyestuff or pigment in the adhesive; a silver colour requires only the natural glitter of the aluminum.

Metallic MC yarns are used almost entirely as decorative materials for the textile industry. The aluminum foil that provides the glitter in a modern metallic yarn is protected from any ambient corrosive materials by the plastic film in which it is enclosed. Polyester types will withstand repeated launderings without losing sparkle. Because metallic yarns are not affected by seawater or the chlorinated water of swimming pools, they are widely used in swimwear. The dyestuffs used in colouring metallic fibres are usually fast to light, thus retaining bright colour.

Because metallic MC yarns are used primarily for decorative purposes, they do not usually contribute significantly to the strength of fabrics or garments. They may, however, be used as weft or warp yarns, and are sufficiently strong to withstand weaving and knitting operations. When necessary, metallic yarns are combined with such supporting yarns as nylon. The plastic film of the metallic yarn is flexible, and the yarns are extensible to a degree depending upon the type.

Aluminum corrodes and tarnishes in air and in contact with seawater, but in metallic fibres it is protected so effectively that it retains its glitter for long periods. The chemical resistance of a metallic filament is generally equal to the chemical resistance of its coating of plastic film. In the case of polyester films, this is outstanding.

If metallic fibres remain in contact with strong alkaline solutions for prolonged periods, the aluminum may be attacked at the unprotected edges of the ribbon. Such fibres should not, therefore, be subjected to alkaline reagents of significant strength. Organic solvents also may attack the laminate adhesive or lacquer coating, and care is required in dry cleaning to ensure the use of an appropriate type of solvent. The plastic films in metallic fibres are thermoplastic, softening at elevated temperatures, and delamination may occur if the fibres are heated. Acetate types in particular should be processed only at low temperatures.

The plastic films may be permanently embossed by heat and pressure, thus allowing the introduction of special effects.

Miscellaneous inorganic synthetic fibres. Fibres play an increasingly important role in engineering, where they are used as constituents of composite materials. Embedded in a matrix, which may be plastic, metal, or ceramic, fibres contribute their unique properties to the composite to function in a manner similar to the steel rods embedded in reinforced concrete.

Composites

Composite materials containing fibres are used in applications in which high strength and low weight are needed, as in aeronautics and astronautics. Fibres retaining high strength at elevated temperatures are valuable in high-temperature applications. Since World War II, many types of fibre have been used in composite materials, and the demand for new materials to meet the demands of modern engineering applications has stimulated development of new types.

Composites are heterogeneous structures consisting of a continuous matrix in which the fibrous material is dispersed and embedded. Efficient performance of the composite requires a strong bond between the fibre and the matrix. Techniques used in producing composites from fibres and matrices depend upon the nature of the materials and the purposes for which the composite is required.

Fibres may be in the form of continuous filaments, masses of staple fibre, yarns formed from continuous filaments or spun from staple fibre, or fabrics produced from yarns.

Continuous filaments may be used to build structures by a winding technique. The structure is built up by winding the filament around some kind of form, the structure being held together by a matrix as it is created. This technique is used in constructing pressure vessels and rocket bodies.

Staple fibre is commonly mixed with the matrix; *e.g.,* in the production of reinforced plastic materials. Yarns and fabrics may be used in layering processes, and glass-fibre-reinforced plastics are used in building boat hulls and car bodies.

Carbon and graphite fibres. Because the element carbon has the ability to form strong bonds between its atoms, carbon atoms hold together in strong three-dimensional structures, such as that of diamonds, and in two-dimensional structures, as in graphite, and they also hold together in strings, forming long, single-dimension polymer structures in the same way that molecules form the basic structural units of fibres.

The fibres used in textiles are based almost entirely on organic polymers; the molecules consist of long chains of carbon atoms, with other atoms or groups of atoms attached as appendages to the carbon backbone. By removing the appendages from an organic fibre molecule it is possible to produce long molecules consisting entirely of carbon atoms. When a chemical operation of this sort is carried out on a highly oriented organic fibre, carbon fibres can be obtained in which the molecules retain a high degree of orientation.

Production of carbon fibres

Such carbon fibres may be produced by treatment of organic fibres under carefully controlled conditions. Cellulosic fibres, such as viscose rayon, or acrylic fibres are the favoured starting materials, and acrylic fibres are widely used in the commercial production of carbon fibres.

The nature of the carbon fibre produced depends upon the conditions under which pyrolysis, the subjection to the action of heat, is carried out. At lower temperatures, carbon fibres are formed; at higher temperatures, the carbon atoms in the fibre are arranged in the crystalline form of graphite.

Carbon fibres have a high strength-to-weight ratio, as well as high modulus and stiffness. They are thermally stable, retaining high strength at temperatures where other engineering materials suffer significant loss of strength. Graphite fibres are softer and have high electrical conductivity.

Composite materials containing carbon fibres are used in making the compressor turbine blades in jet engines and in the construction of aircraft and spacecraft. They are used in deep-sea submersibles, in heavy-duty bearings, and in pressure vessels. In spacecraft re-entry shields, carbon fibre composites are able to retain their structural form at the high temperatures developed.

Boron and boron-derivative fibres. The development of boron fibres has provided the engineering industry with high-modulus continuous filaments used in high-performance composites. Boron filaments may be made by depositing the element boron on a fine core filament of a conducting material such as tungsten or carbon. Boron trichloride is passed over such a filament, which is then

heated electrically, thus causing the boron trichloride to decompose and deposit boron on the core filament.

Boron nitride fibers are commercially useful because they have properties similar to those of graphite fibres, including high strength, low weight, and high elastic modulus. They conduct heat readily, and are thermally stable, but differ from graphite fibres in being good electrical insulators.

Refractory oxide fibres. In modern engineering, the demands made upon materials are of increasing stringency; materials are required to retain such properties as high strength when used at high temperatures and under corrosive conditions. These requirements have stimulated great interest in production of fibres from refractory oxides and similar materials. Alumina or silica, for example, may be used at high temperature and in an oxidizing environment. Many such materials are difficult to spin, however, because of their comparatively sharp melting temperatures and narrow viscosity ranges.

Special techniques have been devised to meet particular problems. Alumina may be spun by packing it into a tube of fused silica. The end of the tube is melted, and the fibres drawn from the molten material consist of crystalline alumina filaments enclosed in silica sheaths.

Silica-carbon fibres may be made by drawing continuous filaments from a solution of viscose rayon and sodium silicate. The filaments are pyrolyzed (heated to a high temperature) in nitrogen, to form a silica-carbon fibre that may be converted to a silicon carbide fibre by heating at temperatures above 1,500° C (2,700° F).

Whiskers. Single crystals of extremely high strength, called whiskers, have been the subject of scientific investigation for many years. They are a form of fibre, and have potential importance in the fibre composite field.

Properties of whiskers

Whiskers produced from alumina and silicon have very high elastic moduli and do not exhibit plastic deformation. Used in composites (*e.g.*, with resins, metals, or ceramics), whiskers provide engineering materials having unique and valuable characteristics. Titanium compressor blades reinforced with silicon-carbide whiskers, for example, retain high strength at elevated operating temperatures.

ECONOMIC ASPECTS

History of fibre development

From the original production of Chardonnet silk in the 19th century, synthetic fibres have had a large and growing economic impact throughout the world. Before World War II, almost all man-made fibres were derived from cellulose, in the form of viscose or acetate. During the 1930s, as a result of intensive fibre research, several new synthetic fibres were produced experimentally. In the late 1930s these tentative steps led to the production of nylon, the first commercially successful synthetic-textile fibre.

Since that time, the development of synthetic-fibre production has continued. The years following World War II were marked by the introduction of new types of synthetic fibre; nylon was joined by the polyesters, acrylics, polyolefins, and others as fibre producers worked to establish themselves in the new technically based industry. Some of the new types of fibre have achieved high acceptance; others have not lived up to their early promise and have disappeared.

Because the commercial development of a synthetic fibre requires substantial expenditure on research, technology, and plant construction, a commitment to produce a particular type cannot be easily changed to produce another of marginally better properties. Manufacturers, who must recoup their original investments, work to improve the chosen fibre. A stage is commonly reached in the development of a new technologically based industry where consolidation becomes necessary, and the synthetic industry reached this stage, to a large degree, in the 1960s. Producers who had committed themselves to the production of polyamides, polyesters, acrylics, and other commercially accepted fibres concentrated much of their attention on the improvement of basic products. There was rapid commercial development of techniques increasing the versatility of established fibre types. The development of texturizing, chemical crimping, multicomponent fibres, and film fibres extended applications of synthetic fibres, and the production of polynosic, or high wet modulus, and other modified rayons stimulated interest in the natural polymer fibres field.

Several new fibres were introduced, however, during this consolidation period. Polyurethanes and polypropylenes captured a substantial share of the textile market, and other new fibres were in the experimental stage.

Man-made fibres are produced almost entirely from organic polymers, constructed from various carbon derivatives. Cellulose is the basic material from which the most important natural polymer fibres are made, and, although ample supplies of cellulose are available, conversion to a form suitable for fibre production presents certain problems. The raw materials of synthetic fibres, however, are simple organic chemicals that come from stores of available carbon compounds. Initially, the industry relied heavily upon coal for such materials, but petroleum has largely superseded it. The rapid development of synthetic-fibre production has made heavy demands on the resourcefulness of the petrochemicals industry.

As world population expands, the demand for textiles may be expected to increase. In addition, a need for greater versatility in textile materials is likely. The man-made-fibre industry is expected to supply an increasingly large share of the total world fibre output, and because man-made fibres can be tailor-made to meet particular needs, a portion of the increased production will consist of specialized types. The greatest production, however, will be in "bread-and-butter" fibres, such as polypropylene, that can be produced cheaply to provide the kind of textiles that have long been produced from cotton. (J.G.Co.)

Paints and varnishes

Paint, a fluid suspension spread in thin coats to decorate and protect surfaces, consists of pigment or colouring matter, and the vehicle in which the pigment is suspended. The function of the vehicle is to form a tough film when applied to a surface and to bind the pigment to the surface. Paint may be applied to metal, wood, stone, paper, leather, cloth, or other surfaces.

Varnish is a liquid coating material containing a resin that dries to a hard transparent film. Though usually clear, varnishes occasionally may contain pigment. Lacquer is a varnish that solidifies by evaporation of the solvents contained within it rather than by one of the more complicated processes of film formation described later in this section. Lacquers may be clear and tinted, or opaque and coloured.

HISTORICAL DEVELOPMENT

Origins. Paints were in use for representational and decorative purposes for thousands of years before the idea of using them as protective coatings appeared. The earliest known paintings, found in the caves of Lascaux, France, and Altamira, Spain, made with iron oxide and applied without binder, date from as early as 15,000 BC. Early peoples of Africa, Oceania, and the Americas also used paints to decorate temples and dwellings. The Egyptians prepared colours from soil and by 1500 BC imported such dyes as indigo and madder to make blue and red pigments. By 1000 BC they had developed a varnish from the gum of the acacia tree (gum arabic) that contributed to the permanence of their art.

First known paintings

Asian art appears to have begun independently, with coloured crayons as pigment and clay as a binder. Natural ores served as sources of the first pigment, but calcined (fired) mixtures and organic pigments were developed at least before 6000 BC. Vehicles were prepared from gum arabic, egg white, gelatin, and beeswax. The use of lacquer in China goes back to prehistoric times. During the Chou dynasty (*c.* 1111–255 BC), it served for the decoration of carriages, harnesses, and weapons. By the 2nd century BC, Chinese buildings were decorated with lacquer both inside and outside; lacquer was also extensively used in Korea and Japan.

The first use of protective coatings was by the Egyptians, who employed pitches and balsams to seal ships, a practice followed by other peoples of the ancient world.

Not until the Middle Ages did paint as a preservative for

exposed wood surfaces have considerable use. The paints developed then were handmade, using such costly raw materials as egg white; the craftsmen kept their formulas secret and their products were expensive. Medieval city streets, nevertheless, bore considerable evidence of the painters' art, with colourful tradesmen's signs and shop facades painted red or blue. In the 17th century white lead paint became widely available; yet ordinary houses and even such important structures as bridges remained unpainted for the most part until the 18th century, when there was an increased availability of both vehicles and pigments. Extensive exploitation of linseed oil from the flax plant and pigment-grade zinc oxide produced a rapid expansion of the paint-manufacturing industry. In the 19th century for the first time the two ingredients, pigment and vehicle, were brought together before the paint was marketed.

Emergence of paint science and technology. The 20th century brought an enormous proliferation of articles requiring protective coating; a corresponding proliferation of paint products was achieved by intensive research efforts. The dye industry contributed the knowledge needed to upgrade and evaluate pigments, and the plastics industry contributed the development of polymers, the giant molecules containing many repeating basic molecules.

The study of heat energy, or thermodynamics, begun in the 19th century, proved a useful tool for understanding chemical reactions of paint components and the more subtle interactions leading to compatibility and pigment wetting. A newer discipline, rheology—the science of the deformation and flow of matter—had a great impact; chemists and engineers appeared in paint-industry laboratories.

New coating materials

Many new types of coating materials have evolved: resins, solvents, plasticizers, pigments, driers, foam-control agents, and adhesion promoters. Fire retardancy, corrosion resistance, and heat stability have been achieved with selected new materials and new methods of application. The art of paint making has become a science.

The most significant change in paint technology in this century is the return to aqueous systems (water-base paints) at a level of sophistication and complexity never conceived by the Egyptians. The industry is stable; the raw materials are virtually unlimited; and the future portends a small but steady growth rate, notwithstanding competition from plastics, decorator fabrics, and ceramic tile.

PAINT MANUFACTURE

Composition. Paint is composed of a pigment that hides the substrate (base on which it is applied) and imparts colour, and a vehicle, or binder, which carries the pigment onto the substrate and binds it there. Inorganic pigments are insoluble, metallic compounds the colours of which depend mainly on the electronic energy levels of the metal. Organic pigments are compounds containing colour-producing (chromophoric) groups of atoms within their molecules. Manufactured in batches of up to 9,500 litres (2,500 gallons), binders are composed of polymeric resins, usually dissolved in a solvent. A resin is a high molecular weight organic compound, soluble in organic solvents but not in water.

Dispersion. Paint manufacture involves bringing pigment particles and vehicle into close contact and achieving a thorough dispersion of the particles throughout the binder. Dispersion is in fact the most important operation in a paint factory; the economics of plant operation are keyed to it. The design of dispersion mills depends on the type of paint manufactured and on the formulations and colours the equipment must handle. The most popular types are so-called ball mills, which grind by tumbling heavy metal or ceramic balls through the paint in a cylindrical container, and sand grinders, which circulate a suspension of sand in paint through a rotor assembly at high speed. A fine grind can be achieved with ball mills operating for eight or more hours on single batches or with sand grinders delivering a nearly continuous output.

Paints are no longer hand mixed. Most of the important properties of a paint depend on the nature and extent of the interfaces (boundary surfaces) between the pigment particles and the vehicle. If the particles are not completely dispersed, with all the pigment particles wetted by vehicle, they link together in a reticulated structure that leads to poor brushability, low gloss, and the appearance of oversize particles in the film.

Oil absorption of a pigment

Oil absorption of a pigment, the volume of oil per unit weight of dry pigment needed to form a stiff coherent paste, indicates how well the pigment will disperse. Oil absorption is determined experimentally.

For each paint, there is an optimum ratio between the amount of pigment and the amount of binder. In the industry, this ratio is called critical pigment volume concentration (CPVC). As this critical figure is approached by adding more binder, gloss and blistering tendency decrease markedly, whereas permeation (penetration) of water and oxygen increase.

Control. Paint is usually made in batches. But there is a trend toward automation, especially in plants that make one product on a given line, even though continuous manufacture of paint is difficult (see Figure 15). If a large variety of colours or types must be made, the scheduling and control of the various operations become more important than the convenience of automation.

Drawing by D. Meighan based on material courtesy of E.F. Wahl

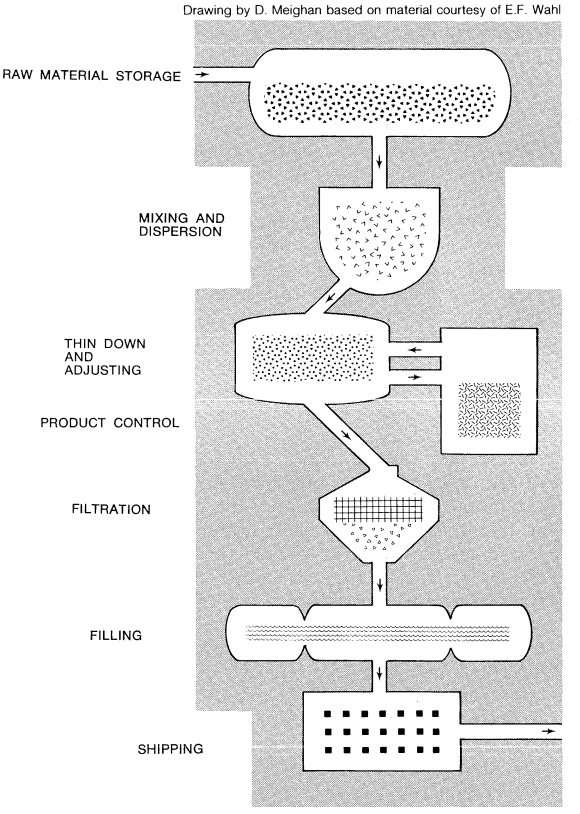


Figure 15: Stages in the paint manufacturing process.

Colour adjustment. Paint products must be uniform in colour, flow, and other physical properties. Driers, fungicides, and tint bases are added during the process of thinning, and the paint is checked against a previously established colour standard. Skilled colour matchers have never been replaced in this phase of manufacture, although instruments are used to quantify the colour values of hue, brightness, and chroma (purity). As a rule, white pigments and extenders are used to provide hiding power by scattering light (see below), and the hue is provided by coloured pigments and toners that absorb light of certain wavelengths. By contrast, tint base used in the adjustment stage is a neutral colour that possesses good hiding qualities.

The technical aspects of the film are established instrumentally insofar as possible, and the observations are recorded instrumentally; but the final check is made by eye.

PIGMENTS AND OTHER COLORANTS

Pigment is the universally accepted term for the particulate matter dispersed in liquid or solid binders. Its two

main functions are to impart colour and hiding and to improve hardness and durability.

Colour and hiding. The absorption of certain wavelengths of light gives rise to colour. The colour is the complement of the wavelengths absorbed.

When light strikes a pigment particle, the portions that are not absorbed are diverted in two ways, both of which result in hiding or obscuring the substrate. The light not absorbed may be reflected from the pigment surface or refracted as it passes through the pigment. Reflected light does not reach the substrate and therefore cannot illuminate it at all. Refracted light, on the other hand, may reach the substrate, but it would do so at such an angle, and over such a tortuous path, that the diffuse reflection would not reach the eye. Reflection and refraction depend on a mismatch of refractive index between the pigment and the binder to scatter the light. Absorption is a characteristic that is present in dyes as well as pigments and that does not rely on light scattering. Hiding power has often been referred to the difference in refractive index between pigment and vehicle, but this difference plays a role in scattering.

Colour-defining systems

Colour is difficult to quantify. The best known colour-order system, the Munsell system, complete with ordered collections of colour chips, was originated in 1915 by an American artist, Albert Henry Munsell, and has since been refined. Another system in wide use that does not depend on collections of physical samples and is most often used in connection with instruments is the CIE (Commission internationale de l'éclairage) system, based on the matching of colours by mixing coloured lights from standard sources and providing a standard observer. In reducing the colour description to a set of quantitative terms, suitable for use in a computer, this system employs a set of standard coordinates, called tristimulus values. These values are used to match and guarantee the uniformity from batch to batch of one of the most important properties of a paint—its colour. For further information on the major colour-order systems, see the article COLOUR.

Pigments. Pigments are classified according to type or colour. Early classifications distinguished between mineral, or natural, pigments, which came from the earth and were either ground directly, or refined to a certain extent; and chemical pigments, which required a conversion of one sort or another. It is more convenient to classify pigments by colour, starting with the whites, then listing the blacks, and following with the more popular primary hues.

White pigments. White lead, the basic carbonate ($2PbCO_3 \cdot Pb[OH]_2$), has been manufactured for over 300 years by subjecting lead sheet to acetic acid fumes and carbon dioxide. The process takes several months.

Basic sulfate of lead ($PbSO_4 \cdot PbO$) is manufactured by heating lead in contact with sulfur. Reaction occurs when both the lead and the sulfur are vapourized. The basic silicate ($PbO \cdot SiO_2$) is formed as a coating on silica.

Zinc oxide (ZnO) is the product of the reaction of zinc vapour with oxygen. Zinc sulfide (ZnS) is formed by precipitation out of aqueous solution and heating to a high temperature.

Lithopone ($ZnS + BaSO_4$) is a composite material, formed by precipitating zinc sulfate ($ZnSO_4$) and barium sulfide (BaS) together. Antimony trioxide (Sb_2O_3) is an exceptionally white pigment obtained by roasting antimony ore in air.

Titania, or titanium dioxide (TiO_2), is the superior pigment from the standpoint of hiding capability. Its preparation requires the removal of iron from the ilmenite ore $FeTiO_3$. Titanium dioxide is heated to a high temperature with small amounts of aluminum, antimony, silicon, or zinc oxide to produce white pigments with improved texture, resistance to light and moisture, and ease of dispersion in specific paint vehicles.

Black pigments. Most black pigments are made of elemental carbon. Carbon blacks are formed by allowing a smoking flame of natural gas to impinge onto a metal surface, usually iron. In the furnace process, the flame is not made to impinge on a surface but simply burns freely; the carbon is precipitated from the cooled gases. Lampblacks are obtained by burning oils in regulated amounts of air and allowing the carbon to fall by gravity from the combustion gases. Their blackness is inferior to carbon blacks, and they produce finishes of lower gloss. Bone blacks are made by charring bones. The char, which forms in the absence of air, consists of only about 10 to 20 percent carbon. Bone blacks are used principally in tinting when compatibility with other pigments is required. Graphite occurs naturally or can be prepared from coal in electric furnaces. It is an inferior pigment used only for specialty finishes. Amorphous graphite occurs in nature and is available at low cost. Mineral blacks derived from shale, peat, and coal dust are used as inexpensive pigments in undercoats.

Obtaining carbon black and lampblack

In addition to the carbon-containing black pigments, there exist iron oxide blacks found in nature or prepared. Blue basic lead sulfate, called blue lead, is a gray pigment that is used in the priming (*i.e.,* first coating) of structural steel.

Coloured pigments. It is convenient to group coloured pigments according to their position in the visible spectrum. Mineral varieties predominate at the red and blue ends and organic varieties in the centre.

Among the most common red pigments are red lead, iron oxide, cadmium reds, cuprous oxide, and various organics. Red lead (Pb_3O_4) is usually applied to steel for its protective value and is not used for tinting. This pigment is usually made by heating lead monoxide, or burning metallic lead. Iron oxide earth pigments (Fe_2O_3) are naturally occurring end products of the weathering of silicate rocks and were used in prehistoric times. The more important of these earth pigments are classified as ochres, the colorant of which is exclusively iron oxide and which tend toward the yellow part of the spectrum; siennas, which contain from 30 to 75 percent iron oxide and range from yellow-brown to reddish-orange depending on the variety and the degree of calcination; and umbers, which have a high manganese content in the form of manganese dioxide and tend toward brown coloration. Iron oxide ores, some of which can be ground directly to produce bright-coloured red pigments, include hematite ores, such as Spanish oxide and Persian Gulf oxide; magnetite (black); limonite (yellow); siderite (brown); and pyrites (dark brown). Synthetic iron oxide can be derived from ferrous sulfate ($FeSO_4$) by means of: calcination (heating to high temperatures) to produce grades ranging from light Turkey reds to dark Indian reds; calcination in the presence of lime (CaO), which produces a mixture of ferric oxide and calcium sulfate ($CaSO_4$) known as Venetian red; and precipitation from aqueous solution, which can produce any of three iron oxides of different hue.

Miscellaneous inorganic red pigments include cadmium reds, which rely on the mixture of cadmium sulfide (CdS) and selenide (CdSe) to produce colours ranging from light red to maroon, and cuprous oxide (Cu_2O), which starts out as a bright red pigment but gradually oxidizes to the cupric form (CuO), which is black. Because cuprous oxide is a fungicide, this pigment is used in spite of its inherent chemical instability.

Organic red pigments include naturally occurring organic pigments, such as carmine and madder (mainly of historical interest, though still used as artists' colours), and synthetics. Synthetics are high-cost items, used because of the infinite variety of hues that can be achieved by minor changes in molecular structure.

Yellow and orange pigments

Yellow and orange pigments are obtained from metallic salts and from organic compounds. Chrome yellows and oranges are low-cost, bright, and stable pigments that display a considerable range of hues. Consisting of various proportions of lead chromate ($PbCrO_4$), lead sulfate ($PbSO_4$), and lead monoxide (PbO), all turn brown on exposure to sulfide gases in the air because of the formation of lead sulfide (PbS). Molybdate oranges, inclining toward the red, are made from lead chromate precipitated together with lead molybdate ($PbMoO_4$) and lead sulfate ($PbSO_4$). These also darken on exposure. Zinc yellow has a greenish cast because of a high content of chromic oxide. It is lightfast, rust inhibitive, and sulfide resistant. Basic zinc chromate ($5ZnO \cdot CrO_3$) is widely used as an undercoat. Cadmium yellow and or-

anges are formed by diluting yellow cadmium sulfide (CdS) with white zinc sulfide (ZnS) and barium sulfate ($BaSO_4$), using small amounts of red cadmium selenide (CdSe) to impart the orange coloration. Yellow-tinted titanium dioxide is produced by calcining it in the presence of nickel. The resulting product is very lightfast. Organic yellows and oranges are expensive compounds.

Blue and green pigments are mostly synthetic today. The natural blues and greens such as lapis lazuli have disappeared from coatings technology. In their place are synthetic ultramarines, produced by fusing china clay in a mixture of sulfur, sodium sulfate, and an organic reducing agent such as rosin. The glassy product is pulverized, leached free of sodium sulfate, and size-separated by sedimentation.

The most widely used blue pigments are the iron blues that contain various amounts of iron, ammonium and sodium ions (electrically charged particles), and water of crystallization (water trapped in the molecular structure as the pigment crystallizes). Two versions of iron blue are produced. Chinese blue has a greenish-blue tint; Milori blue tends toward the red. Hiding power of iron blue is good. Phthalocyanine pigments have a rather complicated organic structure; are reproducible, lightfast, brilliant; and make excellent tinting pigments but are expensive to produce.

The anthraquinone molecule ($C_{14}H_8O_2$), also used in dyes, can be made to form a permanent strong blue pigment that can be obtained in green hues by substituting chlorine atoms for most of the hydrogen atoms. It is an expensive pigment that withstands baking.

Iron blues mixed with chrome yellows in various proportions produce a popular class of chrome greens. Except for the darkening inherent in these lead-containing compounds (and the growing objections in many places to lead in coatings on grounds of health hazard), chrome greens offer the best combination of features for indoor and outdoor applications. Chromic oxide (Cr_2O_3) is a heat resistant, lightfast pigment.

Miscellaneous pigments. Finely divided aluminum, copper, bronze, and zinc are used to create a metallic or bronzing effect in coatings, particularly on metal panels, and also to impart corrosion resistance. The metals are produced in the form of powders or flakes that align themselves in the finished coating roughly parallel with the substrate. The flakes have good hiding power and are highly reflective, whereas spherical particles produce a less brilliant, sometimes milky, effect, with somewhat less protective value.

Fluorescent pigments include materials that absorb short wavelength radiation and re-emit light of longer wavelength. These are exceptionally brilliant if the emitted radiation falls in the same wavelength range as the reflected radiation. The molecular structure of the dyes responsible for this reinforcement must be such that electronic energy levels for the reflected and emitted light lie near each other. Phosphorescent pigments, which glow in the dark, are made from radioactive materials. Pearlescent pigments create a pearly, or nacreous, effect by diffraction of visible light.

Fluorescent pigments

Extenders. Certain inorganic salts that have little hiding power by themselves can be used to extend or conserve the more expensive hiding pigments; these are called extenders, or fillers. Extenders are also used to control the flow properties of paint, improving brushability and mechanical strength of the dried film. They sometimes decrease gloss and increase the resistance of a film to vapour transmission.

The main types of extenders are sulfates, which include natural or precipitated calcium or barium salts; carbonates, which comprise calcium or a mixture of calcium and magnesium salts; oxides, which include silica, either natural, precipitated, or hydrolyzed; and silicates, among which are clay, talc, mica, and calcium silicate. The types of coating that employ large volumes of extenders are architectural and maintenance finishes, particularly interior flat wall paints where hiding is not as critical as uniform coverage.

VARNISHES AND OTHER PAINT BINDERS

A surface coating protects the substrate against external influences such as moisture, light, and corrosive elements in the atmosphere. Thus, the binder for the pigment and whatever extenders are present are characterized by fluidity before application to the surface and rigidity soon afterward. This metamorphosis requires that a surface coating be either a solution from which the solids can be smoothly deposited on the substrate, a solution from which the binder will evaporate, or a liquid capable of solidifying or polymerizing to a rigid mass.

Natural binders. Originally, pigments were suspended in water, with and without binders. As binders the Persians used gum arabic; the Chinese, weak glue; the Asiatic Indians, boiled rice; the East Indians, shellac; and the Japanese, the sap from the wood-oil tree (which they called the varnish tree), possibly the first drying oil. The American Indians used a mixture of salmon eggs and fish oil.

Oil, wax, egg albumen, and glue were used in medieval Europe. During the Renaissance each artist developed his own formula for incorporating resins in drying oils. With the Industrial Revolution, copals and amber became the principal resins, and five classes of varnish were in use, only one of which contained linseed oil, which was to become the backbone of the industry in its formative stage.

The first varnishes were solutions of natural resins, characterized by transparency, hardness, amorphous (noncrystalline) structure, and virtually no long-lasting integrity. They developed cracks on aging, even when used indoors; the utility of an outside coating relied largely on the natural tendency of pigments to adhere to the substrate by mechanical means.

Early varnishes

Versatility is a characteristic of natural resins. They vary not only in solubility but also in transparency, brittleness, and crystallinity. Many are formed as natural secretions of plants or insects. Those resins that are compatible with oils (rosin, copals, amber) generally require heating. By contrast, the spirit-soluble resins (balsams, turpentine, lacs) usually dissolve without application of heat, but the danger of fire was a hazard in preparing their solutions with the highly volatile solvents.

Modern vehicles. Synthetic resins are now the heart of the paint industry. Not subject to variations in availability or quality as were their natural predecessors, these substances can be adapted with great chemical precision to various end uses and can be formed in many different suspending media.

There are two main types of synthetic resin: condensation polymers and addition polymers. Condensation polymers, formed by condensation of like or unlike molecules into a new, more complex compound, include polyesters, phenolics, amino resins, polyurethanes, and epoxies. Addition polymers, formed by direct chemical combination of substances into a single large molecule consisting of repeated units of the simpler original substance, include polyvinyl acetate, polyvinyl chloride, and the acrylates.

Condensation polymer resins. Alkyds (any of several common condensation polymer resins) are the most widely used protective-coating vehicles. Almost 450,000,000 kilograms (1,000,000,000 pounds) of these substances are consumed annually throughout the world. Brittle and insoluble in common solvents, they are always manufactured along with an agent (plasticizer) that makes them soluble and plastic. By manipulating their molecular structure in various ways, manufacturers can make alkyds more or less viscous, more or less hard as a coating, soluble in water or other substances, adherent, and capable of mixing successfully with various pigments.

Phenol (C_6H_5OH) reacts readily with aldehydes (any of various compounds containing the group $-CHO$, composed of carbon, hydrogen, and oxygen) to form oil-soluble compounds called phenolic resins, which are rapid drying, hard, chemically resistant, and abrasion resistant. The outdoor life of phenolic resins is not long, however, and they tend to yellow on aging.

Phenolic resins

Condensation of urea (H_2NCONH_2) or melamine ($C_3H_6N_6$) with formaldehyde (HCHO) produces compounds called amino resins, which are non-yellowing and alkali resistant and have good adhesive properties. When

used in coatings, these amino resins contain an alcohol as a third component.

Both polyurethanes and epoxy resins are produced by means of complex organic reactions. Polyurethane coatings dry hard and inert on a substrate; epoxy resins have the qualities of toughness, adhesion, chemical resistance, rigidity, and thermal resistance.

Silicones are produced when various organic groups are attached to the silicon atom. They are hard and tough, with clarity and high gloss. They resist heat, weathering, and chemical attack and can be modified by blending or by chemical combination with other resins.

Addition polymers. Ethylene (CH_2=CH_2), one of the simplest unsaturated organic molecules, can be made into translucent, inert, waxy polymers (called polyethylenes) at high pressures. Though the low-molecular-weight resins lack toughness and the higher are insoluble and unworkable, the inertness of the polymers gives them special interest in paint applications. When polyethylenes are applied to substrates, the resulting film is resilient and resistant to acids and oxidizers.

Polyvinyl acetate resin

One of the synthetic addition polymers that has largely replaced natural resins is polyvinyl acetate. This substance is a clear transparent resin that softens below the boiling point of water. It is sensitive to water and decomposes in acids and bases. Its solubility in a wide variety of organic solvents is good. It also has excellent adhesion properties.

Vinyl resins are marketed either dissolved or suspended in an organic medium like the natural resins or polymerized and marketed as aqueous emulsions called latexes.

Polyvinyl chlorides, though related to the compounds discussed above, lack desirable qualities as coatings. Acrylic compounds, on the other hand, which are clear, nonyellowing and heat stable, are used in the production of acrylic emulsion paints. Slight variation in the chemical composition of acrylic esters can produce wide variations in the physical properties of the polymer. Manufacture of acrylic emulsion paints consists of dispersing the pigment in water and adding the latex with only moderate agitation.

Solvents. Except for the solventless coatings (linseed oil and unsaturated polyesters) and the aqueous dispersions (vinyls and acrylics), it is necessary to dissolve solid resinous binders. Sometimes solvents are referred to as thinners because their inevitable effect is to reduce the viscosity of the resin solution. Plasticizers belong to the same category except that their volatility is not as pronounced as that expected of a solvent.

FILMS

Film formation. In the sense used by paint scientists, film formation involves conversion of a fluid coating, such as paint, into one with considerable rigidity.

Colloid chemical basis. Before the development of coatings science, little was known about the process of film formation except that a successful conversion required the film to maintain both integrity and adherence. Much of the modern thrust of paint science has been devoted to upgrading appearance and durability of materials that were known from the start to be film formers; little attention was paid to promising materials of good stability and desirable mechanical properties that did not form films.

The process of film formation is still not thoroughly understood. It is a far more complicated phenomenon than simple hardening or freezing. It involves transient and nonreversible conversions that are subject to many interdependent variables such as temperature, humidity, and the nature of the substrate. Film drying involves orientation of molecules at interfaces. Matter with thicknesses in the colloid range of dimensions is involved.

Many of the test methods used by the paint industry are pragmatic: they are supposed to indicate when the coating is dry enough to be put in service. In effect they measure changes in the mechanical properties of the film and signify when they cross a particular threshold. All matter can be characterized by certain physical constants such as viscosity, modulus of rigidity, free volume, and density. Values of these constants, together with such basic chemical properties as bond strength, determine whether a film will form.

Types of film formation. There are two recognized ways to produce a rigid film on a substrate: continued growth of the polymer until it crosses the threshold between liquid and solid, and coalescence of finely dispersed polymer that has already terminated its growth as a molecule. The former method is chemical and the latter, physical. Combinations exist wherein the initial stage is physical and curing is effected chemically.

Physical and chemical film formation

The evolution of new concepts and new materials has broadened the number of variations of the film-forming process. Three strictly physical means of film-forming are: (1) from solution, in which evaporation of solvent leaves a deposit of polymer on the substrate; (2) from dispersions, in which evaporation of nonsolvent results in a deposit of particulate matter that coalesces into a continuous film; and (3) from melts, in which cooling of a supported fluid or plastic mass produces a rigid coating. Three strictly chemical means of film formation are: (1) by reaction with a component of the atmosphere in which the molecule combines with oxygen or water in the atmosphere; (2) by interaction of components already present in formulation, in which heat or radiation induce a reaction between species that coexist in the container without reacting; and (3) by reaction with a reagent or in the presence of a catalyst added just before application.

It is possible for solvent evaporation to accompany any of the three chemical means of film formation in a combination of physical and chemical drying. A combination of the two processes is provided by oil in the form of extended latexes, the initial rapid set of which is provided by evaporation and coalescence but the ultimate development of high rigidity of which is chemical.

Film formation by physical means. A solid resin or synthetic high polymer may be dissolved in a volatile solvent and converted by evaporation from a mobile fluid into a rigid film. Lacquer is an example of a coating that dries exclusively by this mechanism.

One of the standard polymerization techniques uses an emulsion directly as a film former, providing that the particles are deformable and hence are able to pack closely on drying. Only if the interstices and particle interfaces (points of contact) are obliterated will a clear, continuous film result. Acrylic ester emulsions may be formulated at extremely high molecular weights (up to 1,000,000), resulting in a polymer highly resistant to attacks by solvents.

Film formation by chemical means. Chemical reactions that occur during drying and hardening may lead to the formation of molecules of elongated or branched shape. When linseed oil dries and when epoxy resin hardens, the molecules link up and become elongated. In the case of some triglycerides and epoxies, elongation and cross-linking take place indefinitely, converting liquid films to a solid comprising one giant macromolecule. The greater the frequency of cross-links in the molecule, the more rigid the dried film.

Characteristics. Coatings are measured and characterized with instruments, ranging from the spatula of the craftsman and the penknife (or thumbnail) of the early development chemist to highly complex and expensive modern instruments.

Polymer properties. The properties of any substance depend on the size, shape, and composition of the molecules and on the magnitude and nature of the bonding forces within each molecule. Polymer structures are either open-chain or cross-linked; the former are generally thermoplastic and soluble, whereas the latter remain rigid at all temperatures below decomposition and merely swell in the presence of solvents.

Polymer properties and film formation

The properties of polymers that influence the quality of the ultimate film include the strength of bonding between atoms and between repeating groups of atoms in the chain of the polymer molecule; the crystallinity of the polymer—that is, the geometric arrangement of repeating atoms with the chain; and the manner in which polymers mix with various solvents. Measurement of these properties and determination of the exact means by which they influence film quality is complex and even controversial. Many difficult questions remain to be answered.

Paint properties. A paint should be applicable while

liquid; on application all surface irregularities should smooth out, and paint film should flow into cracks and crevices before it develops appreciable rigidity. If a fluid condition is maintained too long, the coating will sag on vertical surfaces; consequently, most paints are so formulated that the time of hardening is kept short. This is done by adding what are called thixotropic agents to the paint that cause the binder to lose its structural rigidity as it is brushed onto the surface and to regain it shortly thereafter.

ECONOMIC ASPECTS

The paint industry of the world has enjoyed a steady growth, notably in western Europe, Japan, and the United States. The number of paint producers in any country is typically large because of the competitive advantage experienced by producers in the immediate geographic area in which the paint is marketed and used. Demand is met by many small producers as well as by branches of large companies; the consumers' requirements for small specialty batches and his call upon technical service staffs are based on many subjective factors requiring local contacts.

Half of the paint market involves industrial firms that sell the coating as an integral part of the finished product, and only half involves over-the-counter sales to final users. Nevertheless, the small paint manufacturer exerts a profound effect on the overall industry; it is common for the leading paint company in any given metropolitan area to be a local firm. To compete with the research efforts of the large companies, small companies often band together to form cooperative research groups.

Over one-third of the cost of paint is the cost of resins and other chemical raw materials; another third is for manufacturing costs and the remainder for packaging and marketing.

Technical societies

Technical developments are disseminated largely by federations of local paint societies made up of individual rather than corporate memberships. The Oil and Colour Chemist's Association in London, the Federation of Scandinavian Paint and Varnish Technicians, and the Federation of Societies for Paint Technology (FSPT) in Philadelphia are continuously active. The countries of western Europe and North America hold a biennial conference under the auspices of a loose international federation known as Fatipec (Federation of Associations of Technicians in the Paint, Varnish, and Printing Ink Industries of Continental Europe).

The surface-coating industry promises to continue to grow in the future as a result of overall world economic growth, technology that will widen the application of surface coatings, and, finally, the important role foreseeable for paints, varnishes, and allied products in the struggle to improve the environment.

Paints will perform more varied functions than decoration and protection against degradation. Sacrificial coatings that protect surfaces from fire by swelling and bubbling are expected to increase appreciably; their use may be required for certain types of installations. Conductive coatings will serve a variety of uses ranging from panel heating to undercoats for electrodeposited finishes. Coatings that change hue on demand will provide a new dimension in decoration.

One of the more rapidly growing markets for coatings has been floor coverings. Having lost much of the market for ceilings, the paint industry has turned to seamless flooring, a liquid system possessing decorative, high-build, and abrasion-resistant qualities. Because of the need for high-impact resistance, seamless flooring must be applied at least 35 mils thick.

Pollution control

As a derivative industry dealing with conversions of raw materials into finished products, the paint industry is concerned more with pollution by its product and related hazards associated with coatings than with pollution during the process of paint manufacture. Research has already been brought to bear effectively on the solution of the production problem. Pollution by the product after application is more difficult to estimate or even to identify. Lead has been removed from coating formulations that are liable to come within the reach of children. Mercury salts were the best fungicides for preventing mildew growth on surfaces until that metal was ruled out of house paint.

Hiding without pigments is an attainable goal. The bloom on fruit, the opacity of flower petals, and blush films all result from the scattering of light at the interface of minute voids in the film. The low refractive index of such voids provides an optical mismatch with the film, as does high-index pigment, so that hiding is just as effective with voids as with solid pigments. Inferior mechanical properties of pigmentless paints, particularly as manifested by loss of hiding power when voids are crushed, must be overcome if these materials are to make deep inroads into paint markets.

Industry-wide basic research is supported by the Paint Research Institute of the FSPT. Through its program of awarding fellowships in chemistry, physics, and engineering, the institute maintains contact with the latest developments in colloid and macromolecular science.

(R.R.M./Ed.)

Papermaking

Paper is the basic material used for written communication and the dissemination of information. In addition, paper and paperboard provide materials for hundreds of other uses, such as wrapping, packaging, towelling, insulating, and photography.

Paper has been defined as a matted or felted sheet formed on a wire screen from water suspension. The word paper is derived from the name of the reedy plant papyrus, which grows abundantly along the Nile River in Egypt. In ancient times, the fibrous layers within the stem of this plant were removed, placed side by side, and crossed at right angles with another set of layers similarly arranged. The sheet so formed was dampened and pressed. Upon drying, the glue-like sap of the plant, acting as an adhesive, cemented the layers together. Complete defibring, an indispensable element in modern papermaking, did not occur in the preparation of papyrus sheets. Papyrus was the most widely used writing material in ancient times, and many papyrus records still survive.

Manufacture of papyrus

THE PAPERMAKING PROCESS

Historical development. Papermaking can be traced to about AD 105, when Ts'ai Lun, an official attached to the Imperial court of China, created a sheet of paper using mulberry and other bast fibres along with fish nets, old rags, and hemp waste. In its slow travel westward, the art of papermaking reached Samarkand, in Central Asia, in 751; and in 793 the first paper was made in Baghdad during the time of Harun-ar-Raschid, with the golden age of Islāmic culture that brought papermaking to the frontiers of Europe.

By the 14th century a number of paper mills existed in Europe, particularly in Spain, Italy, France, and Germany. The invention of printing in the 1450s brought a vastly increased demand for paper. Through the 18th century the papermaking process remained essentially unchanged, with linen and cotton rags furnishing the basic raw materials. Paper mills were more and more plagued by shortages; in the 18th century they even advertised and solicited publicly for rags. It was evident that a process for utilizing a more abundant material was needed.

Improvements in materials and processes. In 1800 a book was published that launched development of practical methods for manufacturing paper from wood pulp and other vegetable pulps. Several major pulping processes were gradually developed that relieved the paper industry of dependency upon cotton and linen rags and made modern large-scale production possible. These developments followed two distinct pathways. In one, fibres and fibre fragments were separated from the wood structure by mechanical means; and in the other, the wood was exposed to chemical solutions that dissolved and removed lignin and other wood components, leaving cellulose fibre behind. Made by mechanical methods, groundwood pulp contains all the components of wood and thus is not suitable for papers in which high whiteness and permanence are required. Chemical wood pulps such as soda and sul-

Mechanical and chemical pulping

fite pulp (described below) are used when high brightness, strength, and permanence are required. Groundwood pulp was first made in Germany in 1840, but the process did not come into extensive use until about 1870. Soda pulp was first manufactured from wood in 1852 in England, and in 1867 a patent was issued in the United States for the sulfite pulping process.

A sheet of paper composed only of cellulosic fibres ("waterleaf") is water absorbent. Hence, water-based inks and other aqueous liquids will penetrate and spread in it. Impregnation of the paper with various substances that retard such wetting and penetration is called sizing.

Before 1800, paper sheets were sized by impregnation with animal glue or vegetable gums, an expensive and tedious process. In 1800 Moritz Friedrich Illig in Germany discovered that paper could be sized in vats with rosin and alum. Though Illig published his discovery in 1807, the method did not come into wide use for about 25 years.

Discovery of the element chlorine in 1774 led to its use for bleaching paper stock. Lack of chemical knowledge at the time, however, resulted in production of inferior paper by the method, discrediting it for some years. Chlorine bleaching is a common papermaking technique today.

Introduction of machinery. Prior to the invention of the paper machine, paper was made one sheet at a time by dipping a frame or mold with a screened bottom into a vat of stock. Lifting the mold allowed the water to drain, leaving the sheet on the screen. The sheet was then pressed and dried (see Figure 16). The size of a single sheet was limited to the size of frame and mold that a man could lift from a vat of stock.

From K.W. Britt (ed.), *Handbook of Pulp and Paper Technology*, 2nd ed.; copyright © 1964 by Litton Educational Publishing, Inc., reproduced by permission of Van Nostrand Reinhold Co.

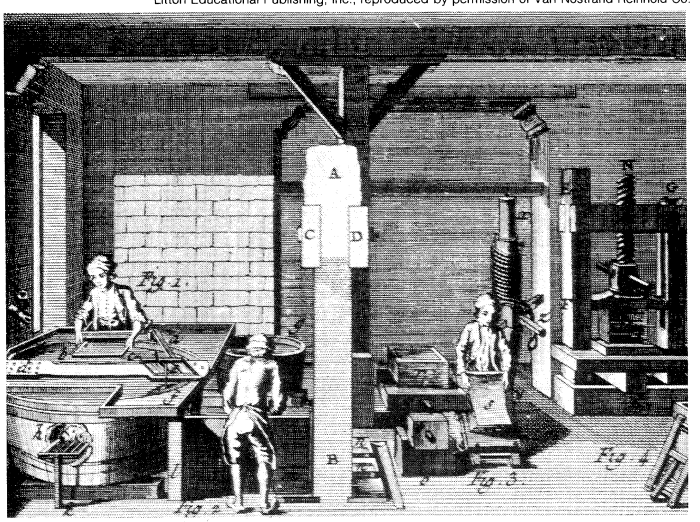


Figure 16: *Making paper by hand in an 18th-century paper mill.* The workman in "Fig. 1," the vatman, is dipping the mold into the vat of stock. The workman in "Fig. 2" is "couching" or pressing the wet sheet while it is still on the mold. The workman in "Fig. 3" is separating the sheets from the molds, making a stack to be pressed by the device in "Fig. 4." Woodcut from the *Encyclopédie*, edited by Denis Diderot and Jean Le Rond d'Alembert, 1751–72.

In 1798, Nicolas-Louis Robert in France constructed a moving screen belt that would receive a continuous flow of stock and deliver an unbroken sheet of wet paper to a pair of squeeze rolls. The French government recognized Robert's work by the granting of a patent.

The paper machine did not become a practical reality, however, until two engineers in England, both familiar with Robert's ideas, built an improved version for their employers, Henry and Sealy Fourdrinier, in 1807. The Fourdrinier brothers obtained a patent also. Two years later a cylinder paper machine (described below) was devised by John Dickinson, an English papermaker. From these crude beginnings, modern papermaking machines evolved. By 1875 paper coated by machinery was being made for use in the printing of halftones by the new photoengraving process, and in 1884 Carl F. Dahl invented sulfate (kraft) pulp in Danzig, Germany.

Although the paper machine symbolizes the mechanization of the paper industry, every step of production, from the felling of trees to the shipment of the finished product, has also seen a dramatic increase in mechanization, thus reducing hand labour. As papermaking operations require the repeated movement of large amounts of material, the design and mechanization of materials-handling equipment has been and continues to be an important aspect of industry development.

Though modern inventions and engineering have transformed an ancient craft into a highly technical industry, the basic operations in papermaking remain the same to this day. The steps in the process are as follows: (1) a suspension of cellulosic fibre is prepared by beating it in water so that the fibres are thoroughly separated and saturated with water; (2) the paper stock is filtered on a woven screen to form a matted sheet of fibre; (3) the wet sheet is pressed and compacted to squeeze out a large proportion of water; (4) the remaining water is removed by evaporation; and (5) depending upon use requirements, the dry paper sheet is further compressed, coated, or impregnated.

Basic operations in papermaking

The differences among various grades and types of paper are determined by: (1) the type of fibre or pulp; (2) the degree of beating or refining of the stock; (3) the addition of various materials to the stock; (4) formation conditions of the sheet, including basis weight, or substance per unit area; and (5) the physical or chemical treatment applied to the paper after its formation.

Fibre sources. The cell walls of all plants contain fibres of cellulose, an organic material known to chemists as a linear polysaccharide. It constitutes about one-third of the structural material of annual plants and about one-half that of perennial plants. Cellulose fibres have high strength and durability. They are readily wetted by water, exhibiting considerable swelling when saturated, and are hygroscopic—*i.e.*, they absorb appreciable amounts of water when exposed to the atmosphere. Even in the wet state, natural cellulose fibres show no loss in strength. It is the combination of these qualities with strength and flexibility that makes cellulose of unique value for paper manufacture.

Properties of cellulose fibres

Most plant materials also contain nonfibrous elements or cells, and these also are found in pulp and paper. The nonfibrous cells are less desirable for papermaking than fibres but, mixed with fibre, are of value in filling in the sheet. It is probably true that paper of a sort can be produced from any natural plant. The requirements of paper quality and economic considerations, however, limit the sources of supply.

Wood. Pulped forest tree trunks (boles) are by far the predominant source of papermaking fibre. The bole of a tree consists essentially of fibres with a minimum of nonfibrous elements, such as pith and parenchyma cells.

Forests of the world contain a great number of species, which may be divided into two groups: coniferous trees, usually called softwoods, and deciduous trees, or hardwoods. Softwood cellulose fibres measure from about two to four millimetres (.08 to .16 inch) in length, and hardwood fibres range from about 0.5 to 1.5 millimetres (0.02 to 0.06 inch). The greater length of softwood fibres contributes strength to paper; the shorter hardwood fibres fill in the sheet and give it opacity and a smooth surface.

Softwood and hardwood fibres

When the sulfite process (see below) was the chief method of pulping in the early days of the pulp industry, spruce and fir were the preferred species. Since that time, advances in technology, particularly the introduction of the kraft process (described below), have permitted the use of practically all species of wood, greatly expanding the potential supply.

Because of the enormous and rapidly growing consumption of wood for pulp, concern regarding the depletion of forest resources has been expressed, even though yearly growth often exceeds the annual harvest. In 1962, for example, though new growth exceeded the harvest by a considerable margin, much of it was inferior in quality and less accessible than the harvested trees. Moreover, wood is now being harvested at a more rapid pace. Approximately 40 percent of the harvest is going into pulp, and that figure is expected to increase. There is also a rising public

demand for withdrawal of forest land from timber production for recreational use and to prevent disturbance to the ecology of certain areas. On the other hand, application of new techniques in fertilization and genetics have brought about enormous increases in the productivity of forest lands in some areas. Universal adoption of these new methods could help to assure a future supply of wood.

Two significant trends in pulpwood utilization deserve mention. Until recently, lumbering and other wood-using industries were operated quite independently of the pulp industry. Since World War II, however, the waste from the wood-using industries, such as sawdust, has increasingly been used for pulp. In addition, more abundant and less desirable hardwoods have been used as a source of pulp. The woodyard of a pulp mill formerly stored pulpwood in the form of roundwood logs, but recently there has been a trend toward storing in the form of chips.

Rags. Cotton and linen fibres, derived from textile and garment mill cuttings; cotton linters (the short fibres recovered from the processing of cottonseed after the separation of the staple fibre); flax fibres; and clean sorted rags are still used for those grades of paper in which maximum strength, durability, and permanence are required, as well as fine formation, colour, texture, and feel. These properties are attributed to the greater fineness, length, and purity of rag fibre as compared with most wood pulp. Rag papers are used extensively for bank note and security certificates; life insurance policies and legal documents, for which permanence is of prime importance; technical papers, such as tracing paper, vellums, and reproduction papers; high-grade bond letterheads, which must be impressive in appearance and texture; lightweight specialties such as cigarette, carbon, and Bible papers; and high-grade stationery, in which beauty, softness, and fine texture are desired.

Properties of rag paper

Rags are received at the paper mill in bales weighing from 200 to 500 kilograms (400 to 1,200 pounds). After mechanical threshing, the rags are sorted by hand to remove such foreign materials as rubber, metal, and paper and to eliminate those rags containing synthetic fibres and coatings that are difficult to remove. Following sorting, the rags are cut up, then dusted to remove small particles of foreign materials, and passed over magnetic rolls to remove iron.

The cut and cleaned rags are cooked (to remove natural waxes, fillers, oils, and grease) in large cylindrical or spherical boilers of about five-ton capacity. About three parts of cooking liquor, a dilute alkaline solution of lime and soda ash or caustic soda combined with wetting agents or detergents, are used with each part of rags. Steam is admitted to the boiler under pressure, and the contents are cooked for three to ten hours.

Once cooked, the rags are washed, then mechanically beaten. The beating shortens the fibre, increases the swelling action of water to produce a softened and plastic fibre, and fibrillates or frays the fibre to increase its surface area. All of these actions contribute to better formation of the paper sheet, closer contact between fibres, and the formation of interfibre bonding that gives the paper strength and coherence.

Wastepaper and paperboard. By using greater quantities of wastepaper stock, the need for virgin fibre is reduced, and the problem of solid waste disposal is minimized. The expansion of this source is a highly complex problem, however, because of the difficulties in gathering wastepaper from scattered sources, sorting mixed papers, and recovering the fibre from many types of coated and treated papers.

Wastepaper may be classified into four main categories: high-grades, old corrugated boxes, printed news, and mixed paper. High-grade and corrugated stocks originate mainly in mercantile and industrial establishments. White paper wastes accumulate in envelope and printing plants, while tabulating cards are supplied by large offices. Much magazine stock comes from newsstand returns, but some comes from homes. Corrugated waste is supplied by manufacturing plants and retail stores. Printed news is derived from newsstand returns and home collections. Mixed paper comes from wastebaskets of office buildings and similar sources. In recent years there has been considerable interest in wastepaper recycling in the interest of ecology.

Converters of paper and paperboard have also increasingly turned to new materials combined with paper and paperboard to give their products special characteristics. Although these new materials have broadened the market for paper, their presence has posed new problems in reusing paper stock. The most common new ingredients are asphalt, synthetic adhesives, metal foils, plastic and cellulose-derivative films and coatings, and some printing inks.

Some objectionable materials can be sorted from wastepaper, and packers generally try to remove them completely. If the producer of wastepaper knows the materials he is using, he can usually segregate trouble-causing substances at the source. Much depends on good cooperation and communication among the papermaker, dealer, packer, and producers so that all may understand what is and what is not acceptable.

There are two distinct types of paper recovery systems: (1) recovery based upon de-inking and intended for printing-grade or other white papers, accounting for about 5 to 6 percent of the total; and (2) recovery without de-inking, intended for boxboards and coarse papers, accounting for the remainder.

Paper recovery systems

In the de-inking recovery process, the bales of wastepaper are opened, inspected, and fed into a pulper, a cylindrical tank with capacity ranging from one to several tons of stock and provided with agitator blades that circulate and agitate the stock. Hot water and various chemicals help the agitator separate and disperse the fibres.

The amount and type of chemicals used vary considerably from mill to mill. Caustic soda is by far the most generally used, but it is often supplemented with soda ash, silicate of soda, phosphates, and surfactants (wetting agents). The temperature range is from 65° to 90° C (150° to 190° F).

The pulpers are aided in the collection and separation of large pieces of trash by special devices. After the stock leaves the pulper, it is screened to remove finer trash particles and washed to remove the dispersed ink and chemicals. In some instances the stock is bleached with hypochlorite to improve its whiteness.

In pulping paper stock where de-inking is not necessary, the equipment is similar to that already described. Hot water is also used in the pulper, but the chemicals for dissolving and dispersing the ink are not needed. The stock is screened and washed to remove trash and dirt.

The use of paper stock in the paper mill presents difficulties because of the presence of foreign materials. Miscellaneous trash has always required operators to be watchful, and its presence depends on the source of the waste and the care with which the paper is prepared for market.

Natural fibres other than wood. Since cellulose fibre is a major constituent of the stems of plants, a vast number of plants represent potential sources of paper; many of these have been pulped experimentally. A rather substantial number of plant sources have been used commercially, at least on a small scale and at various times and places. Indeed, the use of cereal straws for paper predates the use of wood pulp and is widely practiced today throughout the world, although on a relatively small scale of production. Because many parts of the world are deficient in forests, the development of the paper industry in these areas appears to depend to a considerable degree upon the use of annual plants and agricultural fibres.

Agricultural fibres

Nonwoody plant stems differ from wood in containing less total cellulose, less lignin, and more of other materials. This means that pulps of high cellulose content (high purity) are produced in relatively low yield, whereas pulps of high yield contain high proportions of other materials. Papers made from these pulps without admixture of other fibre tend to be dense and stiff, with low tear resistance and low opacity.

The morphology (form and structure) of the cells of annual plants also differs considerably from wood. Whereas the nonfibrous (parenchyma) cells of coniferous wood comprise a minor proportion of the wood substance, in annual plants this cell type is a major constituent. As

hardwoods also often contain considerable amounts of nonfibrous cells, there is a closer resemblance between hardwood pulps and pulps from annual plants.

Pulping reagents

The preferred pulping reagents for nonwood plants are the alkalis: caustic soda, lime and soda ash, and kraft liquor (caustic soda and sodium sulfide). A characteristic of the pulping of annual plants, compared with wood, is the milder treatment necessary to produce pulp. Straw, for example, may be pulped with milk of lime in a spherical digester at a steam pressure of about two kilograms per square centimetre (25 pounds per square inch) and a cooking time of eight to 10 hours. The amount of lime used is about 10 percent of the amount of dry fibre.

In the United States straw pulp was formerly used extensively for corrugating medium (*i.e.,* sheet fluted to form the inner ply of corrugated board). Since then, the use of straw pulp for corrugating medium has been replaced by semichemical hardwood pulp. Straw pulp is still made in several European and Asiatic countries on a small scale.

Bagasse

The residue from the crushing of sugarcane, called bagasse, contains about 65 percent fibre, 25 percent pith cells, and 10 percent water solubles. An essential element in the conversion of bagasse to a satisfactory paper is the mechanical removal of a substantial proportion of the pith prior to the pulping operation. Pulping may be carried out either with soda or with kraft cooking liquor and by batch or continuous systems. Bagasse fibre averages 1.5 to two millimetres (0.06 to 0.08 inch) in length and is relatively fine.

The use of bagasse is substantial in several Latin American countries and in the Middle East. The utilization of bagasse for paper in all the sugar-producing countries that are deficient in forest resources is a practical step.

Esparto

A desert plant of the Mediterranean area, especially in southern Spain and northern Africa, esparto grass has a higher cellulose content than most nonwood plants, with greater uniformity of fibre size and shape. The use of esparto for papermaking was developed in Great Britain in 1856. Consumption rose steadily until the mid-1950s, but since has steadily declined.

Esparto held its own against the competition with wood pulp for some time because of its favourable papermaking properties. The stock forms well on a paper machine because of free drainage and uniform fibre length, compared with rag or wood pulp. Esparto printing papers possess good resilience in contact with the printing plate, have good opacity and smoothness, and are relatively lint-free. Another important characteristic of papers made from esparto is dimensional stability with changes in moisture content.

Bamboo

Botanically, bamboo is classified as a grass, even though it attains a considerable size and the stems or culms resemble wood in hardness and density. It was demonstrated many years ago that satisfactory pulp could be made from bamboo.

Because of the abundance of bamboo in Southeast Asia, where increased production of paper is greatly needed, much interest has been displayed in bamboo pulp development. The growing cycle of bamboo is favourable, for the culms can be harvested without destroying the root system. Under ideal conditions of soil fertility and moisture, an established stand of bamboo probably would produce more fibre per hectare (or acre) per year than any other plant. Wild bamboo, however, is difficult to harvest and transport economically; so far, the interest in it has not been translated into any large-scale production. Pulp mills make use of bamboo in India, Thailand, and the Philippines. Considerable quantities of bamboo pulp are said to be made in China, but details are lacking.

Flax, hemp, jute, and kenaf are characterized by a high proportion of long, flexible bast fibres that are readily separated and purified from the other materials in the plant. Consequently, such fibres have long been used for textiles and rope-making. Most of this fibre reaching the paper industry in the past has been secondary or waste fibre. It has been highly prized because of the strength and durability it imparts to such products as tags, abrasive paper (sandpaper), cover stock, and other heavy-duty paper. It is also used for duplicating and manifold paper, in which extremely light weight must be combined with exceptional strength. Flax is grown expressly for high-grade cigarette paper. Experimental quantities of kenaf have been grown and made into various grades of paper.

Synthetic fibres. The development and use of a great variety of man-made fibres have created a revolution in the textile industry in recent decades. It has been predicted that similar widespread use of synthetic fibres may eventually occur in the paper industry. Active interest has been evident in recent years, both on the part of fibre producers and of paper manufacturers. Many specialty paper products are currently being made from synthetic fibres.

The advantages of synthetic or man-made fibres in papermaking can be summarized as follows:

Advantages of synthetic fibres

Whereas natural cellulose fibres vary considerably in size and shape, synthetic fibres can be made uniform and of selected length and diameter. Long fibres, for example, are necessary in producing strong, durable papers. There are limitations, however, to the length of synthetic fibres that may be formed from suspension in water because of their tendency to tangle and to rope together. Even so, papers have been made experimentally with fibres several times longer than those typical of wood pulp; these papers have improved strength and softness properties.

Natural cellulose fibres have limited resistance to chemical attack and exposure to heat. Because synthetic fibre papers can be made resistant to strong acids, they are useful for chemical filtration. Paper can even be made from glass fibre, and such paper has great resistance to both heat and chemicals.

The natural cellulose fibres of ordinary paper are hygroscopic; *i.e.,* they absorb water from the air and reach an equilibrium depending upon the relative humidity. The moisture content of paper, therefore, changes with atmospheric conditions. These changes cause swelling and shrinkage of fibres, accounting for the puckering and curling of papers. Synthetic fibres not subject to these changes can be used to produce dimensionally stable papers.

Cost of wood pulp and synthetic fibres

The cheapest man-made fibre, rayon, costs from three to six times as much as an equivalent amount of wood pulp, whereas most of the true synthetics, such as the polyamides (nylon), polyesters (Dacron, Dynel), acrylics (Orlon, Creslan, Acrilan), and glass, cost from 10 to 20 times as much. This difference in cost does not preclude the use of existing synthetics, but it limits their use to special items in which the extra qualities will justify the additional cost. The cost factor is increased by the absence in most synthetic fibres of the bonding property of natural cellulose fibres. When beaten in water, natural fibres swell and cement together as they dry. Paper made from synthetics must be bonded by the addition of an adhesive, requiring an additional manufacturing step.

There is a distinct similarity between synthetic fibre "papers" and the class of sheet materials known as nonwovens. As a step in the manufacture of yarn, staple fibres are carded (*i.e.,* separated and combed) to form a uniform, lightweight, and fragile web. Subsequently, this web is gathered together to form a strand or sliver, which is drawn and spun into yarn. If several of these flat webs, however, are laminated together and bonded with adhesive, a nonwoven fabric that has properties resembling both paper and cloth results. In this area it is difficult to draw a clear distinction between what is paper and what is cloth. Processes are now available to form sheet material both by the dry forming method and by the water forming or paper system. When textile-type fibres are formed into webs by either of these processes, the resulting products have properties that enable them to compete in some fields traditionally served by textiles.

Processes for preparing pulp. Mechanical or groundwood pulp is made by subjecting wood to an abrading action, either by pressing the wood against a revolving grinding stone or by passing chips through a mill. The wood fibres are separated and, to a considerable degree, fragmented.

Chemical wood pulp is made by cooking wood chips with chemical solutions in digesters operated at elevated temperature and pressure. The chemicals used are: (1) sulfite salts with an excess of sulfur dioxide, and (2) caustic

soda and sodium sulfide (the kraft process). The lignin of the wood is made soluble, and the fibres separate as whole fibres. Further purification can be accomplished by bleaching. Chemical wood pulp that is purified both by bleaching and by alkaline extraction is called alpha or dissolving pulp. It is used for specialty papers, for rayon and cellulose film production, and for cellulose derivatives, such as nitrate and acetate.

Semichemical pulp is made by treating wood chips with sulfite or alkali in amounts and under conditions that soften the lignin but dissolve only part of it. The softened chips are then defibred.

Mechanical or groundwood pulp. Pulpwood may arrive at the mill as bolts 1.2 metres (four feet) in length or as full-length logs. The logs are sawn to shorter length, and the bolts are tumbled in large revolving drums to remove the bark. The debarked wood is next sent to grinders, where its moisture content is important for ease of grinding and quality of pulp. Moisture content should be at least 30 percent and preferably 45 to 50 percent. Wood of low moisture content is presoaked in a pond or sprayed with water.

Importance of moisture content

Early grinders employed round slabs of natural sandstone 69 centimetres (27 inches) wide and 137 centimetres (54 inches) in diameter, often directly connected to water wheels, to produce five or six tons of pulp per day. The wood was hand-loaded into the grinders.

Today's much larger pulp grinders are usually powered by electric motors and automatically loaded. In a recently built mill, each grinder is gear connected to a 10,000 horsepower motor; the pulpstone, at 360 revolutions per minute, can handle wood 1.5–1.6 metres (60 to 64 inches) long. Hydraulic cylinders produce a pressure of 14 kilograms per square centimetre (200 pounds per square inch) against the stone face. Pulp production from each stone is 130 to 150 tons every 24 hours.

The first artificial grinding stone was produced in 1924; since that time, artificial stones have replaced natural sandstone. Silicon carbide and aluminum oxide are the abrasives used in the manufacture of pulpstones. The abrasive material is broken down into a mixture of sizes that are screened to give fractions of uniform grain size. The abrasive grains are mixed with binder and fired at high temperature (2,300° C or 4,200° F) in the form of segments that are assembled to form the abrasive surface of the pulpstone.

The pulp stock flows from the grinder pit to a series of rifflers and screens, which separate the heavy foreign material and pieces of unfibred wood (shives), knots, bark, and the like.

Most groundwood pulp flows directly to an adjacent paper mill for use as stock. When shipped, it is formed into a sheet on a cylindrical vacuum filter. The sheets are pressed in a hydraulic press to a moisture content of about 50 percent, and the pressed sheets are formed into bales.

An important test to control the quality of groundwood pulp is freeness: the readiness with which water drains from and through a wet pad of pulp. Groundwood pulps are much less "free" than chemical wood pulps.

Freeness characteristic

In groundwood pulp, the fibres are fragmented, and there is considerable debris (fines). Also, groundwood contains all the chemical constituents of wood, including lignin, hemicellulose, resin, and various colouring materials. This means that papers containing groundwood are subject to discoloration (yellowing) upon exposure to light and heat and after aging. The yellowing of newspaper and much book paper is an example of this. Because groundwood fibres are relatively short and have only a moderate ability to bond to each other, papers containing them do not have high strength. On the other hand, papers containing groundwood have good opacity; they are bulky and have good printing qualities.

Groundwood pulp does not have a high whiteness, being limited in this quality by the colour of the wood from which it is made. Though often bleached with peroxide or hydrosulfite to improve whiteness, it does not equal pure cellulose.

Chemical wood pulp. The effect of sulfurous acid (H_2SO_3) in softening and defibring wood was observed by B.C. Tilghman, a U.S. chemist, as early as 1857. Several years later he renewed his experiments and, in 1867, was granted a patent for making paper pulp from vegetable material. He used high temperature and pressure and observed that the presence of a base such as calcium was important in preventing burned or discoloured batches of pulp. His work, however, did not result in commercial use of the process.

During the 1870s the sulfite process for pulping wood was the subject of experimental work in Sweden, England, Germany, and Austria. Within a few years the process was in commercial operation both in Europe and in North America. For many decades, the sulfite process was the leading process for the pulping of wood. Since 1940, however, the kraft process has taken a predominant position, and sulfite mills are no longer being constructed.

Sulfite process

Sulfite cooking liquor, as it is pumped to the digester at the start of a "cook," consists of free sulfur dioxide dissolved in water at a concentration of 4 to 8 percent, together with from 2 to 3 percent in the form of bisulfite. Sulfite digestion is normally carried out as a batch process in a pressure vessel, a steel shell with an acid-resistant lining of ceramic tile set in acid-proof cement or stainless steel. A common digester measures five metres (16 feet) in diameter and 15 metres (50 feet) in height, with a domed top and a conical bottom. It has a capacity of 12 to 15 tons of pulp per batch. Digesters of up to 35 tons capacity have been constructed. Pulp mills normally have a series of digesters arranged in a digester building.

After the blow valve is closed at the bottom, the wood chips are allowed to flow into the top opening and are distributed to fill the digester completely. Hot acid from the accumulator is pumped into the digester unit, completely filling it and replacing the air. Steam provides the heat.

At the end of the cook, the contents of the digester are blown to a blowpit by rapid opening of the bottom valve. The violence of the blow defibres the cooked chips.

From 1 to 6 percent of the digester charge is undesirable material such as knots, uncooked chips, dirt, bark, fibre bundles, and shives. The screen room separates the unwanted particles from accepted fibre, normally on the basis of particle size; there is an increasing use of the centrifugal principle, which separates particles on the basis of density.

The sulfite cooking liquor does not "cook out" or disintegrate bark and other foreign material to the same degree as kraft liquor (described below), and hence more care must be used in selecting and cleaning wood chips for sulfite.

In the conventional sulfite cook using softwood, the typical yield is 44 to 46 percent, based on wood and with a lignin content of 2 to 5 percent. At that point, a relatively light-coloured pulp with good strength properties is obtained, suitable for use in the unbleached state, especially in mixture with groundwood for a variety of printing papers. For pulps in which high brightness (whiteness) is desired, the residual lignin is removed by bleaching.

In 1851 paper pulp was experimentally produced from wood by cooking it with caustic soda at elevated temperature and pressure. Though this soda process attained commercial importance, soda pulp was of relatively low strength; and use of the process was limited to manufacturing filler pulps from hardwood, which were then mixed with a stronger fibre for printing papers. Because this process consumed relatively large quantities of soda, papermakers devised methods for recovering soda from the spent cooking liquor; recovery has remained an integral part of alkaline pulping ever since.

Soda process

In 1884 a German chemist, Carl F. Dahl, employed sodium sulfate in place of soda ash in a soda pulping recovery system. This substitution produced a cooking liquor that contained sodium sulfide along with caustic soda. Pulp so produced was stronger than soda pulp and was called "kraft" pulp, so named from the German and Swedish word for "strong." The process has also been termed the sulfate process because of the use of sodium sulfate (salt cake) in the chemical makeup. Sulfate, however, is not an active ingredient of the cooking liquor.

Kraft process

Many soda mills were converted to kraft because of the greater strength of the pulp. Kraft pulp, however, was

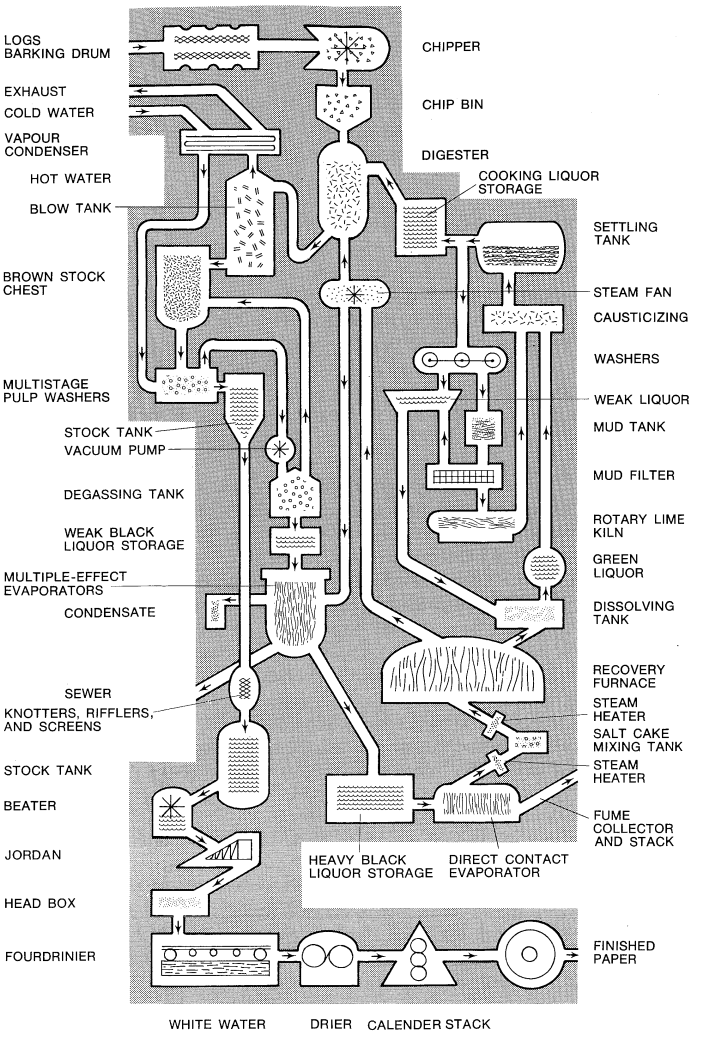


Figure 17: Stages in the kraft pulping process.
Drawing by D. Meighan

dark in colour and difficult to bleach; for many years the growth of the process was slow because of its limitation to papers in which colour and brightness were unimportant. In the 1930s, bleached kraft became commercially important with the discovery of new bleaching techniques. The availability of pulp of high whiteness and the expanding demand for unbleached kraft in packaging resulted in rapid growth of the process, making kraft the predominant wood-pulping method.

Paper produced by the kraft process is particularly strong and durable. Acceptable pulp can be produced in the kraft process from many species of wood not suitable for sulfite. The various pines, for example, especially southern yellow pine, contain large amounts of wood resin or pitch. Chemically altered and dissolved in the kraft process, this material is removed from the pulp and becomes a valuable by-product. The wood pitch is not removed to the same degree in the sulfite process, and hence high-resin woods, such as pine, are not suitable.

Continuous digesters

In the cooking operation, wood chips are prepared and fed to the digesting equipment by methods previously described. The cooking vessels are still widely used as batch digesters. In the past 25 years, however, continuous digesters have been developed and are being widely adopted by the kraft industry. These huge cylindrical towers, more than 60 metres (200 feet) in height, have a number of zones or compartments. Wood chips and cooking liquor are fed into the top and injected into successive zones of high pressure and temperature, where impregnation and cooking takes place as the chips progress downward. Additional zones wash the spent liquor from the chips. Continuous digesters are capable of producing 600 tons of pulp per day.

In batch cooking, after the digester is charged with chips, a mixture of "black liquor," the spent liquor from a previous cook, and "white liquor," a solution of sodium hydroxide and sodium sulfide from the chemical recovery plant, is pumped in. The digester is heated either by direct injection of steam or by the circulation of the cooking liquor through a heat exchanger.

Figure 17 shows a flow diagram of a kraft pulping operation. After completion of the cook, the spent cooking liquor is washed from the pulp; the latter is then screened and sent to the bleach plant or directly to the paper mill if it is to be used unbleached. Some of the spent liquor (black liquor) is used for an admixture with white liquor to charge new cooks; the remainder is sent to the recovery plant to reconstitute cooking chemicals.

All the sodium used for digestion is contained in the spent liquor, mostly in the form of sodium salts and sodium organic derivatives. The amount of sodium present is such that its re-use is economically necessary.

Semichemical pulp. For semichemical pulping, wood preparation and chipping are essentially the same as that for other wood-pulping processes. The chips are steeped and impregnated with inorganic chemical solutions similar to those used for full chemical pulping, but in smaller amounts and with less severe conditions. Probably the most common is the solution of sodium sulfite in the neutral range, between acidity and alkalinity. Other agents used in some cases are acid sulfite, caustic soda, and kraft cooking liquor.

After the impregnation operation, the chips are fed into one or more disk refiners (described below) in series. The attrition action of refiners reduces the softened chips to pulp. The yield of semichemical pulp based on wood is 66 to 90 percent. The higher fibre yield pulps are usually termed chemimechanical pulps.

The semichemical pulps have chemical and strength properties intermediate between softwood, groundwood, and full chemical pulps. These are used in a fairly wide range of papers and boards. The major tonnage of semichemical pulps goes into the light board, termed corrugating medium, which is fluted to serve as the interior layer of corrugated boxboard in heavy-duty containers. Stiffness and adequate strength are the important properties. Semichemical pulp is used extensively in a variety of low-cost printing papers.

Bleaching and washing. The use of calcium and sodium hypochlorites to bleach paper stock dates from the beginning of the 19th century. In the early days of sulfite pulp manufacture, a single-stage treatment of pulp at low consistency, using calcium hypochlorite (chlorinated lime), satisfied most requirements.

This simple bleaching treatment, however, is not practical for kraft that is difficult to bleach, nor can it retain maximum pulp strength. Accordingly, multistage bleaching systems have evolved in which various sequences of chemical treatment are employed, depending upon the type of unbleached pulp and special requirements.

During the normal first stage in a modern bleach plant, the unbleached pulp is chlorinated. Three to four percent of gaseous chlorine is rapidly mixed with pulp at a temperature of 21° to 27° C (70° to 80° F); the mixture is quite acid due to the acidity of the chlorine. Chlorine is absorbed largely by reaction with the non-carbohydrate components of pulp, with no brightening effect and with only slight dissolution of lignin.

Alkaline extraction

In the following stage an alkaline extraction with dilute caustic soda dissolves chlorinated compounds, which are then washed out.

In its simplest sequence the final stage consists of a treatment with a very alkaline hypochlorite to neutralize the solution, followed by a final wash.

In recent years the compound chlorine dioxide (ClO_2) has become available for on-site preparation; it is too unstable to be shipped for wood pulp bleaching. By the use of small amounts of ClO_2 in later bleaching stages, it is possible to achieve high degrees of purification and brightness without the degradation of cellulose.

The brightness of paper and other materials is determined by special reflection meters containing photoelectric cells

that measure the amount of light of selected wavelength reflected by the surface. Freshly prepared pure magnesium oxide is considered to be 100 on the brightness scale. On this scale unbleached sulfite and groundwood cover the range from about 50 to about 62; peroxide bleached groundwood, 66 to 72; single-stage hypochlorite sulfite, 80 to 85; multistage bleached pulp, 85 to 88; and multistage with chlorine dioxide, 90 to 94.

Manufacture of paper and paperboard. *Preparation of stock.* Mechanical squeezing and pounding of cellulose fibre permits water to penetrate its structure, causing swelling of the fibre and making it flexible. Mechanical action, furthermore, separates and frays the fibrils, submicroscopic units in the fibre structure. Beating reduces the rate of drainage from and through a mat of fibres, producing dense papar of high tensile strength, low porosity, stiffness, and rattle.

Hollander beater

An important milestone in papermaking development, the Hollander beater consists of an oval tank containing a heavy roll that revolves against a bedplate. The roll is capable of being set very accurately with respect to the bedplate, for the progressive adjustment of the roll position is the key to good beating. A beater may hold from 135 to 1,350 kilograms (300 to 3,000 pounds) of stock, a common size being about 7 metres (24 feet) long, 4 metres (12 feet) wide, and about 1 metre (3.3 feet) deep. A centre partition provides a continuous channel.

Pulp is put into the beater, and water is added to facilitate circulation of the mass between the roll and the bedplate. As the beating proceeds, the revolving roll is gradually lowered until it is riding full weight on the fibres between it and the bedplate. This action splits and mashes the fibres, creating hairlike fibrils and causing them to absorb water and become slimy. The beaten fibres will then drain more slowly on the paper machine wire and bond together more readily as more water is removed and the wet web pressed. Much of the beating action results from the rubbing of fibre on fibre. Long fibres will be cut to some extent.

The beater is also well-adapted for the addition and mixing of other materials, such as sizing, fillers, and dyes. By mounting a perforated cylinder that can rotate partially immersed in the beater stock, water can be continuously removed from the beater, and the stock can be washed.

Though many design modifications have been made in the Hollander beater over the years, the machine is still widely used in smaller mills making specialty paper products. For large production modern mills have replaced the beater by various types of continuous refiners.

Refiners

In mills that receive baled pulp and use refiners, the pulp is defibred in pulpers. While there are a number of variations in basic design, a pulper consists essentially of a large, open vessel, with one or more bladed, rotating elements that circulate a pulp–water mixture and defibre or separate fibres. The blades transform the pulp or wastepaper into a smooth mixture. Unlike beaters and refiners, pulpers do not reduce freeness and cause fibrillation in the fibres. A typical pulper has a capacity of 900 kilograms (2,000 pounds) of fibre in 6 percent solution and requires 150 horsepower to drive it.

The original continuous refiner is the Jordan, named after its 19th-century inventor. Like the beater, the Jordan has blades or bars, mounted on a rotating element, that work in conjunction with stationary blades to treat the fibres. The axially oriented blades are mounted on a conically shaped rotor that is surrounded by a stationary bladed element (stator).

Like other refiners, the disk refiner consists of a rotating bladed element that moves in conjunction with a stationary bladed element. The disk refiner's plane of action, however, is perpendicular to the axis of rotation, simplifying manufacture of the treating elements and replacement. Since the disk refiner provides a large number of working edges to act upon the fibre, the load per fibre is reduced and fibre brushing, rather than fibre cutting, may be emphasized.

Sizing

Sizing has been described above as the treatment given paper to prevent aqueous solutions, such as ink, from soaking into it. A typical sizing solution consists of a rosin soap dispersion mixed with the stock in an amount of 1 to 5 percent of fibre. Since there is no affinity between rosin soap and fibre, it is necessary to use a coupling agent, normally alum (aluminum sulfate). The acidity of alum precipitates the rosin dispersion, and the positively charged aluminum ions and aluminum hydroxide flocs (masses of finely suspended particles) attach the size firmly to the negatively charged fibre surface.

Filling

Paper intended for writing or printing usually contains white pigments or fillers to increase brightness, opacity, and surface smoothness, and to improve ink receptivity. Clay (aluminum silicate), often referred to as kaolin or china clay, is commonly used, but only in a few places in the world (Cornwall, in England, and Georgia, in the United States) are the deposits readily accessible and sufficiently pure to be used for pigment. Another pigment is titanium dioxide (TiO_2), prepared from the minerals rutile and anatase. Titanium dioxide is the most expensive of the common pigments and is often used in admixture with others.

Calcium carbonate ($CaCO_3$), also used as a filler, is prepared by precipitation by the reaction of milk of lime with either carbon dioxide (CO_2) or soda ash (sodium carbonate, Na_2CO_3). Calcium carbonate as a paper filler is used mainly to impart improved brightness, opacity, and ink receptivity to printing and magazine stocks. Specialty uses include the filling of cigarette paper, to which it contributes good burning properties. Because of its reactivity with acid, calcium carbonate cannot be used in systems containing alum.

Other fillers are zinc oxide, zinc sulfide, hydrated silica, calcium sulfate, hydrated alumina, talc, barium sulfate, and asbestos. Much of the filler consumed is used in paper coatings (see below).

Since most fillers have no affinity for fibres, it is necessary to add an agent such as alum to help hold the filler in the formed sheet. The amount of filler used may vary from 1 to 10 percent of the fibre.

The most common way to impart colour to paper is to add soluble dyes or coloured pigment to the paper stock. Many so-called direct dyes with a natural affinity for cellulose fibre are highly absorbed, even from dilute water solution. The so-called basic dyes have a high affinity for groundwood and unbleached pulps.

Bonding agents

Various agents are added to paper stock to enhance or to modify the bonding and coherence between fibres. To increase the dry strength of paper, the materials most commonly used are starch, polyacrylamide resins, and natural gums such as locust bean gum and guar gum. The most common type of starch currently used is the modified type known as cationic starch. When dispersed in water, this starch assumes a positive surface charge. Because fibre normally assumes a negative surface charge, there is an affinity between the cationic starch and the fibre.

The natural cellulose interfibre bonding that develops as a sheet of paper dries is considered to be due to interatomic forces of attraction known to physical chemists as hydrogen bonding or van der Waals forces. Because these attractive forces are neutralized or dissolved in water, wet paper has practically no strength. Though this property is convenient for the recovery of wastepaper, some papers require wet strength for their intended use. Wet strength is gained by adding certain organic resins to the paper stock that, because of their chemical nature, are absorbed by the fibre. After formation and drying of the sheet, the resins change to an insoluble form, creating water-resistant bonds between fibres.

Formation of paper sheet by machines. In a paper machine, interrelated mechanisms operating in unison receive paper stock from the beater, form it into a sheet of the desired weight by filtration, press and consolidate the sheet with removal of excess water, dry the remaining water by evaporation, and wind the travelling sheet into reels of paper. Paper machines may vary in width from about 1.5 to 8 metres (5 to 26 feet), in operating speed from a few hundred metres to 900 metres (about 3,000 feet) per minute, and in production of paper from a few tons per day to over 300 tons per day. The paper weight (basis weight) may vary from light tissue, about 10

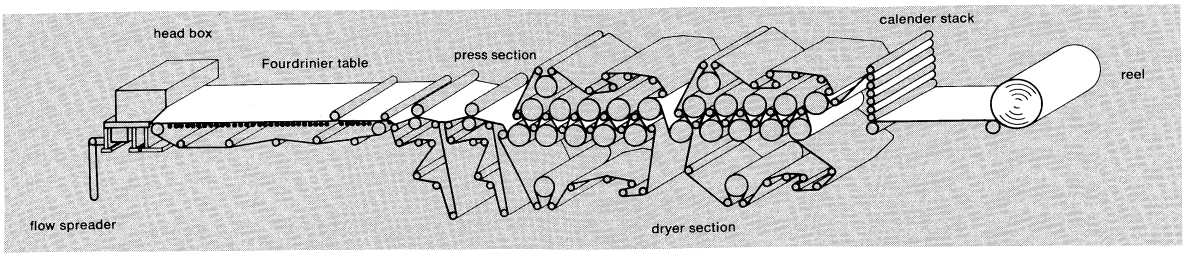


Figure 18: Elements of a Fourdrinier-type paper machine.
From K.W. Britt (ed.), *Handbook of Pulp and Paper Technology*, 2nd ed.; copyright © 1964 by Litton Educational Publishing, Inc., reproduced by permission of Van Nostrand Reinhold Co.

grams per square metre (0.03 ounce per square foot), to boards of over 500 grams per square metre (1.6 ounces per square foot).

'linder d Four- inier achines

Traditionally, paper machines have been divided into two main types: cylinder machines and Fourdrinier machines. The former consists of one or more screen-covered cylinders, each rotating in a vat of dilute paper stock. Filtration occurs by flow action from the vat into the cylinder, with the filtrate being continuously removed. In the Fourdrinier machine, a horizontal wire-screen belt filters the stock. In recent years a number of paper machines have been designed that depart greatly from traditional design. These machines are collectively referred to as "formers." Some of these formers retain the travelling screen belt but form the sheet largely on a suction roll. Others eliminate the screen belt and use a suction cylinder roll only. Still others use two screen belts with the stock sandwiched between, with drainage on both sides.

Figure 18 shows a typical modern Fourdrinier machine. The various functional parts are the headbox; stock distribution system; Fourdrinier table, where sheet formation and drainage of water occur; press section, which receives the wet sheet from the wire, presses it between woolen felts, and delivers the partially dried sheet to the dryer section; dryer section, which receives the sheet from the presses and carries it through a series of rotating, steam-heated cylinders to remove the remaining moisture; size press, which permits dampening the sheet surface with a solution of starch, glue, or other material to improve the paper surface; calendar stock, for compressing and smoothing the sheet; and the reel.

eadbox

The function of the headbox is to distribute a continuous flow of wet stock at constant velocities, both across the width of the machine and lengthwise of the sheet, as stock is deposited on the screen. Equal quantities of properly dispersed stock should be supplied to all areas of the sheet-forming surface. The early headbox, more commonly called a flowbox or breastbox, consisted of a rectangular wooden vat that extended across the full width of the machine behind the Fourdrinier breast roll. The box was provided with baffles to mix and distribute the stock. A flat metal plate extending across the machine (knife slice) improved dispersion of the fibre suspension, providing distribution of flow across the machine, and also metered the flow to produce a sheet of uniform weight. To accommodate increased speed in modern headboxes, the knife slice is designed to develop a jet of liquid stock on the moving wire. Modern headboxes are enclosed, with pressure maintained by pumping.

ɔrming ction

The Fourdrinier table section of a paper machine is a large framework that supports the table rolls, breast roll, couch roll, suction boxes, wire rolls, and other Fourdrinier parts. The wire mesh upon which the sheet of fibre is formed is a continuous rotating belt that forms a loop around the Fourdrinier frame. The wire, not a permanent part of the machine, is delicate and requires periodic replacement. It is a finely woven metal or synthetic fibre cloth that allows drainage of the water but retains most of the fibres. The strands of the Fourdrinier wire are usually made of specially annealed bronze or brass, finely drawn and woven into a web commonly in the range of 55 to 85 mesh (strands per inch). Even finer wires are used for such grades as cigarette paper, coarser wires for heavy paperboard and pulp sheets. Various types of weave are used to obtain maximum wire life.

The table rolls, in addition to supporting the wire, function as water-removal devices. The rapidly rotating roll in contact with the underside of the wire produces a suction or pumping action that increases the drainage of water through the wire.

The dandy roll is a light, open-structured unit covered with wire cloth and placed on the wire between suction boxes, resting lightly upon the wire and the surface of the sheet. Its function is to flatten the top surface of the sheet and improve the finish. When the dandy roll leaves a mesh or crosshatch pattern, the paper is said to be "woven." When parallel, translucent lines are produced, it is said to be "laid." When names, insignia, or designs are formed, the paper is said to be "watermarked." Paper watermarks have served to identify the makers of fine papers since the early days of the art. A watermark is actually a thin part of the sheet and is visible because of greater transmission of light in its area compared with other areas of the sheet. Because light transmission can be varied by degrees, it is possible to produce watermarks in the form of portraits or pictures.

The final roll over which the formed sheet passes, before removal from the Fourdrinier wire, is the "couch roll." Prior to the transferring operation, the couch roll must remove water from and consolidate the sheet to strengthen it. In modern machines the couch roll is almost always a suction roll.

Press section

The press section increases the solids content of the sheet of paper by removing some of the free water contained in the sheet after it is formed. It then carries the paper from the forming unit to the dryer section without disrupting or disturbing sheet structure and reduces the bulk or thickness of the paper.

The first two functions are always necessary. Pressing always results in compaction, and this may or may not be desirable depending upon the grades being made.

Felts for the press section act as conveyor belts to assist the sheet through the presses, as porous media to provide space and channels for water removal, as textured cushions or shock absorbers for pressing the moist sheet without crushing or significant marking, and as power transfer belts to drive nondriven rolls or parts.

Woven felts of wool, often with up to 50 percent synthetic fibres, are made by a modified woolen textile system. Selected grades of wool are scoured, blended, carded, and spun into yarn. The yarn is woven into flat goods, leaving a fringe at each end. The ends are brought together and joined to produce an endless, substantially seamless belt.

Paper machine felts have a limited life ranging from about a week to several months. Their strength and water-removal ability is gradually lost through wear, chemical and bacterial degradation, and by becoming clogged with foreign material.

Press rolls must be strong, rigid, and well-balanced to span the wide, modern machines and run at high speed without distortion and vibration. Solid press rolls consist of a steel or cast iron core, covered with rubber of various hardnesses depending upon the particular service required. Suction press rolls consist of a bronze or stainless steel shell two inches (five centimetres) or more in thickness and usually covered with one inch of rubber.

Dryer

Paper leaving the press section of the machine has a solids content or dryness of 32 to 40 percent. Because of the relatively high cost of removing water by evaporation, compared with removing it by mechanical means, the

sheet must be as dry as possible when it enters the dryers. The dryer section of a conventional paper machine consists of from 40 to 70 steam-heated drying cylinders. After passing around the cylinders, the sheet is held in intimate contact with the heated surfaces by means of dryer felts.

Until recent years, relatively heavy, rather impermeable cloths composed of wool, cotton, asbestos, or combinations of these materials covered the dryer portion of the paper machine. Such cloths are termed dryer felts, though felting or fulling process is rarely used in their manufacture. Relatively lightweight, highly permeable cloths called dryer fabric also are employed.

For conventional dryer felts, cotton is still the most commonly used fibre, although it is seldom used alone. The main difference between the conventional dryer felt and the open-mesh dryer fabric is air or vapour permeability. High permeability is desirable because it allows the escape of the water vapour from the sheet.

For every ton of paper dried on the paper machine, approximately two tons of water are evaporated into the atmosphere. About 50 to 60 tons of air are required to remove the water vapour, with about 2,700 kilograms (6,000 pounds) of steam required by the dryers.

Finishing and converting. The rolls of paper produced by the paper machine must still undergo a number of operations before the paper becomes useful to the consumer. These various operations are referred to as converting or finishing and often make use of intricate and fast-moving machinery.

There are two distinct types of paper conversion. One is referred to as wet converting, in which paper in roll form is coated, impregnated, and laminated with various applied materials to improve properties for special purposes. The second is referred to as dry converting, in which paper in roll form is converted into such items as bags, envelopes, boxes, small rolls, and packs of sheets. A few of the more important converting operations are described here.

Coating

Paper has been coated to improve its surface for better reproduction of printed images for over 100 years. The introduction of half-tone and colour printing has created a strong demand for coated paper. Coatings are applied to paper to achieve uniformity of surface for printing inks, lacquers, and the like; to obtain printed images without blemishes visible to the eye; to enhance opacity, smoothness, and gloss of paper or paperboard; and to achieve economy in the weight and composition of base paper stock by the upgrading effect of coating.

The chief components of the water dispersion used for coating paper are pigment, which may be clay, titanium dioxide, calcium carbonate, satin white, or combinations of these; dispersants to give uniformity to the mixture or the "slip"; and an adhesive binder to give coherence to the finished coating. The latter may be a natural material such as starch or a synthetic material such as latex.

Equipment installed between dryer sections on the paper machine can apply the coating (on-machine coating), or it can be done by a separate machine, using rolls of paper as feed stock (off-machine coating).

The extrusion coating process, a relatively new development in the application of functional coating, has gained major importance in the past 20 years. The process is used to apply polyethylene plastic coatings to all grades of paper and paperboard. Polyethylene resin has ideal properties for use with packaging paper, being waterproof; resistant to grease, water vapour, and gases; highly stable; flexible in heat sealing; and free from odour and toxicity.

In the extrusion-coating machine, the polyethylene resin is melted in a thermoplastic extruder that consists of a drive screw within an electrically heated cylinder. The cylinder melts and compacts the resin granules and extrudes the melt in a continuous flow under high pressure. The resin is discharged through a film-forming slot die. The die has electric heaters with precision temperature controls to give uniform temperature and viscosity to the plastic melt. The slot opening can be precisely adjusted to control film uniformity and thickness.

The hot extruded film is then stretched and combined with paper between a pair of rolls, one of which is a rubber-covered pressure roll and the other a water-cooled, chromium-plated steel roll. The combination takes place so rapidly that a permanent bond is created between the plastic film and the paper before they are cooled by the steel roll.

Corrugating

The most widely used package for commodities and manufactured products is the corrugated shipping container. A corrugated box consists of two structural elements: the facings (linerboard), and the fluting structure (corrugating medium).

Linerboard facings are of two general types: the Fourdrinier kraft liner is made of pine kraft pulp, usually unbleached, in an integrated mill as a continuous process from the tree to the paper web, and the cylinder liner is made from reprocessed fibres, generally from used containers, providing a content of about two-thirds kraft.

The operation begins by unwinding the single-face liner and corrugating medium from holders, threading the medium into the fluting rolls, applying adhesive to the tips, and bringing the medium in contact with the liner to form a single-face web. Next, the single-face web passes another glue roll that applies adhesive to the exposed flute tips of the medium. The second face liner is brought in contact with the single-face web, and the combined board travels through a hot plate section between belts to set the bond, to a cooling section, and then to a slitter–scorer.

The world paper industry. The paper industry tends to be concentrated in those countries that are industrially advanced and have abundant supplies of fibrous raw material, especially wood. There is a large-scale international trade in wood pulp, pulpwood, and paper flowing from those countries with large forest resources to those countries with less or that are as yet undeveloped.

PAPER PROPERTIES AND USES

Used in a wide variety of forms, paper and paperboard are characterized by a wide range of properties. In the thousands of paper varieties available, some properties differ only slightly and others grossly. The identification and expression of these differences depend upon the application of standard test methods, generally specified by industry and engineering associations in the papermaking countries of the world.

Substance and quantity measurement. Weight or substance per unit area, called basis weight, is a fundamental property of paper and paperboard products. From the first uses of paper in the printing trades, it has been measured in reams, originally 480 sheets (20 quires) but now more commonly 500 sheets (long reams). The term ream weight commonly signifies the weight of a lot or batch of paper. Since the printing trades use a variety of sheet sizes, there can be numerous ream weights for paper having the same basis weight.

Measurement in reams

Table 2 gives basis weight ranges for some common papers.

To determine basis weight, the sample is brought to equilibrium under standard conditions (24° C or 75° F;

Table 2: Basis Weight Ranges for Common Papers

paper grade	basis weight*
Carbonizing tissue†	5¾–25
Facial tissue	9–10
Manifold tissue‡	10–20
Tea-bag tissue	8–12
Blotting paper	114–266
Bond	25–60
Book, uncoated	27–91
Book, coated two sides	45–109
Bristol, postcard	135
Envelopes	16–40
Greaseproof (glassine)	20–50
Newsprint	30–35
Multiwall bag stock	40–70
Kraft wrapping, most common	25–80
Cover stock, file folders	66–216

*Basis weight in pounds per 480 sheet ream, cut to 24 by 36 inches. †Paper to be impregnated with wax and pigment to make carbon paper. ‡Employed for making several carbon copies.

50 percent relative humidity). The paper specimens must consist of at least 10 sheets with a total area of not less than about 600 square centimetres (100 square inches). Since the properties of paper change with moisture content, all tests are conducted under standard conditions.

The caliper (thickness) of paper or paperboard in fractions of a millimetre or inch is measured by placing a single sheet under a steady pressure of 0.49 to 0.63 kilogram per square centimetre (seven to nine pounds per square inch) between two circular and parallel plane surfaces, the smaller of which has an area of 1.6 square centimetres (0.25 square inch).

The density or specific gravity of paper is calculated from the basis weight and caliper and may vary over wide limits. Glassine, for example, may be 1.4 grams per cubic centimetre and creped wadding, used for packaging breakables, only 0.1 gram per cubic centimetre. Most common papers are in the range of 0.5 to 0.7 gram per cubic centimetre.

Strength and durability. The strength of paper is determined by the following factors in combination: (1) the strength of the individual fibres of the stock; (2) the average length of the fibre; (3) the interfibre bonding ability of the fibre, which is enhanced by the beating and refining action; and (4) the structure and formation of the sheet.

Resistance to rupture when subjected to various stresses is an important property in practically all grades of paper. Most papers require a certain minimum strength to withstand the treatment received by the product in use; but even where use requirements are not severe, the paper must be strong enough to permit efficient handling in manufacture. Tensile strength is the greatest longitudinal stress a piece of paper can bear without tearing apart. The stress is expressed as the force per unit width of a test specimen.

Since the weight of the paper and the width of the test specimen affect the force of rupture, a conventional method of comparing inherent paper strength is the breaking length; that is, the length of a paper strip in metres that would be just self-supporting. This value varies from about 500 metres for extremely soft, weak tissue to about 8,000 metres for strong kraft bag paper, and to about 14,000 metres for sheets of paper made under ideal laboratory conditions.

Because some paper products such as towels, sanitary tissues, and filter paper are subjected to wetting by water in their normal use, wet tensile testing has become important. This test is essentially the same as that for dry tensile strength, except that the specimen is wetted. Paper that has not been specifically treated to produce wet strength possesses from about 4 to about 8 percent of its dry strength when completely wetted. By treating paper as described above, wet strength can be raised to about 40 percent of the dry strength.

Bursting strength

One of the oldest and most widely used strength tests for paper and paperboard is the bursting test, or Mullen test. It is defined as the hydrostatic pressure (caused by liquids at rest) necessary to cause rupture in a circular area of a given diameter. Other strength tests for which standard methods exist are tearing strength and folding endurance.

The resistance of paper to a bending force is evident in the various operations of its manufacture and in its many uses. The range in this property extends from very soft, flexible tissues to rigid boards. Thicker and heavier sheets tend to be stiff, whereas soft, flexible sheets are light and thin. Even at the same weight there is a considerable difference in stiffness, chiefly due to the compactness and the amount of bonding of the sheet.

Permeability

Because paper is composed of a randomly felted layer of fibre, the structure has a varying degree of porosity. Thus, the ability of fluids, both liquid and gaseous, to penetrate the structure is a property both highly significant to the use of paper and capable of being widely varied by the conditions of manufacture.

Sizing paper with vegetable materials and rosin-like substances has already been described. When paper began to be used for wrapping, consumers demanded sizing treatments that could protect the contents of the package from the effects of fluid transfer through the paper wrapping. In some instances complete impermeability was required. In another direction the use of paper as an absorbent medium for wiping up liquids, for filtering, and for saturating has created a demand for maximum wettability and permeability toward water and other fluids.

In certain types of packaging, paper must resist grease and oil penetration. The resistance of paper to the penetration of water can be increased by treatment of fibre with materials that lack affinity for water, with little effect upon sheet porosity, but the penetration of oil materials is little affected by such treatment. Oil and grease resistance is attained, in fact, by reduction in porosity. So-called greaseproof paper is made by beating an easily hydrated pulp to extremely low freeness, which results in a dense sheet with very little void space.

Absorbent papers such as towelling, sanitary tissue, and blotting and filter paper are normally made from lightly beaten stock. Since cellulose is naturally hydrophilic (*i.e.,* has a strong affinity for water), absorbent papers have a minimum of foreign materials associated with the fibre. Of particular importance are the wood rosins that may be present in pulp and produce a self-sizing effect, especially upon aging.

Optical properties. The most important optical properties of paper are brightness, colour, opacity, and gloss.

The term brightness has come to mean the degree to which white or near-white papers and paperboard reflect the light of the blue end of the spectrum (*i.e.,* their reflectance). This reflectance is measured by an instrument that illuminates paper at an average angle of incidence of 45° and a wavelength of 457μ (microns). Brightness measured in this way is found to correlate closely with subjective estimates of the relative whiteness of paper.

Opacity is one of the most desired properties of printing and writing papers. Satisfactory performance of such papers requires that there be little or no "show-through" of images from one side of the sheet to the other. Satisfactory opacity in printing papers requires that white mineral pigments be incorporated with the paper stock or applied as a coating.

The terms gloss, glare, finish, and smoothness are used in describing the surface characteristics of paper. The broad term finish refers to the general surface characteristics of the sheet. Smoothness refers to the absence of surface irregularities under either visual or use conditions. Gloss refers to surface lustre and connotes a generally pleasing aspect. Glare is used for a more intense reflection and a more unpleasant effect. Calendering and coating are important paper-treating methods that affect gloss. Gloss of paper is determined by measuring percent reflectance at a low angle of incidence, 15 degrees (75 degrees from the perpendicular).

Paper grades. *Bond paper.* Bond is characterized by a degree of stiffness, durability for repeated handling and filing, resistance to the penetration and spreading of ink, bright colour, and cleanliness. There are two groups of bond papers: rag content pulp and chemical wood pulp. Rag content bond may vary from 25 to 100 percent cotton fibre content. The principal uses of bond paper are for letterhead stationery, advertising pieces, announcements, leases, deeds, writs, judgments and other legal documents, currency, certificates, and insurance policies.

Types of bond papers

Book paper. Most book papers are made of various combinations of chemical wood pulp; and for lower priced grades, groundwood, semichemical, and de-inked wastepaper are also used. In addition to pulp, the "furnish" from which book papers are made contains various amounts of sizing, fillers, and dyes.

Uncoated book paper comes in four finishes: (1) antique or eggshell, (2) machine finish, MF, (3) English finish, EF, and (4) supercalendered. Antique has the roughest surface. High bulking pulps, such as soda pulp, are used and only slightly beaten in stock preparation. The sheet is lightly calendered (pressed between rollers) to provide a degree of surface smoothness while preserving the antique or eggshell appearance. Machine finish has a medium-smooth surface obtained for this finish from a calender stock at the dry end of the machine. Machine finish book is a relatively inexpensive general utility paper. It is princi-

pally used for books, catalogs, circulars, and other matter using line etchings. Machine finish book may be used for halftones up to a 100-line screen. English finish is a step higher in the book paper scale; this finish is distinguished from machine finish by a higher degree of stack beating, by greater pressure between the rollers of a machine calender, and by calendering at a greater moisture content of the sheet. Supercalendered book is the smoothest surface that can be obtained without coating. The finish is obtained by a special calendering operation after the paper leaves the paper machine. The supercalender presses the paper between successive sets of iron and compressed fibre rolls that make a smooth, compact printing surface. It is used for books, brochures, and magazines where halftone printing in the range of a 100–120 line screen is required.

Coated book papers are produced to create surfaces suitable for the printing of fine-screen halftones. Coated book paper must be uniformly smooth, receptive to printing inks, have high brightness and gloss, and be capable of folding without cracking.

Bible paper, as the name implies, was developed for lightweight, thin, strong, opaque sheets for such books as bibles, dictionaries, and encyclopaedias, which require minimum bulk. Bible papers are pigmented (loaded) with such pigments as titanium dioxide and barium sulfate and contain long fibres and artificial bonding agents to maintain strength.

Bristol. The general term bristol refers to a group of stiff, heavy papers with thicknesses ranging from 0.15 millimetre (0.006 inch) upward. These grades are made from various combinations of chemical wood pulp. The stock is beaten to a medium degree and usually well sized to prevent penetration of moisture. Increasingly important in recent years has been the use of bristols for the punch cards used in tabulating and sorting machines.

Groundwood and newsprint papers. These are printing and converting grades containing varying amounts of groundwood pulp, together with small percentages of chemical wood pulp for strength and durability.

For many years newsprint was virtually the only use for groundwood pulp, but more recently, due to improvements in the pulping process and to the introduction of a bleaching process for this pulp, a class of printing papers of broad utility has been developed. Magazines, paperbound books, catalogs, directories, and general commercial printing consume large quantities of these papers.

Groundwood papers are noted for an even, uniform formation and a high degree of opacity. These papers tend to be bulky and are receptive to printing ink. They do not have high whiteness and tend to turn yellow when exposed to light and after long aging.

Kraft wrapping. Kraft wrapping, a heavy stock used for paper bags, is used in greater volume than all other wrapping papers combined. It is composed of wood pulp in unbleached condition made from softwoods, usually pine. It is distinguished by outstanding tensile and tearing strength. Kraft wrapping is sized to retard wetting when exposed to water. For wrapping of wet materials, the paper may be given wet strength by treatment with special resins. Multiwall sacks of kraft paper are used for shipment of bulk materials.

Paperboards. The term paperboard is a general term that is descriptive of products which are 0.30 millimetre (0.012 inch) or more in thickness, made of fibrous materials on paper machines. Paperboard is commonly made from wood pulp, straw, wastepaper, or a combination of these materials.

Types of paperboard

There are three main types of paperboard: (1) boxboards, used for such products as food board, food trays, plates, and paper boxes; (2) container boards, for the manufacture of corrugated and solid fibre shipping containers; and (3) paperboard specialties, including such items as binders board, electrical pressboard, and building boards.

Sanitary papers. The group of papers known collectively as the sanitary grades include toilet tissue, towelling, facial tissue, and napkins. These grades are made from various proportions of sulfite and bleached kraft pulps with relatively little refining of the stock to preserve a soft, bulky, absorbent sheet. This sheet is further softened by machine creping, in which the wet sheet is pressed upon a smooth drying roll and subsequently removed by running against a flat stationary metal blade (doctor blade). The sheet is piled up upon itself, thus producing a creped effect. Facial tissue is dry-creped; that is, drying is complete on the drying roll before the creping doctor blade. Other papers are creped while still partially wet. Towelling is generally of heavier weight than the tissues and is usually creped while still wet. Napkins are of somewhat heavier weight than tissues. The plastic nature of paper fibres when slightly moist permits the reproduction of surface patterns by embossing to a remarkable degree. Paper napkins are an example of this art.

Because of the soft, bulky texture of sanitary papers, they are relatively weak. Since they are often exposed to wetting in use, they are often treated with resins to increase wet strength. (K.W.Br.)

BIBLIOGRAPHY

General. Encyclopaedic coverage of every aspect of the chemical industry is provided by: the *Kirk-Othmer Encyclopedia of Chemical Technology,* 2nd ed., 22 vol. (1963–70, and suppl. 1971); *Ullmanns Enzyklopaedie der technischen Chemie,* 3rd ed., 19 vol. (1951–69); and *Thorpe's Dictionary of Applied Chemistry,* 4th ed., 12 vol. (1937–56). The first edition of Thorpe, in 2 vol. (1890–93), is a historical classic.

Books covering the technology of the chemical industry include: R.N. SHREVE, *Chemical Process Industries,* 3rd ed. (1967); and E.R. RIEGEL, *Industrial Chemistry,* ed. by JAMES A. KENT (1962). Works covering the chemical industry from the industrial and statistical point of view include: CONRAD BERENSON (ed.), *The Chemical Industry: Viewpoints and Perspectives* (1963); JULES BACKMAN, *The Economics of the Chemical Industry* (1970); ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, *The Chemical Industry* . . . (annual); UNITED NATIONS, ECONOMIC COMMISSION FOR EUROPE, *Market Trends and Prospects for Chemical Products,* 3 vol. (1969); and ALFONS METZNER, *Die chemische Industrie der Welt,* vol. 1, *Europa,* vol. 2, *Uebersee* (1955).

Good coverage of sections of the chemical industry is provided by: W.L. FAITH, D.B. KEYES, and R.L. CLARK, *Industrial Chemicals,* 3rd ed. (1965), a description of the economics and technology of 137 chemical products; P.H. GROGGINS, *Unit Processes in Organic Synthesis,* 5th ed. (1958); and E. KILNER and D.M. SAMUEL, *Applied Organic Chemistry* (1960).

(A.St.)

Soaps and detergents. I. DAVIDSOHN *et al., Soap Manufacture,* vol. 1 (1953); A.M. SCHWARTZ, J.W. PERRY, and J. BERCH, *Surface Active Agents and Detergents,* vol. 2 (1958); A. DAVIDSOHN and B.M. MILWIDSKY, *Synthetic Detergents* (1967); J.G. KANE, *Soaps: Their Chemistry and Technology* (1959); W.W. NIVEN (ed.), *Industrial Detergency* (1955); J.L. MOLLIET, B. COLLIE, and W. BLACK, *Surface Activity: The Physical Chemistry, Technical Applications, and Chemical Constitution of Synthetic Surface-Active Agents,* 2nd ed. rev. (1961). Information may also be found in chapters on soaps and detergents in A.E. BAILEY, *Industrial Oil and Fat Products,* 3rd ed. by D. SWERN (1964); and in R.N. SHREVE, *Chemical Process Industries,* 3rd ed. (1967).

(A.S.D.)

Dyes. *History:* Authoritative articles on dyes and dyeing from classical antiquity to the present may be found in the journal CIBA *Reviews* (1937–). The commemorative volume, *Perkin Centenary, London: 100 Years of Synthetic Dyestuffs* (1958), contains a chapter on the life and work of Sir William Perkin by J. READ. The article by C. PAINE surveys the development of the synthetic-dye-making industry.

Industrial dye chemistry: E.N. ABRAHART, *Dyes and their Intermediates* (1968); and H.E. FIERZ-DAVID and L. BLANGEY, *Fundamental Processes of Dye Chemistry* (1949, Eng. trans. from the 5th Austrian ed. of 1943), are textbooks written from the industrial viewpoint. Articles on special topics (*e.g.,* wool dyes, acetate silk dyes, etc.) may be found in the *Kirk-Othmer Encyclopedia of Chemical Technology,* 2nd ed., 22 vol. and supplements (1963).

(E.N.A.)

Pharmaceuticals. *Pharmacopoeias:* WORLD HEALTH ORGANIZATION, *Pharmacopoea Internationalis,* 2nd ed. (1967, suppl. 1971); *The British Pharmacopoeia* (1968 and addenda); and *The United States Pharmacopeia XVIII* (1970), mainly as reference works and for examples.

Textbooks: L.M. ATHERDEN, *Bentley and Driver's Textbook of Pharmaceutical Chemistry,* 8th ed. (1969), an excellent source for chemical formulas and synthetic methods of production; ALFRED BURGER (ed.), *Medicinal Chemistry,* 3rd ed., 2 vol. (1970), a comprehensive work that includes the historical development of the subject and synthetic routes for many drugs; E.P. CLAUS

et al., Pharmacognosy, 6th ed. rev. (1971), includes historical material and information on production of crude drugs and of glandular extracts and purified hormones therefrom; WYNDHAM DAVIES, *The Pharmaceutical Industry* (1967), a study of the importance, scope, and economic aspects of the worldwide industry; ARTHUR GROLLMAN and E.F. GROLLMAN, *Pharmacology and Therapeutics,* 7th ed. (1970), covers the therapeutic uses of drugs; H.E. SCHULTZE and J.F. HEREMANS, "Analytical Methods in Protein Chemistry," in *Molecular Biology of Human Proteins,* ch. 2 (1966), gives more detail on human blood products; GLENN SONNEDECKER, *Kremers and Urdang's History of Pharmacy,* 3rd ed. (1963), a detailed treatment of the historical development of pharmacy and pharmaceuticals.

(L.M.At.)

Explosives. A.P. VAN GELDER and H. SCHLATTER, *History of the Explosives Industry in America* (1927), a highly detailed book covering the origin and development of all explosives of any importance throughout the world up to the date of its publication; M.A. COOK, *The Science of High Explosives* (1958), an advanced mathematical work devoted almost exclusively to theory, with a brief, interesting section on the history of explosives; C.H. JOHANSSON and P.A. PERSSON, *Detonics of High Explosives* (1970), an outstanding book describing the behaviour of high explosives, with emphasis on experimental data; E.I. DU PONT DE NEMOURS AND COMPANY, INC., *Blasters' Handbook,* 15th ed. (1969), a practical discussion of commercial blasting; INSTITUTE OF MAKERS OF EXPLOSIVES, "Safety in the Transportation, Storage, Handling and Use of Explosives" (1970), a pamphlet primarily designed for the guidance of the consumer; T.L. DAVIS, *The Chemistry of Powder and Explosives,* 2 vol. (1941–43), a general treatment of explosives with excellent coverage of pyrotechnics; T. URBANSKI, *Chemistry and Technology of Explosives,* 3 vol. (1967), an excellent treatment of these subjects, highly recommended; N.B. WILKINSON, *Explosives in History: The Story of Black Powder* (1966), a popular account, written primarily as a science supplement for high school students; U.S. BUREAU OF MINES, *Apparent Consumption of Industrial Explosives and Blasting Agents in the United States* (annual).

(N.G.J.)

Rubber. H.J. STERN, *Rubber: Natural and Synthetic,* 2nd rev. ed. (1967), is the most up-to-date comprehensive work on the subject. Another good basic text is M. MORTON (ed.), *Introduction to Rubber Technology* (1959). Two classic, but somewhat dated works are C.C. DAVIS and J.T. BLAKE (eds.), *The Chemistry and Technology of Rubber* (1937); and G.S. WHITBY (ed.), *Synthetic Rubber* (1954). See also L. BATEMAN (ed.), *The Chemistry and Physics of Rubber-Like Substances* (1963), which contains an excellent article on the botany of the rubber tree; and W.J.S. NAUNTON (ed.), *The Applied Science of Rubber* (1961). For history, see P. SCHIDROWITZ and T.R. DAWSON (eds.), *History of the Rubber Industry* (1952). Current statistical information may be found in the monthly periodicals, *Rubber News* and *Rubber Statistical Bulletin.*

(L.R.M.)

Plastics and resins. M. KAUFMAN, *Giant Molecules* (1968), a superbly produced introduction to the technology of plastics, fibres, and rubbers; E.G. COUZENS and V.E. YARSLEY, *Plastics in the Modern World,* rev. ed. (1968), the best known work on plastics; M.L. MILLER, *The Structure of Polymers* (1966), a monumental reference book; F.W. BILLMEYER, *Textbook of Polymer Science* (1962), a comprehensive classic, suitable for first courses in polymer chemistry at the graduate level; G. ODIAN, *Principles of Polymerization* (1970), a well-balanced treatment of all the types of reactions used to synthesize polymers; H. LEE, D. STOFFEY, and K. NEVILLE, *New Linear Polymers* (1967), a description of 10 new commercially available high-temperature polymers, developed since the mid-60s; W. MAYO SMITH (ed.), *Manufacture of Plastics,* vol. 1 (1964), information on all stages of the manufacture of commercial plastics, from monomer production to polymer finishing and shipping; J.A. BRYDSON, *Plastics Materials* (1969), one of the most readable, useful, and comprehensive books on plastics yet written; P.D. RITCHIE (ed.), *Physics of Plastics* (1965), a collection of 8 monographs by members of the Plastics Institute; *Modern Plastics Encyclopedia,* one of the most popular and well used reference book on plastics.

(E.G.C.)

Man-made fibres. H.F. MARK, N.G. GAYLORD, and N.M. BIKALES (eds.), *Encyclopedia of Polymer Science and Technology: Plastics, Resins, Rubbers, Fibers,* 9 vol. (1964–68), provides a comprehensive survey of fibre science and technology. J.G. COOK, *Handbook of Textile Fibres,* 4th ed., 2 vol. (1968), treats the production, properties, processing, and applications of natural and man-made textile fibres. *Industrial Fibres* (Commonwealth Secretariat 1968), is a survey of production, trade, and consumption relating to both natural and man-made fibres. Works primarily of interest to textile technologists include: W.E. MORTON and J.W.S. HEARLE, *Physical Properties of Textile Fibres* (1962); and J.W.S. HEARLE and R.H. PETERS (eds.), *Fibre Structure* (1963). Useful works for the general reader include: I.B. WINGATE, *Textile Fabrics and Their Selection,* 6th ed. (1970), and (ed.), *Fairchild's Dictionary of Textiles* (1967).

(J.G.Co.)

Paints and varnishes. R.R. MYERS and J.S. LONG, *Treatise on Coatings,* vol. 1, *Film-Forming Compositions,* 3 pt. (1967–71); successive volumes covering *Characterization, Vehicles,* and *Pigments,* offer the most comprehensive coverage of coatings since J.J. MATTIELLO (ed.), *Protective and Decorative Coatings,* 4 vol. (1941–42). The best monographs on coatings are D.H. PARKER, *Principles of Surface Coatings Technology* (1965); AND P. NYLEN and E. SUNDERLAND, *Modern Surface Coatings* (1965). See also the *Federation Series of Coatings Technology,* of the Federation of Societies of Paint Technology, Philadelphia, consisting of 20 pamphlets on basic subjects by various authors; and A.K. DOOLITTLE, *The Technology of Solvents and Plasticizers* (1954), for background. A *Raw Materials Index* is compiled by the National Paint, Varnish and Lacquer Association, Washington, D.C.; and *Physical and Chemical Examination of Paints, Varnishes, Lacquers and Colors* is produced by the H.A. Gardner Laboratory, Inc. in Bethesda, Maryland. Journals devoted to coatings science and technology include: *The Journal of Paint Technology,* Federation of Societies for Paint Technology, Philadelphia, Pennsylvania; JOCCA, the journal of the Oil and Colour Chemists' Association, London; *Farg och Lack,* Skandinavisk Tidskrift för Medlemsblad för Skandinaviska Lackteknikers Forbund (in Swedish), Copenhagen, Denmark; and an international journal entitled *Progress in Organic Coatings,* by Elsevier Sequoia S.A., containing review articles.

(R.R.M.)

Papermaking. AMERICAN PAPER AND PULP ASSOCIATION, *The Dictionary of Paper,* 3rd ed. (1965), is a compilation of terms and definitions relating to paper and paper manufacture. K.W. BRITT (ed.), *Handbook of Pulp and Paper Technology,* 2nd ed. rev. (1970), contains 61 articles grouped in 9 sections describing the most important features of pulp and paper manufacturing technology. A multivolume textbook that has gone through several editions is J.N. STEPHENSON (ed.), *Pulp and Paper Manufacture,* 4 vol. (1950–55). Each volume covers a particular area of the industry: pulp manufacture, stock preparation, paper manufacture, and finishing and converting. JAMES P. CASEY, *Pulp and Paper: Chemistry and Chemical Technology,* 2nd ed. rev., 3 vol. (1960–61), emphasizes the chemistry of papermaking and is noteworthy as a review of published literature with extensive references. A two-volume work intended as a textbook for students and prepared under the auspices of the Joint Textbook Committee of the Technical Association of the Pulp and Paper Industry is *Pulp and Paper Science and Technology,* ed. by C. EARL LIBBY (1962). S.A. RYDHOLM, *Pulping Processes* (1965), places primary emphasis on the chemical reactions of pulping and is a standard reference work in the field, of interest primarily to specialists in pulping. DARD HUNTER, *Papermaking: The History and Technique of an Ancient Craft,* 2nd ed. rev. (1957), is a classic account of handmade paper over the centuries since its invention in China early in the Christian era.

(K.W.Br.)

Extraction and Processing Industries

The extraction and processing industries are those industries concerned with the removal from the Earth of raw materials such as metals, minerals of organic origin, and certain nonmetallic nonorganic substances and with the conversion of these minerals into refined products. Coal and nearly all metals and other nonorganic industrial minerals are taken from the Earth's crust by either underground or surface mining. Oil and natural gas, on the other hand, are recovered from wells and boreholes drilled to considerable depths. In most cases, the raw materials obtained are processed for concentration, purification, or upgrading.

This article surveys basic mining and drilling operations and delineates refining or treatment processes for a variety of substances. Other functions, such as the transportation and storage of refined products, also are treated. For details on the properties and uses of metals, coal, petroleum, and gas, see CHEMICAL ELEMENTS and FUELS. (Ed.)

This article is divided into the following sections:

Metals and other nonorganic industrial minerals 334
- Ore deposits 334
  - General considerations
  - Nature, genesis, and classification of ore deposits
- Mining 339
  - Development of mining techniques
  - Prospecting and exploration
  - Underground mining
  - Surface mining
  - The mining industry
- Metallurgy 349
  - Development of metallurgy
  - Mineral dressing
  - Process metallurgy
  - Principal areas of metal processing
- Iron production 360
  - History
  - Mining and preparation for processing
  - Ore processing
  - The metal and its alloys
  - Iron powder and compounds
- Steelmaking 365
  - Mass production of steel
  - Modern steel technology
  - Properties of carbon and alloy steels
  - Steel metallurgy
  - World steel production
- Aluminum production 388
  - History
  - Ores and mining
  - Ore processing
  - Aluminum metal and its alloys
  - Aluminum powder and particles
  - Aluminum chemicals
  - Aluminum fabrication
- Copper production 396
  - History
  - Ores and mining
  - Ore processing
  - The metal and its alloys
  - Chemical compounds
- Zinc production 399
  - History
  - Ores and mining
  - Refining and recovery
  - The metal and its alloys
  - Chemical compounds
- Lead production 403
  - History
  - Ores and mining
  - Ore treatment and refining
  - Lead products
  - Lead alloys and compounds
- Production of other important industrial metals 406
  - Calcium 406
  - Chromium 407
  - Cobalt 409
  - Gold 410
  - Magnesium 413
  - Mercury 415
  - Nickel 417
  - Platinum-group metals 419
  - Silver 420
  - Sodium and potassium 422
  - Tin 424
  - Titanium 430
  - Tungsten 432
  - Uranium 433

Petroleum and natural gas 436
- Prospecting and exploration 436
- Drilling and completing wells 436
  - Rotary drilling
  - Well logging and drill-stem testing
  - Well completion
- Recovery of oil and gas 438
  - Secondary recovery of oil
  - Gas cycling
- Petroleum refining 439
  - General considerations
  - Separation into components
  - Alterations of molecular structure
  - Purification processes
  - Petroleum products and their uses
  - Refinery plant and facilities
- Processing, transportation, and storage of natural gas 447
  - Treatment and purification
  - Transportation and distribution
  - Gas storage

Coal 448
- Mining 448
  - History of coal mining
  - Underground coal mining
  - Surface coal mining
  - Transportation of coal
  - Productivity
- Processing 456
  - Development of coal-processing technology
  - Classification and evaluation of coals
  - Cleaning and other preparatory steps
  - Pyrolysis
  - Other processing techniques and products
  - Coal as a chemical

METALS AND OTHER NONORGANIC INDUSTRIAL MINERALS

Ore deposits

GENERAL CONSIDERATIONS

The word ore is used to mean a concentration of metals or metalliferous minerals that can be separated from the associated rock and mined at a profit. Present usage includes elements such as sulfur, when the sulfur is extracted from the mineral pyrites, and nonmetalliferous minerals such as fluorite and barite that occur in veins and require separation from the associated rock.

The profitability of mining any concentration of minerals can, and frequently does, change from year to year depending on a number of technological and economic factors. For these reasons the quantity and quality of ore in any particular concentration, or association of concentrations, must be defined on the basis of tonnages and

Grades

grades of proven, probable, and possible ore. Grades refer to the actual metal content of the ore, commonly expressed in pounds, ounces, or metric units of weight of a particular metal per unit weight of ore. Proven ore refers to a known volume of ore of defined limits that has been blocked by drilling. Probable ore relates to extensions of proven ore, and the limits are not precisely defined. Possible ore is classed as a prospective volume where spatial or geological relationships to known ore deposits warrant the assumption that mineral concentrations of economic value may be found.

Associated with the ore minerals in an ore deposit is worthless material called gangue, consisting of rock and unwanted minerals, which is mined with the ore minerals, subsequently separated in the milling processes, and discarded as dumps of waste. The unwanted minerals contained in these dumps may eventually become valuable, and, if the grade of such dumps is sufficiently high, they may be reworked at a profit. Also associated with an ore deposit is the country rock, or host rock, enclosing the deposit. The country rock can be of any type, although igneous and metamorphic rocks (those formed at high temperatures and pressures) are generally better host rocks for metalliferous ore deposits than are sedimentary rocks. Many exceptions are now known, but regardless of rock type, mineral prospecting and development generally require a certain amount of drilling, opencut operations, or underground mining before a reliable assessment can be made of the reserves present and the feasibility of mining them for profit.

Metals and metalliferous minerals

Some metals, including copper, silver, gold, and platinum, are mined in the native state. Others are rarely found in this state and only in quantities too small to be of economic importance. The outstanding deposits of native copper are in the Lake Superior area of Michigan, where the copper occurs in interbedded conglomerate (sedimentary rock with particles of diverse size) and basaltic lava flows, many of which are vesicular or amygdaloidal (that is, the minerals were deposited in cavities, or amygdules, after the flow formed). In the conglomerate the copper fills interstices and replaces matrix and pebbles; in the lava flows the copper fills the vesicles. Originally discovered by Indians, who manufactured ornaments from the malleable copper, and later noted by the Jesuits in the 17th century, these native copper deposits were first mined on a commercial basis in 1845.

Native silver was found in quantity in the Cobalt District of Ontario during the early years of the 20th century. Some of the ore was phenomenally rich, ranging from 200 to 1,000 ounces of silver per ton. Slabs of up to 725 kilograms (1,600 pounds) containing 40 percent by weight have been extracted from the Cobalt silver mines. The silver occurs in veins that have filled fissures in conglomerate and other sedimentary rocks.

Gold commonly occurs in the native state alloyed with copper and silver. It originally formed in fracture-filling veins associated with chalcopyrite, pyrite, and other sulfides, with silicate minerals such as quartz, sericite, and tourmaline, and with the carbonate minerals calcite and ankerite. Erosion of such veins results in the eventual transportation of the released gold by streams and rivers in the form of fine particles, flakes, and nuggets; these become concentrated, as in crevices of the bedrock underlying the gravel or at the upstream end of sandbars, and are called placer deposits.

Platinum is mined chiefly in the native state; it occurs alloyed with other metals of the platinum group, namely, osmium, iridium, palladium, rhodium, and ruthenium. It also occurs as platinum arsenides and sulfides. Invariably associated with nickel and chrome mineralization in basic igneous rocks (those rich in iron and magnesium), it is mined mainly in the Ural Mountains of the U.S.S.R., in the Sudbury District of Ontario, and in the Bushveld District of South Africa. Production from the U.S.S.R. is derived principally from placer deposits that have been derived from magmatic concentrations of platinum in olivine and pyroxene-rich igneous rocks called dunite and pyroxenite. In the Sudbury District, the platinum is associated with massive copper-nickel sulfide ores, which are the principal metals sought. In South Africa, platinum is mined from magmatic concentrations associated with copper-nickel sulfides in dunite pipes.

Metalliferous minerals are chemical compounds of metals, commonly sulfides and oxides, that are found in the rocks. Some are common and some are rare. Copper, nickel, lead, and zinc are chiefly mined as sulfides, some of which have a metallic appearance; iron, aluminum, and uranium are mined as oxides, many of which have an earthy appearance. Where these minerals are concentrated they form metalliferous deposits that can be defined as ore deposits.

Nonmetallic industrial minerals

Nonmetallic industrial minerals include those extracted from ore, such as diamonds and asbestos. They also include minerals that form deposits in themselves, such as beds of gypsum, salt, and sulfur. Other nonmetallic minerals used in industry are feldspar, quartz, mica, spodumene, gemstones, fluorite, barite, magnesite, sylvite, and many others. Spodumene and magnesite are also metalliferous because they are mined for the extraction of the metals lithium and magnesium, respectively.

A distinction must be drawn between industrial minerals and industrial materials such as soapstone, pumice, limestone, and shale, which are rocks, and clays such as kaolin, which are mixtures of hydrous aluminum silicates derived from the weathering of rocks.

Strategic minerals

Essential minerals that are in short supply, or that are found mainly in certain areas of the world from which they must be imported, are known as strategic minerals. The list of such minerals will vary from country to country and from time to time. Some countries are dependent on others for supplies of strategic minerals or their concentrates. Withholding such minerals from the world markets can cause serious shortages and consequent price rises on the international metal exchanges. The movement of minerals is affected not only by the amounts produced but also by political and economic considerations. Stockpiles of strategic minerals, concentrates, or the metals derived from them are hedges against the possibility of world shortages and high prices.

Worldwide exploration by mining companies, and the consequent discovery of new ore deposits, tends to modify the dependence of one country on another for strategic minerals. Antimony and titanium are examples of essential metals derived from strategic minerals. For many years China supplied two-thirds of the world's requirements for antimony, a metal that has the property of expanding upon cooling, a property that is used for type metal and hard lead alloys. China is still an important producer on the world market, but discoveries of the principal antimony sulfide mineral (stibnite) in other parts of the world have tended to lower the dependence on exports from China. Titanium is a lightweight, strong, extremely heat-resistant, low-corrosion, structural metal that is of increasing importance to the aerospace industry. It is mainly derived from the titanium oxide mineral rutile, which is supplied largely from beach sand-mining operations in Australia, South Africa, and Sierra Leone.

Conservation and future sources

Conservation of an ore deposit means the utilization of the deposit in such a way that the maximum amount of ore is ultimately mined. This means further that within reasonable economic limits the lower grades of ore are not sacrificed by mining methods that deplete too rapidly the higher grades for the sake of immediate monetary gain. Higher grades of ore are usually mixed with lower grades to average a uniform grade for the daily intake of a mill. The milling of high-grade ore only is seldom justified because it frequently results in mining practices that preclude the subsequent recovery of large tonnages of lower grade ore. Such practices are infrequent, and mine management practices now generally tend to encourage the long life of a mine.

Conservation is particularly important because ore deposits are nonrenewable natural resources. Although new ore deposits may continually be formed, they do so over periods of many thousands of years, commonly at depths of hundreds or thousands of metres; thus it is probable that the existing ore deposits are the only ones that man will ever be able to use. As known deposits have been

depleted in the past, new ones have been found. To date these have been either extensions of known deposits, deposits that crop out in explored parts of the world but which hitherto have escaped discovery, or unexposed deposits lying just below a surface covered by soil or vegetation. As time goes on fewer of the exposed deposits will be found, and discovery will depend increasingly on geophysical and geochemical methods to find the hidden deposits. In those parts of the world that have not been explored and mapped in some detail on the ground, exposed ore deposits may be found for many years to come. Whether or not such deposits are mined will depend on their economic viability at the time of discovery.

Technological developments increasingly have made possible the economic mining of low-grade deposits. This trend may continue and in the future make available large low-grade deposits that are not mined today. Technological developments involving dredging, suction, and other methods may in the future be used to mine phosphate nodules from deeper parts of the ocean floor, and rutile, zircon, ilmenite, and gold from ancient beaches now under a shallow sea. Magnesium has been commercially extracted in large quantities from seawater for many years, but the vast volume of seawater required to produce relatively small quantities of most metals is an inhibiting factor.

NATURE, GENESIS, AND CLASSIFICATION OF ORE DEPOSITS

Size, shape, and depth distribution. Ore deposits show considerable variations in size, shape, and depth of occurrence. They can range in size from a few hundred to many hundreds of millions of tons of ore. High-grade vein deposits of the copper minerals chalcopyrite, bornite, and chalcocite, for example, will generally have considerably less tonnage than low-grade copper porphyry deposits containing the same minerals. The latter commonly contain tens of millions (some contain hundreds of millions) of tons of ore averaging a fraction of 1 percent copper by weight. This fraction, or average grade, is an important factor in evaluating the worth of the deposit and the viability of developing it as a mine. The range of grades found that can economically be mined are commonly 0.4–0.8 percent copper, but the economic limit or cut-off grade may subsequently be lowered by technological developments.

Fractures, fissures, and veins

Ore deposits have a variety of shapes. They may be formed as discrete veins and lenses, irregular masses, pipes or carrot-shaped bodies, networks of fractures, disseminations of minerals in crushed rock and sedimentary beds, or as sedimentary deposits themselves. Their shapes depend on their nature and genesis and on the stratigraphic and structural conditions that existed during their time of formation. Subsequent earth movements such as faulting and folding may displace or contort ore bodies, resulting in the development of zones of fracturing or crushing within the ore body in which further mineralization can take place. This sequence of formation of the minerals of an ore deposit is termed paragenesis.

A fissure vein (Figure 1A) is a fracture in the country rock that has been filled with minerals. Some of these minerals may be valuable, whereas others constitute gangue. The vein material and the minimum quantity of adhering wall rock that must be mined with it constitutes ore. Mineralization of the space formed by a fracture may result in some cases from the forceful intrusion of magmatic differentiates; but most examples are hydrothermal in origin; that is, they have been formed by deposition of minerals from hot water and gaseous solutions moving through fractures in the country rock at depths of hundreds or thousands of feet. The force exerted by growing crystals, particularly by those aligned perpendicular to the walls of the fracture, can force apart the walls of a fracture, thus creating more space for growth. The asbestos minerals chrysotile, crocidolite, and tremolite commonly form in this way.

Because the original fractures occupied by veins were formed by tension within the rock, they commonly show a relationship to the vectors of tension, or directions of forces, that caused them. This arrangement, known as a fracture pattern, indicates the orientation of these vectors and, consequently, the sense of displacement of the country rock by faulting or folding. Where there have been several periods of earth movements it may be difficult or impossible to unravel the sequence of fracture patterns. A common type of vein fills tension or gash fractures, which may have an en echelon arrangment (essentially parallel and overlapping in plan view) due to the tension created by rotation of the rock mass in which they were formed. Other types of vein fillings are anastomosing (braided), subparallel, or fan-shaped, and commonly they are offset by numerous minor faults. The grade of ore within larger veins varies spatially. It may be higher near the footwall or hanging wall, and it may show variations with depth. Where the ore deposit consists of an interlacing network of veinlets less than an inch wide and a few feet long, the entire rock mass is mined. Concentrations of mineralization within the veins that can be mined as richer ore are termed ore shoots.

From A.M. Bateman, *Economic Mineral Deposits* (1954); John Wiley & Sons, Inc.

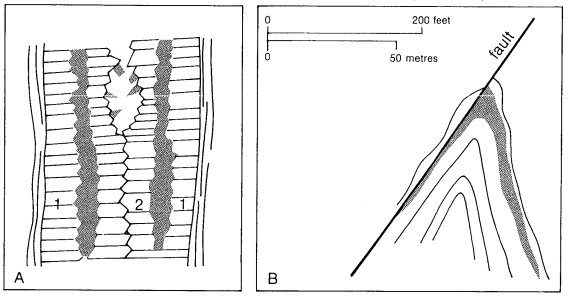


Figure 1: *Ore deposit shapes.*
(A) Fissure vein of (1) white quartz, sphalerite (shaded), and (2) amethyst formed in that order, illustrating paragenesis. Sphalerite crystals have formed in the vug (cavity) at top centre. (B) Saddle reef of gold ore (shaded) within a faulted fold of quartzite and slate.

Because they may be conspicuous at the surface, vein deposits have been mined since ancient times, particularly for copper and gold. Other minerals found in vein deposits in important mines in various parts of the world include native silver, lead, zinc, antimony, nickel, cobalt, mercury, uranium, and tungsten minerals.

Lenticular, irregular, and pipelike bodies

All veins terminate and many pinch out to form irregular, sheetlike bodies that are markedly elongate but lenticular in cross section. For this reason a clear-cut distinction cannot always be made between veins and lenses of ore. In some circumstances the ore deposits may be formed in distinct lenses or in lenticular bodies of quartz, examples being the saddle reefs (Figure 1B) of Bendigo, Australia. These gold-bearing deposits form inverted V-shaped lenses in the domes of anticlinal folds of Lower Ordovician (about 500,000,000 years ago) quartzites and slates. The quartzite tends to fracture rather than bend with folding, and the saddle reefs containing the ore are commonly found at or near the intersection of a fault with the trend of the fold. Saddle reefs containing gold ore are also found in association with folded beds of slate in the Salmon River District of Nova Scotia.

In some cases ore deposits form irregular masses that do not show any clear-cut stratigraphic or structural relationships. Some of these are the result of secondary and tertiary periods of deformation and mineralization that have obscured the original pattern. Other irregular masses, including some deposits of the manganese oxide minerals hausmannite and psilomelane, and the aluminum oxide minerals diaspore, gibbsite, and boehmite, which are collectively known as bauxite, have formed by a process of lateritization, or long-continued weathering *in situ* near the surface (Figure 2). Such irregular masses may form disconnected pockets in the underlying bedrock, partly connected pockets, or irregular sheets commonly ranging in thickness up to 15 metres (50 feet). Lateritic deposits of the nickel silicate minerals garnierite and noumeite are also found.

Important lateritic deposits of manganese, aluminum, and

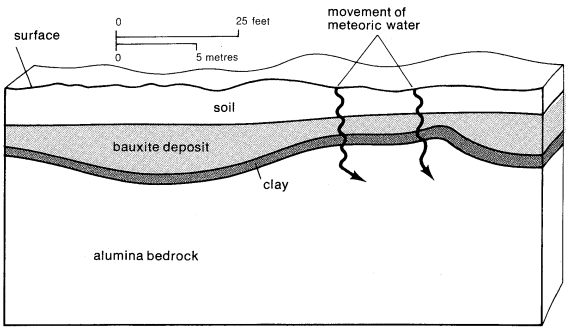


Figure 2: Relationship of bauxite deposit to alumina bedrock and overlying soil. Arrows indicate water percolation.

nickel are mined in the following countries: manganese in the U.S.S.R., India, Brazil, South Africa, Morocco, China, Australia, and Ghana; aluminum in Australia, Guinea, and Jamaica; and nickel in New Caledonia and Cuba.

Pipes are irregular, cylindrical or funnel-shaped ore bodies that commonly extend vertically, or nearly so, to depths of hundreds or thousands of feet. They have been formed either as volcanic necks (central conduit of a former volcano) or as crushed zones bordering the intersection of the planes of two steeply dipping faults, fissures, or dikes. The diamond pipes of South Africa are carrot-shaped volcanic necks, up to 700 metres (2,300 feet) deep, tending to be oval near the surface but becoming narrower at depth. They consist of fragmental and weathered basic igneous rock termed kimberlite in which the diamonds are disseminated. Mining was carried out initially by open-pit operations and subsequently by underground methods. Other pipes, sometimes referred to as chimneys, are formed by various combinations of structure and stress that cause disruption of the country rock, by upward stoping (removal of rock) caused by solutions, or by the washing of rock debris into limestone caves. Consisting of brecciated rock, or rubble, these pipes have been channels for mineralizing solutions. Ore-bearing pipes are commonly less than 30 metres (100 feet) in diameter but may be several hundred feet in length.

Ore bodies formed by mineralizing solutions also can consist of disseminations of minerals in crushed or finely fractured rock and in porous limestone and sandstone. Porphyry copper deposits are in the former category. These consist of disseminated copper minerals in the crushed or cracked periphery of intrusive igneous bodies of granitic rocks such as quartz monzonite and diorite porphyry. Mineralization also extends into the intruded and disrupted country rock. These porphyry copper bodies, which commonly contain valuable amounts of molybdenum sulfide minerals, are large, low-grade deposits that are usually mined by giant open-pit operations. Irregular in shape, but tending to be tabular, porphyry copper deposits may have dimensions ranging up to several thousand feet horizontally and several hundred feet vertically. Large porphyry copper deposits are mined in the United States, the Solomon Islands, and Canada.

Ore bodies in sedimentary rocks

Ore bodies are also formed as irregularly shaped mineralized masses that have developed in zones of higher permeability through which mineralizing solutions have permeated beds of limestone or sandstone. In the U.S., Missouri, Oklahoma, and Kansas have notably large deposits of sphalerite, associated with galena, a lead sulfide, in Cambrian dolomite. The ore deposits occur as patchy replacements in silicified zones of the dolomite; they are commonly tabular bodies 3 to 4.5 metres (10 to 15 feet) thick and a few hundred metres wide.

Lead–zinc deposits of similar shape are found in Triassic dolomites of the Silesian region of Czechoslovakia and Poland. An irregular and tabular body of lead–zinc ore is also mined from a Devonian (345,000,000 to 395,000,000 years ago) limestone reef at Pine Point in northern Canada.

Ore deposits also are found in sandstone beds where they have formed by precipitation of ore minerals in the interstices between the grains of sandstone. The loci of deposition have been determined by variations in permeability of the sandstone, the mineralizing solutions moving more freely through certain portions of the bed and possibly in some cases by selective precipitation caused by the catalytic action of particular variations in the sedimentary content of the sandstone. As these factors are stratigraphic, they result in the formation of fairly regular, usually elongate and tabular bodies of ore lying parallel to the bedding.

In Zambia the disseminated copper sulfide mineralization, chiefly chalcocite, chalcopyrite, and bornite, within Lower Roan Group feldspathic sandstones, is characterized by uniformity over considerable distances, the ore deposits following the configuration of the folded sandstone beds. Other examples of ore deposits formed by mineralization of particular portions of a sandstone bed are the uranium—vanadium deposits in the Jurassic (136,000,000 to 190,000,000 years ago) Morrison Formation and the Triassic Chinle Formation and Shinarump Formation of the Colorado Plateau in the United States. The deposits were selectively developed within fluvial channels, forming sinuous belts of elongate ore bodies. Commonly associated with carbonized fossil wood fragments and plant material, the minerals form a complex association that includes the uranium minerals uraninite and brannerite and the vanadium minerals roscoelite and montroseite. The ore bodies are elongate and tabular, or podlike, commonly ranging in thickness up to 6 metres (20 feet) and in length up to 304 metres (1,000 feet). They are more or less parallel to the bedding but locally discordant.

Some very large ore deposits consist of minute particles of minerals disseminated in layers parallel to the bedding planes of a sedimentary rock. A few examples are the lead-zinc deposits containing the sulfide minerals galena and sphalerite disseminated in argillaceous sedimentary rocks, in the Gulf of Carpentaria and Mt. Isa areas of Australia; the Kupferschiefer copper deposits of Germany, which consist of black, bituminous shale that mainly contains finely divided particles of the copper sulfide minerals bornite, chalcocite, and chalcopyrite but also contains galena and sphalerite; and the cherty banded iron formations called taconite, itabirite, or jaspilite, consisting of disseminated hematite, magnetite, siderite, or greenalite, found in many parts of the world, including the Lake Superior region. These banded iron formations (BIF) are tabular to sheetlike in shape, usually only a few feet thick, and in places may be traced along the layers of sediments for several kilometres.

Little is known about the precise depths at which ore deposits are formed or the maximum ranges of depth. Diamonds probably were formed as crystals in a basic kimberlite magma that was extruded upward, possibly from a depth of several kilometres, into volcanic pipes that form the ore deposits today. Sulfide minerals are formed within a depth range of several hundred to many thousands of metres. The limit of depth probably varies geographically and certainly depends on the limiting depth of fracturing and circulation of water vapour. Many ore deposits probably form within a temperature range of 300°–500° C (570°–930° F) and at depths of up to 16 kilometres (10 miles).

Geochemistry and mineralogy. The geochemistry of an ore deposit includes the physicochemical reactions involved in the formation of individual minerals, and mineralogy is the study of those minerals and their physical and crystallographic properties. Studies of the genesis of ore deposits, of the relationships of ore deposits to the depths and temperatures at which they formed, and to the petrology of country rock, therefore involve geochemistry and mineralogy. Many variations of physicochemical processes can be involved in the genesis of an ore deposit, which may involve several episodes of mineralization. A few examples are discussed here.

Physicochemical processes of importance

The Kiruna iron ore deposits in Sweden consist of magnetite and fluorapatite that appear to have first formed as a magmatic differentiate in the form of an immiscible magnetite-apatite liquid melt. Forceful intrusion of this melt into fractures in the country rock resulted in the formation of the ore deposits. The geochemistry of this deposit is clearly related to the physicochemical properties

of solutions in molten rocks and the minerals that can crystallize from them under certain conditions of temperature and pressure.

The hot volatile gases, including water vapour, given off by molten magma intruding the country rock are important mineralizing agents. Permeating the country rock through fissures and brecciated zones above the intruding magma, these aqueous and gaseous solutions deposit minerals as they become cooler. There are many examples of ore deposits formed as veins, stockworks, or disseminations by high-temperature (hypothermal), medium-temperature (mesothermal), and low-temperature (epithermal) solutions.

Residual ore deposits of the manganese oxides psilomelane and hausmannite, and of the aluminum oxides collectively referred to as bauxite, are formed by chemical processes in the weathering *in situ* of soil, subsoil, and underlying alumina-rich or manganiferous rock. These processes are not thoroughly understood but are known to be the result of meteoric water (water of atmospheric origin) penetrating downward, removing silica and leaving a concentration of manganese or aluminum oxides. The drainage of meteoric waters through the soil and into streams and rivers results in the solution and transportation of minute amounts of elements such as copper, lead, and zinc that are in the surficial material derived from the weathering of underlying mineralized rocks.

Geochemistry also is a factor in the formation of sedimentary ore deposits, referred to as syngenetic deposits. Some of these may have formed originally as mechanical concentrations of mineral particles within the sediments. There is evidence, for example, that the gold in the conglomerate beds of Witwatersrand in South Africa was emplaced as a placer deposit, although it subsequently has been partly redistributed. Similar evidence can be shown for the origin of the uranium in conglomerate beds of the Blind River area in Canada. But other deposits, with the possible inclusion of the lead–zinc mineralization in the Gulf of Carpentaria and Mt. Isa areas of Australia, appear to have formed by the contemporaneous deposition of metallic sulfides with the sediments. Deposition of lead and zinc sulfides in mud are known to be taking place in certain low areas of the Red Sea floor where water of higher salinity and density is concentrated. The physiochemistry and possible biochemistry involved in such processes is under study, and the stratigraphic implications, insofar as exploration for ore deposits is concerned, are far-reaching.

Temperatures and pressures of formation. Ore deposits have been formed under great ranges of temperature and pressure, depending on depth. Magmatic segregation deposits resulting from the differentiation within a molten magma of immiscible melts, such as that believed to have formed the magnetite–apatite deposits at Kiruna in Sweden, have originated under conditions of very high temperature and pressure. Another example of an ore magma is a sill-like body of magnetite and hematite with minor amounts of apatite at Lago Sur in northern Chile. These metallic facies (aspects) of magmas are not common at the eroded surface of the Earth's crust but may be more abundant at depth. It is not known whether they also include metallic sulfides or are confined to the oxides.

Allied to deposits formed by ore magmas are the tabular to lenticular chromite deposits of Cuba, which are thought to have originated as a crystal mush of ore magma and chromite crystals injected into the country rock. Other chromite deposits that seem to have had a similar origin are those of South Africa and Turkey.

The temperature ranges under which ore magmas are emplaced exceed 500° C (930° F), in many cases probably by several hundred degrees. The pressure ranges are in the order of several thousand pounds per square inch. Temperature and pressure ranges obtained during the formation of ore deposits of hydrothermal origin are generally lower.

T.S. Lovering, a U.S. economic geologist, says

> Of many physical factors that may induce precipitation, changes in temperature and pressure seem the only ones likely to be important during ore deposition. It is generally assumed that pressure increases directly with depth, and rock pressure at any specified depth below the surface is usually calculated as equal to the weight of the column of rock above it. Similarly, it has been assumed that the pressure of a mineralizing solution at any point below the surface corresponds to the hydrostatic head existing at that depth. It should be noted that pressure on a small area is transmitted in an outward flaring cone of diminishing intensity and it is thus possible for pressure to build up in a localized area that is greatly in excess of the rock pressure calculated for that depth. Although the pressure gradient in an open channel filled with liquid is approximately the hydrostatic head of any given depth, this is not true of a hypogene solution passing through a conduit that has marked constrictions between the point of entrance and the exit of the fluid. At any place the pressure is that of the hydrostatic head plus the driving force required to overcome friction. Below a marked constriction the pressure may be very much in excess of that above it, resulting in a rapid change of pressure. In any conduit the steepest pressure gradient would be at the greatest constriction. The places of greatest constriction may change from time to time because of mineral deposition, mineralization stoping, or intramineralization movements, and the sudden access to a new channel may move the position of the steepest pressure gradient either toward or away from the source of the solution.

Hydrostatic gradients and mineral precipitation

From the foregoing it is evident that the pressure–temperature relationships obtained during the formation of an ore deposit can be complex. Geothermal gradients vary from place to place and may be in excess of 1° C (2° F) per 30 metres (100 feet) of depth. Hydrostatic gradients depend on the density, or salinity, of the underground water and are commonly about 3 kilograms per square centimetre per 30 metres (roughly 40–45 pounds per square inch per 100 feet) of depth. Geostatic or rock pressures are two to three times the hydrostatic pressures at the same depth.

Paragenesis of ore minerals. Paragenesis means the sequential order in which various minerals within an ore deposit are formed. Variations in temperature, pressure, and chemical constituents of a mineralizing solution will result in the precipitation of different minerals at different times within the same deposit. Further complications in paragenesis arise where the ore deposit has been formed by more than one period of hydrothermal activity. A simple example of paragenesis can be seen where minerals line the walls of a fissure or cavity (Figure 1A). The earliest formed minerals grow inward from the wall rock forming an outer layer, and later formed minerals grow as inner layers or rings filling the void. More complex relationships are seen by examination of polished sections of ore under the microscope, which shows the manner and sequence in which crystallization and replacement of one ore mineral by another have occurred.

Sequence of mineral deposition

Studies of many hydrothermal ore deposits throughout the world have established a general sequence of mineral deposition based on mineral stability ranges. Commonly, the first minerals formed are silicates and carbonates, which constitute gangue. Of these, quartz and calcite, both of which are common in ore deposits, continue to be deposited during the later stages of development. The minerals chlorite, sericite, tourmaline, albite, adularia, barite, siderite, fluorite, ankerite, and rhodochrosite are also formed at an early stage. Secondly, the oxide minerals magnetite, ilmenite, chromite, specularite, and uraninite are formed. Thirdly, the minerals pyrite, arsenopyrite, cassiterite, tantalite, wolframite, molybdenite, pyrrhotite, pentlandite, and the nickel and cobalt arsenides are crystallized. Fourthly, the minerals chalcocite, bornite, chalcopyrite, sphalerite, galena, native silver, gold, the tellurides, stibnite, and cinnabar are precipitated. This arrangement shows that in general the oxides are deposited early; that the sulfides and arsenides of iron, nickel, cobalt, and molybdenum are contemporaneous with or slightly younger than the oxides; and that the sulfides of lead and zinc are even younger. Later, the native metals and tellurides are formed, and the last to be deposited are antimony and mercury sulfides.

Related to paragenesis is the zoning of ore deposits (Figure 3). As mineralizing solutions move through channels in the rock they undergo changes in temperature, pressure, and chemical composition that may result in the deposition of different mineral concentrations at increasing distances along the course of deposition, and presumably

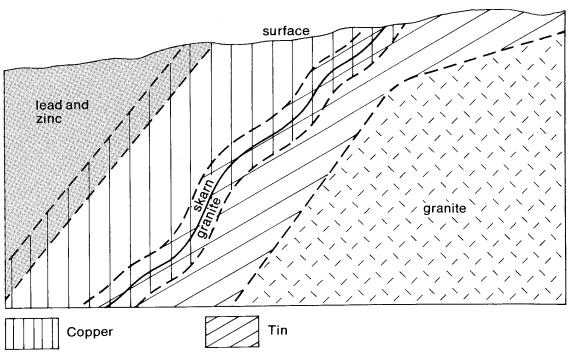


Figure 3: Relationship of mineralized zones to the granite from which the mineralizing solutions are believed to have emanated.

from the magmatic source. This zoning is common but is not always present in ore deposits. In general, minerals of tin, tungsten, and bismuth lie closer to the magmatic sources of the mineralizing solutions than do those of the copper minerals. Lead and zinc minerals are more distant; native gold and silver, and minerals of gold and silver, are still farther away; and the most distant of all from their source are the minerals of antimony and mercury.

Genetic and other classifications. To be useful, a classification of ore deposits must be simple and usable in the field. Unfortunately, ore deposits cannot always be clearly defined by type. In many cases they have been formed by complex physicochemical processes, and in some cases their genesis is still in question. Classifications can be derived from the agents that form ore deposits, such as magmas, hydrothermal solutions, pneumatolytic solutions, syngenetic processes such as precipitation in a body of water, weathering, and mechanical processes. Other classifications can be built on temperature and depth relationships, on stratigraphic and structural relationships, and on mineralogical and chemical relationships. Each classification may satisfy certain categories of ore deposits, but any comprehensive classification of all ore deposits must of necessity be general and, consequently, of limited application.

Bases and limitations of classifications

There are three broad classifications of ore deposits in use. The Niggli classification has two main divisions, namely plutonic, or intrusive, and volcanic, or extrusive. The former division has three categories: orthomagmatic, pneumatolytic to pegmatitic, and hydrothermal. The Schneiderhöhn classification has four main divisions: intrusive and liquid-magmatic, pneumatolytic, hydrothermal, and exhalation. The Lindgren classification has two primary divisions: deposits produced by chemical processes and deposits produced by mechanical processes. The first division contains three categories: ore deposits formed by differentiation in magmas, ore deposits formed in bodies of rock, and ore deposits formed in bodies of water.

Whatever classification is used, the critical point is that it must serve a purpose. Each classification has its limitations and is subject to modifications. (C.E.B.C./Ed.)

Mining

DEVELOPMENT OF MINING TECHNIQUES

Early history. Archaeological discoveries indicate that mining was practiced in prehistoric times. Apparently, the first mineral used was flint, largely for weapons. Certain minerals, such as ochre, black pigments (manganese ore), or chalk, were also mined and used as pigments.

Placer deposits

Gold was one of the first metals utilized, being mined from streambeds of sand and gravel where it occurred as a pure metal (placer deposits) because of its chemical stability. Although chemically less stable, copper occurs in native form and was probably the second metal discovered and used. Silver was also found in a pure state and at one time was valued more highly than gold.

According to historians, the Egyptians were mining copper on the Sinai Peninsula as long ago as 3000 BC, although some bronze (copper alloyed with tin) is dated as early as 3700 BC. Iron is dated as early as 2800 BC; Egyptian records of iron-ore smelting date from 1300 BC. Found in the ancient ruins of Troy, lead was produced as early as 2500 BC.

One of the earliest evidences of building with quarried stone was the construction (2600 BC) of the great pyramids in Egypt, the largest of which (Khufu) is 236 metres (776 feet) along the base sides and contains approximately 2,300,000 blocks of two types of limestone and red granite. The limestone is believed to have been quarried from across the Nile. Blocks weighing as much as 14,500 kilograms (32,000 pounds) were transported long distances and elevated into place, and showed precise cutting that resulted in fine-fitting masonry.

Agricola's *De re metallica*

One of the most complete early treatments of mining methods in Europe is by the German scholar Georgius Agricola in his *De re metallica* (1556). He describes detailed methods of driving shafts and tunnels. Soft ore and rock were laboriously mined with a pick and harder ore with a pick and hammer, wedges, or heat (fire setting). Fire setting involved piling a heap of logs at the rock face and burning them. The heat weakened or fractured the rock because of thermal expansion or other processes, depending upon the type of rock and ore. Crude ventilation and pumping systems were utilized where necessary. Hoisting up shafts and inclines was done with a windlass; haulage was in "trucks" and wheelbarrows. Timber support systems were employed in tunnels.

Development of modern excavating technology. Great progress in mining was made when the secret of black powder reached the West, probably from China in the late Middle Ages. This was replaced as an explosive in the mid-19th century with dynamite, and since 1956 both ammonium nitrate fuel-blasting agents and slurries (mixtures of water, fuels, and oxidizers) have come into extensive use. A steel drill with a wedge point and a hammer were first used to drill holes for placement of explosives, which were then loaded into the holes and detonated to break the rock. Experience showed that proper placement of holes and firing order are important in obtaining maximum rock breakage in mines.

The pneumatic hammer

The invention of mechanical drills powered by compressed air (pneumatic hammers) increased markedly the capability to mine hard rock, decreasing the cost and time for excavation by several-fold. It is reported that the Englishman Richard Trevithick invented a rotary, steam-driven drill in 1813. Mechanical piston drills utilizing attached bits on drill rods and moving up and down like a piston in a cylinder date from 1843. In Germany in 1853, a drill that resembled modern air drills (see below) was invented. Early drills required frequent repairs, as many as 200 being required to keep 16 in operation. Piston drills were superseded by hammer drills run by compressed air, and their performance improved with better design and the availability of quality steel.

Developments in drilling were accompanied by improvements in loading methods, from handloading with shovels to various types of mechanical loaders. Haulage likewise evolved from human and animal portage to mine cars drawn by electric locomotives and conveyers and to rubber-tired vehicles of large capacity. Similar developments took place in surface mining, increasing the volume of production and lowering the cost of metallic and nonmetallic products drastically. Large stripping machines with excavating wheels used in surface coal mining are employed in other types of open pit mines.

Underground excavation technology continued to advance during the 20th century with the development of the continuous miner—a machine that combined excavating, loading, and hauling into an ongoing mining process. The drilling and blasting steps of conventional mining were eliminated by the continuous miner, as were the subsequent time lags between the loading and hauling sequences. The diffusion of the continuous miner coincided with the widespread use of the longwall-mining system in which ore was sliced, or sheared, loaded, and hauled in a continuous, or near continuous, manner by a conveyor-driven machine that paralleled a wall of rock. The later introduction of self-propelled continuous miners was instrumental in the development of the shortwall-mining

system and also caused continuous mining operations to supplant the conventional approach used in many room-and-pillar mining systems.

Underground and surface mining both experienced increased use of hydraulic power in the 20th century. Water had been used in mining to separate valuable pieces of ore from dirt and rock, particularly in placer mining operations for gold (see below *Production of other important minerals: Gold*). During the late 20th century hydraulic power began to be used for extracting and transporting ore in underground and surface operations.

The Americas. Perhaps the first mining by Europeans in the Americas after Columbus was for copper in Cuba in 1554. Many small deposits of copper and iron were discovered in the 13 colonies that became the United States; all were producing small amounts of iron by the end of the Revolutionary War. Iron was first produced from the extensive Lake Superior deposits in 1846. Though much of the richer iron ore had been exhausted by the early 1970s, lower grade taconite ore is being mined and concentrated by special processes.

California gold rush

While gold has been found in several places in North America, and the existence of gold in California was previously known, its discovery in 1848 had marked effects upon immigration and economics in the United States. The first gold mined was in placer stream deposits that proved to be extensive; some deposits were later mined by dredges. The gold deposits were traced upstream to the source from whence they had been eroded, which led to the discovery of the Mother Lode, a rich quartz vein of unusual length in the Sierra Nevada Mountains. Gold was later discovered in Alaska in 1880 on the beaches and in streams and in vein deposits in the mountains inland, in both Alaska and British Columbia.

One of the richest silver mines in America was located at the Comstock Lode in 1859, near what is now Virginia City, Nevada, an area that had been worked for placer gold. Several valuable deposits were discovered in Colorado beginning in 1859, yielding gold and silver from districts known as Leadville, Cripple Creek, and others.

Lead mines existed in what is now Missouri as early as 1687. The mines of southeast and southwest Missouri have supplied large amounts of lead and other metals. More recent discoveries in south central Missouri indicate the possibility of a lead belt across the state.

Lead was discovered in Bingham Canyon, Utah, in 1863, leading ultimately to the discovery and development of the great Utah copper open-pit mine. The successful exploitation of massive low-grade copper deposits of the western United States and elsewhere in the world was made possible by the discovery of the process of mineral flotation, which permits the economical concentration of the ore before it is smelted.

Large-scale development of mining in Canada has been fairly recent. Iron ore was found in Newfoundland and gypsum in Nova Scotia. Much of the country's mineral wealth is found in the Canadian Shield (Precambrian rocks, older than 570,000,000 years). Large deposits of ilmenite (titanium-iron ore, Tahawus, New York), gold (Kirkland Lake), copper (Sudbury Basin), and iron (Lake Superior) are currently being mined, as well as gold deposits in the West (Yellowknife).

In South America, Chile is noted for its deposits of sodium nitrate (extensively exploited 1880–1930) and copper. The economic manufacture of synthetic sodium nitrate has, however, decreased the value of Chilean deposits. Copper was mined by the Spaniards as early as 1601, but most production now comes from massive low-grade deposits. Chile also has several rich iron-ore deposits, as well as some sulfur. Bolivia is the world's second-largest producer of tin. Gold, silver, and other metals and minerals are produced in other Latin American countries.

Australia. Gold was discovered in 1851 in New South Wales and later in Victoria, but after 1857 opportunities for miners to make a living decreased. The Broken Hill, Mount Isa, and Kalgoorlie gold mines have been most productive. Copper, silver, zinc, and tin are also produced in varying amounts. Australia leads the world in production of opals. After 1960, active exploration led to discovery of extensive new deposits of metallic and nonmetallic mineral deposits.

Europe. France has notable iron deposits, while England's deposits of coal and iron long made it a principal producer of steel. Open-pit operations in western Europe include mining of iron, limestone, granite, and shale. In France the Minette iron deposit, previously exploited by open-pit methods, is now mined by underground techniques. Ammonium nitrate fuel oil and slurry blasting agents have been widely adopted in some European countries, though their use is restricted in some. Stripping, mining, loading, and transport are practiced in much the same manner as in the Americas, with some differences due to geography and local conditions.

Africa. The continent of Africa is the source of about 70 percent of the world's diamonds (major producers are South Africa, Botswana, South West Africa/Namibia, Angola, Zaire, Ghana, Sierra Leone) and more than one-half of the gold (South Africa). Cobalt (more than one-half of world supply; Zaire and Zambia), chromite (about one-quarter; South Africa), and manganese (about one-tenth; South Africa and Zimbabwe), all important in strengthening steel, are other products of this area. The huge copper-mining operations in the Shaba Province of Zaire became a source of political difficulty following that nation's independence in 1960. In western Africa, high-grade iron deposits, magnetite and hematite, are mined by the open-pit method. Other mining practices generally resemble those employed in Europe.

Soviet Union. Besides extensive reserves of coal, the Soviet Union probably has a greater variety of minerals than any country outside the United States. The Krivoy Rog deposits, similar to the Lake Superior deposits, have produced large tonnages of iron, which is also found in the Ural and Crimea areas. The Soviet Union leads the world in manganese reserves and has deposits of copper, lead, zinc, and aluminum, as well as nickel, gold, and platinum. Nonmetals include phosphate salts for production of fertilizer, asbestos, talc, magnesite, salt, and gems.

PROSPECTING AND EXPLORATION

In the past, many deposits of valuable minerals were found by means of outcrops, but few such deposits near the surface remain to be discovered. Currently, most new deposits are located by means of geological techniques, often supplemented by geophysical and geochemical methods. Prospecting is also being conducted to find deposits under the ocean. An understanding of the processes by which deposits were formed is a valuable aid in prospecting, since geologic controls can furnish information as to the likely position of ore.

The first step in modern prospecting is the compilation of all existing geological information on the area of interest. A detailed study and mapping of surface features follows; these are correlated with known structures that are projected to depth where possible. All evidences of mineralization, favourable structures, or environment for deposits are carefully recorded and employed as indicators of ore. If evidence of the possible presence of valuable deposits is found, a drilling program is carried out to obtain more information about subsurface geology and to intersect ore structures if possible. Data obtained in prospecting may also be employed in the exploration or evaluation stage that follows.

Indicators of ore

Types of geological evidence are both regional (of broad extent) and local (of limited extent). For metallic deposits, the presence of batholiths (large, irregular masses of igneous rock that have melted or forced their way into surrounding strata), certain volcanic and igneous rocks, the age of rock relative to mineralization periods, presence of favourable rocks, and climatic conditions may be important. Local surface features that can provide valuable clues are presence of rock alteration, trace mineralization in surface rocks, and quartz and other gangue (non-ore-bearing) minerals. Rock fractures, particularly fault zones, are important, as well as beds or zones that are favourable for mineralization. Many deposits "bottom out" with depth because pressure or other factors were unfavourable for mineral deposition.

Geochemical prospecting is the chemical analysis of rocks to determine the presence, usually in very small amounts, of possible trace elements. The patterns of dispersion of certain elements have proven to be valuable in conjunction with geologic data in locating and defining the limits of valuable ore bodies.

Geophysical methods employed vary with the local conditions. Seismic techniques, which involve transmission of low-frequency waves through the Earth's crust and recording of their behaviour, are utilized to determine underground structure, particularly in horizontal-bedded deposits. Magnetic methods determine the presence of magnetic bodies. Electrical methods indicate the presence of conductive materials. Gravity methods detect the presence of more dense bodies. Geiger counters ascertain the presence of radioactive minerals.

Obtaining cores

Prospect drilling is often performed with diamond drills to obtain cylindrical specimens of the rock and ore, called cores. The drill bit is a hollow cylinder embedded with commercial-grade diamonds on one end and attached to a core barrel. The drill and core barrel are rotated with the bit against the rock by means of hollow drill rods through which water flows to keep the bit cool and wash out the cuttings. Diamond drilling is of greatest value where the rock can yield intact cores that can then be employed to map underground geology or to give solid samples of valuable rock or ore.

When a valuable deposit is discovered, prospecting phases into exploration, which is planned to yield necessary data concerning size, shape, and value to furnish information to begin development. Careful trenching (digging a trench to obtain samples of ore lying near the surface), drilling, and sampling are carried on to give reliable information before large capital expenditures are made or the necessary plant is purchased.

Modern computers are often employed to calculate possible risks, profits, and other results of further exploration. Sample values obtained for the material to be mined are usually evaluated by use of the theory of probability, with average values weighed by the amount of ore they represent or by a similar method.

A mineral deposit is a "wasting asset," and, hence, its present value is calculated on the basis of the fact that the worth of the mine will be zero at the end of a certain period of time. The cost of capital equipment must be amortized accordingly, and a reliable estimate of the life of the mine is made. Rates of interest and return of investment are calculated to ensure adequate profits before the time a deposit is mined out.

UNDERGROUND MINING

Choosing a mining method. A space in an underground mine from which ore has been removed is called a stope. Depending upon conditions, stopes require varying amounts of support to prevent the earth and rock above (called overburden, or capping) from collapsing. In some mines, ore is extracted from several different underground levels. In this case, intervals must be left between the levels to prevent collapse. Figure 4 is a composite sketch illustrating various terms used in underground mining.

From R.S. Lewis, *Elements of Mining*, 3rd ed. (© 1964); used with permission of John Wiley & Sons, Inc.

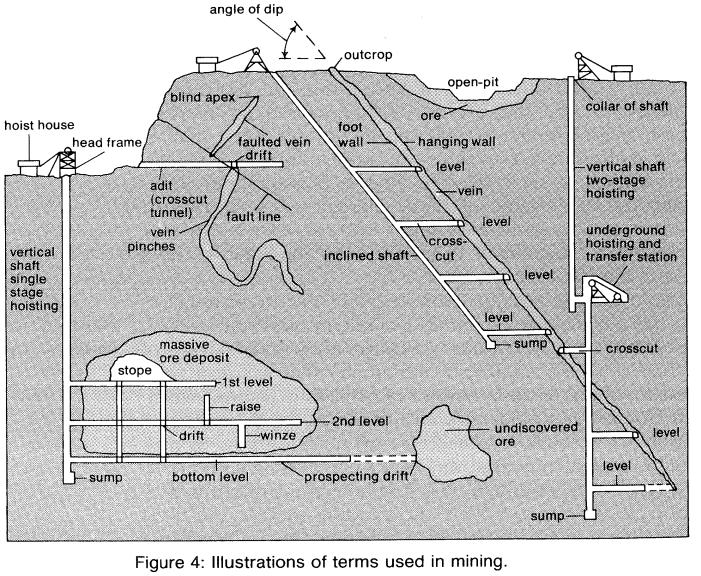


Figure 4: Illustrations of terms used in mining.

Rock mass stability

The stability of rock masses in which ore deposits occur in many cases determines the method of mining employed, although other factors may dictate the method used. Geologic and rock-mechanics data (see below) are the most reliable guides for evaluation of the adaptability of mining methods to a given deposit. Ideally, a method of mining is employed that will yield the largest net financial return consistent with safety and optimum extraction of valuable minerals.

Rock mechanics. The term rock mechanics is used to define the study of the behaviour of rock materials as they relate to man-made engineering projects that involve rock in place. Thus, studies are made of the physical properties of rock specimens that are taken from a rock mass. Many intact rocks behave elastically under load, and the theories of elasticity and the strength of materials are applied. Compressive and tensile strength are measured, as well as the strength under confining pressure.

Dynamic rock mechanics deals largely with waves caused by explosives and their effects in rock breakage, with rock breakage in drilling, with mechanics of rock bursts (sudden collapses), and with effects of large, underground explosions.

Methods have been devised to measure stresses, strains, and displacements in rock, both in open pits and underground. These studies are made to determine the stability of rock structures, support requirements, the behaviour of subsiding (settling) masses in caving and longwall methods of mining, and for increasing the efficiency and safety of mining.

Evaluating geologic features. Geologic details are carefully evaluated so that preliminary openings may be placed to best advantage; permanent openings may be located in stable rock; stopes may be planned for optimum stability and ground control; mine openings may be maintained at minimum cost; underground water may be avoided; overall mining methods, including caving procedures, may be planned; and safety can be achieved.

Geologic features that influence the choice of mining method include the size and shape of the ore body; the depth and type of overburden; the location, strike (horizontal course), and dip (the angle at which the deposit is inclined from the horizontal) of the deposit; the strength and physical character of the ore; the strength and physical character of the surrounding and overlying rock; the presence or absence of aquifers (water-carrying rocks); and the grade and type of ore.

Engineering-design factors. Important factors in engineering design of underground mines include strength of ore and wall rocks; shape, horizontal area, volume, and boundary regularity of the ore body, and the thickness, dip or pitch or both of the deposit and ore shoots; continuity of the ore; depth below the surface; nature of the overburden; and position of the deposit relative to surface improvements, drainage, and other mine openings. It is usually necessary to minimize those factors of lesser importance in order to satisfy the requirements of the more critical ones.

The strength of both ore and wall rocks fixes the safe size of openings, the length of time they will stand, and the support required. Stability depends not only on the strength of the solid rock but also on the existence of fractures, joints, and other discontinuities and their geometric arrangement.

Ore body shapes

Shapes of ore bodies vary from massive to tabular (having two parallel surfaces) and from bedded deposits to pipes, shoots, dikes (fissures in older rock filled with igneous material), and sills (strata of igneous rock found between beds of other rock), depending on their origin and the character of the host rock structure. The attitude (inclination) of a deposit directly influences the mining method. A tabular deposit dipping at a high angle, for

example, may be removed without use of supports if the walls are moderately strong. The roof of a similar deposit dipping at a low angle, however, may need support.

A regularly shaped deposit may require less support. Mining operations in an irregular deposit will produce stopes with uneven walls, overhanging slabs, and projecting columns of wall rock that may require support if the rock is fractured or broken.

A small ore body in firm rock may be mined without support, but larger ones in the same kind of groundmass may require it. There is a limit to the length or width of an unsupported rock arch or span that will possess sufficient strength to stand.

Usually, as depth increases, a method must be changed because openings require more support. Shallow deposits are subject to relatively low lithostatic (weight of overlying rock) pressures. The vertical pressure on mine openings usually increases proportionately with depth. The direction of pressure in walls or roofs of an excavation depends on the attitude of the deposit and local geologic factors and must be considered in stope support. In a steeply dipping tabular deposit, waste (broken-rock) filling will furnish adequate support against side pressure, whereas in a wide deposit, in which the critical pressure acts downward, filling cannot be relied on to prevent all movement. Thus, a cut and fill method can be employed where the pressure is from the sides and the back is strong and self-supporting, but if the critical pressure is vertical, square setting (a heavy square framework of timbers) supplemented by filling or some other method of support is necessary (see below *Underground opening support*).

Geology, rock structure, and geometry. In addition to their value in exploration, geological theories of the genesis of mineral deposits are also valuable guides in mine operation, influencing prospecting, exploration, development, and exploitation. In prospecting, geology is a vital tool in finding new ore bodies and is most effective in the hands of an experienced mining geologist. In exploration and development, the mine engineer should have a knowledge of strength of materials, rock mechanics, and of the structural features of the deposit that will affect the production of ore and the margin of profit. Most of these factors can be treated in only a semiquantitative manner, because the behaviour of rock structures is complex. Design details may also be considered in relation to bedded deposits, veins, and massive deposits.

Bedded and sedimentary deposits that have not been disturbed by faulting, folding, or other action are, like flat-lying beds of coal, simple to mine. Support problems are often easily solved, and a systematic method of pillar support is employed. Poor rock in the immediate roof, however, such as shale, creates hazardous conditions. The most serious problem in this type of mine may be determination of the number, size, and position of pillars required for support.

Dangers of shale roofs

Many commercial mineral deposits are found in or closely associated with veins in fractures, shear, and fault zones. Deposits vary from tabular veins to pipelike bodies and shoots and may be regular or irregular, wide or narrow, continuous or discontinuous. Thus, many different methods are used to mine them.

The large, disseminated copper-ore bodies of the western United States are typical massive deposits; those with enough overburden or capping to prohibit mining by open-pit methods are often mined by block-caving (see below). Deposits of porphyry coppers are usually extensive, flat-lying, and relatively regular in outline; though low-grade, their reserves are usually measured in millions of tons. The concentration of valuable minerals has been the result of geologic changes that left behind a weak, leached overlying rock mass.

Mining operations. These include use of explosives, drilling, hoisting, loading and haulage, and ventilation and lighting.

Explosives. The theory of detonation of high explosives is based upon the laws of conservation of mass, momentum, and energy applied to plane shock waves, combined with thermodynamics, equations of state for gases, chemical equilibria, and material-balance equations. In a detonation wave, the energy of chemical reaction is utilized to heat the detonation products and to maintain the thermodynamic processes of shock-wave propagation at high temperatures (to 5,500 K [9,400° F]) and pressures (to 200,000 times atmospheric pressure). The distinguishing feature of a detonation is that it is a reactive shock—*i.e.,* a shock wave supported by an accompanying chemical reaction.

Most common commercial explosives are made up of compounds containing carbon, hydrogen, nitrogen, and oxygen, some of the compounds serving as fuels and some as oxidizers (agents that bring about chemical combination with oxygen). The compounds must be of such a character that they will react rapidly, releasing chemical energy at a very high rate. Thus, one stick (225 grams, or eight ounces) of an average-energy dynamite may have a work capacity of 40,000 horsepower.

Ammonium nitrate–fuel oil explosives

Development of ammonium nitrate-fuel oil (AN–FO) blasting agents in the late 1950s has reduced explosive energy costs to as little as one-third the cost of dynamite. Though it had been known since the late 1800s that ammonium nitrate mixed with a fuel would explode in large diameters, the corrosiveness, affinity for water, and tendency of ammonium nitrate to cake, plus its large critical diameter (the minimum diameter at which it will propagate a stable detonation), made it unusable. In the 1950s it was found that commercial-grade ammonium nitrate fertilizer prills (small spherical grains) would detonate in reasonably small diameter holes when they were mixed with finely divided carbon or about 6 percent by weight of fuel oil.

Whereas solid ammonium nitrate crystals are difficult to detonate, prills are cracked and fissured because the last 5 percent of moisture is evaporated after the ammonium nitrate is in prilled form. This leaves much more surface area exposed to chemical reaction, which is the primary reason for the success of ammonium nitrate–fuel oil as a blasting agent; *i.e.,* it results in a smaller critical diameter. Equivalent in strength to about a 40-percent dynamite, ammonium nitrate–fuel oil cannot be detonated by blasting caps but requires a strong booster to initiate an explosion. This feature classifies it as a blasting agent rather than an explosive. Thus, it is safe to handle, can be mixed at or near the blasting site, can be loaded in dry holes in bulk from automatic loaders, and costs much less than dynamite. Disadvantages mentioned above include its affinity for water, corrosive action on metals, and tendency to cake. It cannot be used in wet holes because of its solubility except in waterproof bags or as a slurry.

While the solid density of ammonium nitrate is 1.725 grams per cubic centimetre (0.06 pound per cubic inch), the bulk density of prills varies from 0.85 to 0.95 gram per cubic centimetre (0.03 to 0.04 pound per cubic inch). Detonation velocity increases almost linearly with bulk density, and the question arose during its development whether the increase in density by filling pore spaces between prills with water could be compensated by other thermodynamic factors. It was found that, if ammonium nitrate, trinitrotoluene (TNT), and aluminum were mixed with water, the increase in density caused an increase in detonation pressure, and the favourable heat of reaction of aluminum plus other thermodynamic factors gave detonations of high velocity, higher pressure, but lower temperature. Because of these characteristics, slurries of various compositions have found wide application in rock and ore that is difficult to break. They may also be used in wet blast holes.

Blasting caps

Explosives are usually detonated with electric or nonelectric blasting caps that contain a small amount of powerful, relatively sensitive explosive. Detonating fuse, a train of explosive with a small critical diameter that will propagate a detonation, is also employed. Blasting agents, which are insensitive to blasting-cap initiation, are detonated by means of a booster, a charge of up to 225 grams of high explosive, which is itself initiated by a blasting cap.

An explosive confined in a borehole transmits much of its explosive energy to the surrounding rock in the form of a compressive "stress pulse." When this reflects from a rock surface in tension, it causes the rock to break in

slabs, inasmuch as rock is only about one-tenth as strong in tension as it is in compression. The type, amount, and spacing of explosive charges are determined from experience or by experimentation.

Drilling. Drilling for blasting purposes (Figure 5) is performed with several types of drills. Percussion drills impact the rock with a chisel-shaped or button-faced bit (containing button-shaped inserts of extremely hard tungsten carbide) that chips the rock with repeated blows as it is rotated. This type of drill usually employs compressed air for power. Drag-bit rotary drills cut by shaving, plowing, or abrading with a rotating cutting head using blades, diamonds, or hard steel shot as cutting agents. Drag action is provided by thrust and rotation. Rotary-bit drills utilize a cutting head mounted with roller cones studded with teeth or hard metal buttons; they cut the rock by indenting, shaving, fracturing, and crushing. Rotary–percussion is a combination of the two methods described above.

Factors influencing mechanical drilling are the condition of the equipment, depth, and diameter of the hole, rock properties, geologic conditions, and the stress in the rock.

In open pits, larger holes are drilled for explosives than underground; most large holes in surface mining are made with rotary bits. Some drilling a few decades ago was performed with churn drills, which fragment the rock in the bottom of a hole by repeatedly raising and dropping a heavy drill bit. Cuttings are removed from the hole with a long pipe with a closure device on one end. Until 1949 most rotary drilling employed drag bits. Modern rotary drills attack the rock with roller cones with hard teeth or knobs, the bit assembly being supplied with energy by means of rotating drill rods. Rock is broken by processes of abrasion, chipping, and scraping, or a combination of these. Circulating water is used to clean the hole and carry the cuttings to the surface. The drill is operated by a rig that may be mounted on a truck, trailer, or crawler. Power is furnished by a gasoline or diesel engine or electric motor. Hole diameters for surface mines may vary from 23 to 38 centimetres (nine to 15 inches), with the spacing between holes being five to 12 metres (15 to 40 feet).

Flame-jet drilling

In hard rock that spalls (chips at the surface) readily when heated rapidly, flame-jet or jet-piercing drilling may be employed. The drill consists of a jacketed blowpipe that is water cooled. Fuel oil and oxygen are burned in a combustion chamber generating a high-temperature flame. The heat spalls the rock, and the cuttings are carried from the hole by the steam from the cooling water. Quartzitic rocks containing 20 percent or more of quartz are usually most favourable for spalling. Holes cut by flame jets may be somewhat irregular and require a free-flowing dry explosive or slurries to fill them completely. Flame jets may also be used to cut deep channels in dimension stone rock that spalls readily when attacked by heat.

Most underground drilling and blasting utilizes small holes with diameters of three to eight centimetres (one to three inches), which are cut with either pneumatic or noncoring diamond drills, the latter often being employed for deeper holes. Also, both of these types of equipment can drill in any direction.

Pneumatic drills, powered by compressed air, consist of a heavy cylinder with a reciprocating piston that strikes a hollow drill rod with a bit on the opposite end. The drill rod fits into a chuck that rotates; water is conducted through the machine and drill rod to cool the bit and wash cuttings from the hole. Pneumatic drills may be hand-held (jackhammer) or mounted on a pneumatic leg (stopers) or on a column or jumbo (drifter). Modern jumbos for mounting drills are mobile, and the drilling by two or more machines may be controlled from a console on the vehicle. Pneumatic drills are employed primarily for shorter holes because of the inefficiency of transmitting impact energy through long rods.

For drilling very large holes, such as the shaft in a uranium mine, drills of the type shown in Figure 6 may be employed.

By courtesy of Gardner-Denver Company, Quincy, Ill.

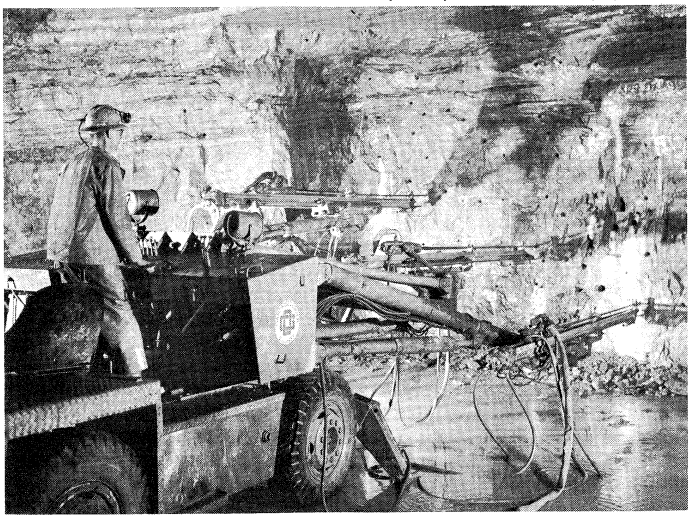


Figure 5: Automated multiple-drill mounting with drills working at a mine face.

From *Engineering and Mining Journal* (June 1970); © McGraw-Hill Inc.

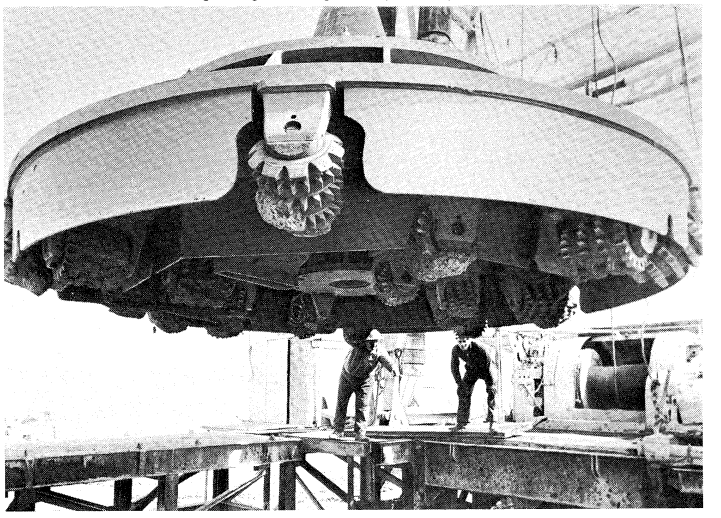


Figure 6: Shaft drilling rig for moderately hard rock. Drilling bit is five metres (16½ feet) in diameter and weighs nearly 40 tons. It is being used here to bore the shaft of a uranium mine.

Shearing and cutting. In soft rock or bedded deposits such as coal, ore often can be broken without drilling or blasting; the bedded deposits can instead be sheared or sliced by giant cutting machines. The mechanization and automation of mining techniques during the 20th century led to the development of cutting and shearing machines that extract ore without drilling or blasting. The longwall mining technique employs a large plow or shear that is dragged across the face of a seam of ore in a parallel manner, slicing off rock in depths of five to 75 centimetres (three to 30 inches). The continuous miners employed in the shortwall and some room-and-pillar techniques act like giant chainsaws, cutting into a seam of ore two to three metres (six to nine feet) in a perpendicular, or near perpendicular, fashion (see below *Coal: Mining: Underground coal mining*).

Hoisting. Hoisting from underground mines was first accomplished with a windlass. A later type of hoist was the horse whim, comprising a rope or cable around a drum turned by a bevel gear attached to a sweep to which a horse was harnessed. Mechanical hoists are powered by internal combustion engines, compressed air motors, or—usually—by electric motors.

Most hoists have a double drum to hoist from a two-compartment shaft. The cables are wound on the two independent but mechanically connected drums in opposite directions to permit hoisting in balance. The cables run up over sheaves (grooved pulley wheels) on the top of a headframe to cages (see below) or skips attached on each. As one cage or skip is hoisted, the other descends, thus utilizing the weight of the empty to assist in hoisting the

loaded one. To keep the peak power demand at a lower level, a control device that utilizes energy stored in a large flywheel is employed on some hoists.

Hoisting drums may be either single or double and cylindrical or conical in shape. A conical shape permits a low starting torque on the motor and thus helps hold down the peak power load. If a single-compartment shaft is used, a counterweight may be employed, similar to that in a building elevator system, which in turn permits a smaller size motor.

Hoisting ropes

Most hoisting ropes are made of multiple-strand, twisted steel wire that must be strong enough to lift loaded skips and cages in addition to sustaining the acceleration and friction forces of the hoisting system. The hoisting drum must be large enough to keep the bending stresses of the cable around the drum below critical values. Hoisting ropes must be inspected frequently and discarded before they are weakened below the allowable strength.

Loading and haulage. Early loading of ore and rock was done by hand into baskets or similar containers that were then carried out of the mine. Wheelbarrows were employed in early mining and are currently used in small openings; animal haulage, which found early use, is still employed in small mining operations. Hand loading, or mucking, is likewise used at the present time, particularly where the cost of mechanical loaders is not justified.

Modern loading in horizontal room-and-pillar mines (in which a large horizontal ore body is mined, leaving some ore behind in the shape of pillars to support the roof), drifts (horizontal openings in or near an ore body and parallel to its long dimension), and tunnels is accomplished with various types of mechanical loaders mounted on either track, tractor tread, or rubber tires. Manual pushing of small cars on narrow-gauge tracks is called hand tramming. Broken ore and rock may also be moved along the floors of stopes or along tunnels by means of a scraper, a large hoe drawn back and forth by means of a cable through an anchored sheave (pulley wheel), the system being powered by a small hoist. Broken material is scraped into chutes or over ramps into mine cars.

Mine cars are heavy sheet-metal boxes on wheels with capacities from about 900 to 9,000 kilograms (2,000 to 20,000 pounds). They are usually drawn by electric locomotives, either battery-powered or powered through a trolley to an overhead wire. All are designed so that the broken rock or ore can be dumped outside of the mine or into a bin cut in rock, often lined with timber (skip pocket) at the hoisting shaft. For large mines on one level, truck haulage may be used. Sufficient ventilation must be provided to dilute or carry out the exhaust gases from the internal-combustion engines.

Ventilation and lighting. Underground mines must be adequately ventilated to provide a breathable atmosphere, to carry out noxious or explosive gases, and to provide cool air in hot mines. One primary purpose of mine ventilation is to make certain that there is enough oxygen in the working places for people to breathe. In gassy bituminous-coal mines, large amounts of air are required to dilute methane gas released from the coal. Noxious gases are generated in some metal mines by the action of acid water on rocks and minerals as well as by explosives used for blasting. Radon gas must be removed from uranium mines because of its radioactivity. Many mines are excavated in rock masses where temperatures are as high as 60° C (140° F) or more, sometimes with high humidity conditions. The air pumped through many such mines is cooled and dehumidified, a costly process.

The main ventilation system of a mine usually requires a minimum of two openings to the surface, with a fan at or near one opening. Air may either be forced through the mine by a positive pressure or pulled through by exerting suction on one opening. The fresh air is guided through the mine by stoppings, doors, and bulkheads. Auxiliary ventilation by small fans and ventilation ducts picks up fresh air and forces it into dead-end drifts, raises, stopes, and other areas that cannot be ventilated by the main air current. If fires are burning in a mine (old timbers, etc., in mined-out areas), these areas must be sealed off from the oxygen supply in the air currents and objectionable gases contained by bulkheads or stoppings.

Auxiliary ventilation

Costs of mine ventilation and air conditioning may constitute a significant part of the cost of producing ore, thus affecting the cost of the final mine product to the consumer, the grade of mineral that can be mined, and the margin of profit to the mining company.

For primary ventilation of a mine, a fan must be chosen that will deliver the necessary amount of air at the proper pressure. The amount of air a fan can deliver varies with the resistance to airflow encountered in the mine. This resistance changes with the extent of excavation, so fans have been designed whose characteristics can be changed by altering the pitch of their blades, or the fans may be designed to operate efficiently over a fairly wide range of pressure and quantity of air flowing.

Natural ventilation will take place in a mine if the two openings are at different elevations at the surface and if the outside temperature is different from the mine temperature, the flow being caused by the difference of weight of cold and warm air (thus producing convection currents). Natural air currents often reverse direction from summer to winter.

Health hazards

Air in workings must be treated to minimize the presence of harmful dust. Fine coal dust may cause "black lung" disease, while fine silica and asbestos may cause silicosis or asbestosis if breathed over a long period. Some dusts may be settled with a water spray, but others are difficult to remove from the atmosphere. Maintaining the quality of mine air is essential to the health of the miners and has been the subject of intensive research and development. Conditions have improved markedly in the past few years. Most countries have comprehensive legislation regarding health and safety in mines; labour unions may refuse to work under dangerous conditions. In view of the vital role that mining plays in the economic and strategic welfare of a country, the importance of supporting mining operations with competent scientific and engineering research is evident.

Most of the shaft stations and main haulage ways in mines are lighted with electric lights. Mine locomotives, trucks, and similar vehicles have lights comparable to those of transports operating on the surface at night. Workmen underground have lights, usually worn on their caps. Early lights were candles, wick lamps, or carbide lights. Most miners currently have electric lights powered from a battery on their belts through an electric cord. Good lighting is essential to safe working conditions as well as to efficient production.

Mining methods. In mining, the intervals between different underground levels and underground opening supports must be carefully selected. Mining methods include open stoping, sublevel stoping, supported stopes, top slicing and sublevel caving, and block caving.

Intervals between levels. In ore deposits of hundreds of feet in vertical dimension, the interval between haulage levels should be the distance that gives the least cost per ton; that is, the value of ore to be mined from a given level must exceed costs of driving the level. The main cost items include sinking shafts and drifts (horizontal openings in or near an ore body, parallel to the course of its long dimension); making crosscuts to the ore; level equipment such as fans, track, air and water lines, and power lines; mining pillars; raises (vertical mine openings driven from one level to the other) and ore passes needed; maintenance of drifts, crosscuts, and raises for the life of the level; hoisting ore to the surface; pumping; handling ore and supplies in stopes; ventilation; waste filling; supervision; and interest on capital represented by broken ore in the stopes.

Underground opening support. In underground openings where the rock in the roof and walls is too weak to stand alone, it must be supported. The simplest type of timber support is a one-piece set called a stull or post (Figure 7A and B). Two-, three-, and four-piece sets (Figure 7C and F) are employed in small drifts and tunnels. In stopes that require close support, square sets (Figure 7D) are employed, either in narrow veins, wide veins, or larger, irregular deposits. Timber support may also be required

where horizontal openings meet the shaft at a shaft station that must usually provide for two hoisting compartments, a manway, and compressed air and water lines in the shaft, in addition to chutes to load ore and support for headroom for unloading materials. Many modern shafts are circular; support may be provided with steel or timber. Concrete may be used in shafts and permanent openings as well as steel sets. In very heavy grounds, collapsible steel arches or sets have been employed.

Rock bolting

Rock bolting has found broad application, particularly in openings that must be supported for long periods. A rock bolt is a steel rod, threaded on one end and provided with a means for anchoring in a hole in the rock on the other end. The bolt is anchored in the hole and a plate or other retainer pulled against the rock with a nut on the threaded end. Rock bolts are employed in drifts and other types of openings to hold the rock structure together, to reinforce strata, and to hold key pieces of rock in place and thus stabilize the structure.

A fine aggregate mixture of portland cement, sand, and water called shotcrete may be used to stabilize the rock around an opening when it is sprayed on in sufficient thickness, and it hardens in place. The shotcrete fills in small voids, holds key fragments in place, and furnishes arch support for the opening. It is used mostly in tunnels.

When a drift is driven in running (loose) ground, a method of timbering called spiling, or forepoling (Figure 7E), is used to provide close support to the ground.

Open stopes. Pillars of ore or waste rock are left to support the back of stopes in deposits of low-grade ore, in which it is cheaper to leave such pillars than to use artificial support. The required support depends on the depth of the deposit or the total weight on the pillars; the nature of the roof, which controls the length of unsupported span; the nature of the floor, which controls the bearing area and consequently the size of the pillar; and the strength of the ore, which determines the minimum size of pillars.

This method is applicable to geologic conditions that require a minimum of artificial support, the walls and roof span or both being self-supporting. Small ore bodies may be mined from wall to wall without pillars. Where ore bodies are larger, pillars of ore are left to keep the roof span to a safe dimension. Pillars may be spaced regularly, as in deposits of salt or phosphate, or randomly spaced in irregularly occurring ore (Figure 8), and the method may be applied to either horizontal or inclined deposits. Its greatest application is in flat-lying, bedded deposits.

From R.S. Lewis, *Elements of Mining*, 3rd ed. (© 1964); used with permission of John Wiley & Sons, Inc.

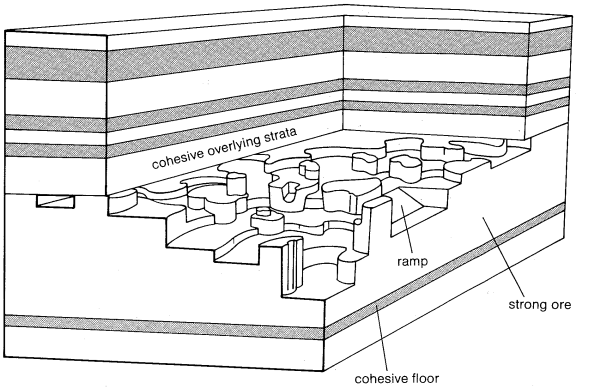


Figure 8: Room-and-pillar method of mining with random pillars.

The pillars are usually roughly circular, square, or rectangular in section and flared at the top and bottom to increase contact area with the roof and floor. In weak ground, the spacing is less, and the pillars are larger than in strong ground. The ore left in the pillars is sometimes recovered later. Some extraction of high-grade pillars has been accomplished with the roof supported by concrete columns, which must be preloaded by forcing them against the roof before it collapses.

In the development and mining of a flat-lying deposit, one or more openings are driven in a selected direction, and rooms are excavated usually at 90° to the openings. The pattern of layout of rooms may vary considerably, however. Pillar heights vary from a few to over 30 metres (100 feet), depending upon the thickness of the deposit.

The geology of typical open-stope mines is notable for simplicity of structure, lack of faulting, lack of alteration, and competency of the beds and the strength of the overall structure. Although dimensions of the deposits are variable, the method has uniform application throughout the deposits as long as the ore and rock are strong enough to support large openings. The placing or size of pillars is governed almost entirely by the presence of faults, fractures, or weak beds. Where weaknesses do not occur, the pillars may be placed regularly. Natural pillars may also be left to support the walls of a deposit that dips steeply. In this case, levels are driven from a shaft, and the ore is mined between the levels with natural inclined pillars left for support. Pillars may later be taken out ("robbed") and the walls allowed to collapse.

Robbing pillars

Sublevel stoping has many variations but is best suited to deposits that are steeply inclined and have strong ore and stope walls. In this method, there are two horizontal drifts (the main levels) 30 to 60 metres (100 to 200 feet) apart vertically. These drifts are connected by raises (vertical openings) that are separated about the same distance. At intervals of six to eight metres (20 to 25 feet), sublevel drifts, between the main horizontal drifts, are driven from the raise to the end of the stope. The main volume of ore is mined from the sublevel drifts. The ore and stope wall must be strong after the ore is excavated to furnish footing for workers and support for sublevel drifts. After the main blocks of ore have been mined, the pillars are usually removed, after which the walls may collapse.

Supported stopes. In shrinkage stoping the back (roof) of the stope, which is usually ore, is unsupported during mining, but the walls are supported by broken ore. De-

From R.S. Lewis, *Elements of Mining*, 3rd ed. (© 1964); used with permission of John Wiley & Sons, Inc.

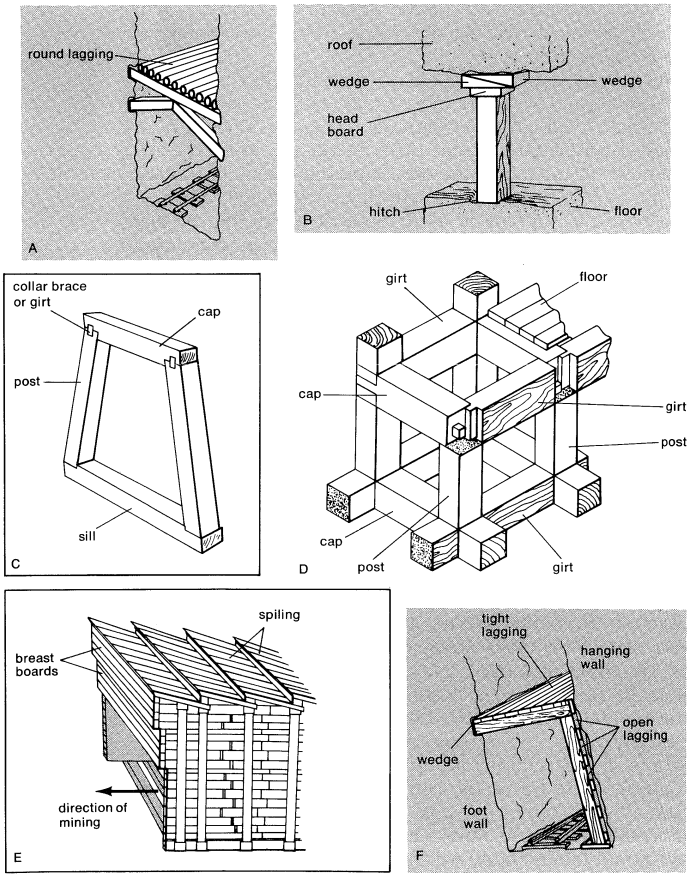


Figure 7: *Types of mining timbering.*
(A) Reinforced stull. (B) Post. (C) Four-piece set, sawed timber. (D) Square set. (E) Timbering in running or loose ground. (F) Two-piece stull set.

posits may or may not need support during mining. The ore should be strong, but the walls may be weaker than those mined by sublevel stoping.

The first stages of shrinkage stoping are similar to those of sublevel stoping. There are no sublevels, however, and horizontal sections of the ore are mined from the bottom upward. The broken ore is left in the stope to support the walls during mining, and to serve as a floor for miners removing the ore overhead. Since the ore takes up more space when broken, about a third is drawn off as it is mined. After a stope is completed, the remaining ore is taken out.

Since the ore is unsupported during mining, it must be competent (*i.e.,* it must not yield or fail under stress). If walls are too weak, rock will slough off, diluting the ore. Hence, in deposits minable by shrinkage methods, the overall structure is slightly weaker than the ore and rock in which sublevel stoping is used. The footwall (the wall of rock under a vein of ore) of the stope should be steep enough to permit drawing the ore by gravity.

Shrinkage stoping has the advantage that blocks of ore may be quickly developed and mined, but considerable broken ore must be left in the stopes during mining, sorting may be difficult, and some sulfide ores may catch fire before they are removed.

Cut-and-fill method

In the cut-and-fill method (Figure 9), the ore is removed in parallel or stepped slices, and, as each slice is removed, a layer of waste (broken rock) fill is placed in the stope, leaving working headroom above it. Its applicability is similar to that of shrinkage stoping. The ore must be strong so that the miners can work under it. The walls do not need to be as strong, since the fill provides support. The fill may be in the form of broken rock from the mine. In some mines a sand and water mixture is pumped through pipelines into the stopes and the water is drained off. Because it settles less, sand fill is usually more satisfactory than broken rock. Sometimes the ore is of such a type that dilution cannot be allowed, and, if the ore oxidizes readily, causing difficulty in subsequent processes, shrinkage stoping cannot be used.

From R.S. Lewis, *Elements of Mining*, 3rd ed. (© 1964); used with permission of John Wiley & Sons, Inc.

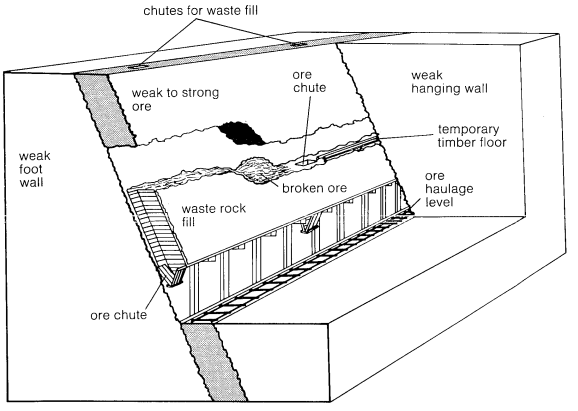


Figure 9: Cut-and-fill stoping, which is adapted to moderately weak ore and wall structures.

Shrinkage stoping is also used widely in steeply inclined deposits where broken ore can be removed by gravity and the stopes filled from the level above. In such deposits, even minimum subsidence (gradual settling) should not occur adjacent to the ore body; thus side support is given by immediate filling. Occasional slabs or loose sections of the back are supported by timbers placed on the fill. In large ore bodies, cut and fill stopes are kept small by spaced vertical pillars that are mined later, commonly by square-set stoping.

Before blasting, a horizontal or inclined layer of plank (lagging) is often laid over the waste to keep it from becoming mixed with the ore. The ore is drawn off at the level below and the waste is brought down from the level above.

Cut-and-fill stoping permits sorting and mining of irregular deposits and keeps ore dilution to a minimum. Ore production is intermittent, however, since filling is required, and the method is costly.

Square-set stoping, which employs timber frameworks (square sets) for support underground, is used most with ores that are weak, and the surrounding rock is also weak. The method is used when caving methods cannot be used, either because of the deposit or because of its high value or both. Sets can be extended vertically or horizontally depending upon the shape of the ore body, thus lending flexibility to operations. Basically the sets provide temporary support and a means of reaching the working face. Permanent support is accomplished by means of fill placed in the sets after a tier is worked out. Besides being used in mining heavy ground, square sets are commonly employed in mining pillars between cut-and-fill stopes. Sets are standardized as far as possible to simplify timber framing.

Among the advantages of square-set stoping are that irregular ore bodies may be worked while masses of waste or lean ore can be left in place; waste may be sorted out and left in the stopes; the grade of ore can be controlled, since each new face can be sampled and assayed before the ore is drilled (this flexibility is important in mines where the ore varies in value); and filling the sets as soon as possible leaves only a small space open at a time. Disadvantages include high cost; slow ore extraction and low productivity per man; large timber requirements and consequent fire hazard; and the highest accident rate of any underground-stoping method.

Disadvantages of square-set stoping

Under many geological conditions in which square-set mining is employed, the ore and rock structure approach the behaviour of unconsolidated materials. The instability is caused by faulting, fracturing, and folding and the destructive alteration that accompanies mineralizing processes.

Top slicing and sublevel caving. In top slicing, horizontal slices are made, work proceeding downward from the top of the deposit. The capping or overburden, immediately above the top of the deposit, caves after each slice is removed. During ore excavation, the capping is supported by timber.

Top slicing is most adapted to deposits that are large, horizontal, and weak. It is essential that the capping be such that it will cave. Development openings consist of a series of drifts and crosscut drifts below the mining level and raises to the top of the ore. Sublevel caving is similar, but in sublevel caving, alternate slices are mined, and the weight of the capping is used to assist in mining of the ore. The capping should be somewhat stronger than that for top slicing.

For both methods, the rock capping must be weak enough to cave when undermined. The caving formations should not be such that there could be sudden, dangerous collapses of large masses. Sublevel caving is more adaptable to irregular deposits than in top slicing.

Since considerable timber is used for supports in top slicing, an adequate supply of tough, low-cost timber should be available.

Development of top slicing

Top slicing was developed in the United States from caving methods used in iron mines of northern England. In the deposits, soft hematite was covered with glacial drift (loose soil and rock deposited when glaciers retreated). Initially, a longitudinal drift was driven through the ore body and raises containing a chute and a manway were placed at 30-metre (100-foot) intervals. From the top of each raise, a drift was driven a distance of 15 metres (50 feet) in either direction. The drifts were timbered. Slicing took the form of a succession of crosscuts and caves, a process that was repeated until an entire horizontal slice was removed.

Top slicing is safe for mining heavy ground and is economical where timber is plentiful and inexpensive. High extraction of ore is possible with little dilution. If the mine shuts down for any reason, stopes may be blasted down, and the ore body will remain in good condition for resumption of operations. Top slicing can be employed under loose material and does not require as clean a mat (mixture of broken rock and timber) as sublevel caving. On the other hand, top slicing cannot usually be em-

ployed without subsidence. Ventilation of such a mine is difficult, and large amounts of old timber create a fire hazard. Furthermore, the period of development prior to production is relatively long, and the rate of production is inflexible. Timbering and covering down (laying a rough timber floor) consume a large part of the work shift; hence, rock drills and scrapers are idle much of the time. Finally, sudden collapse of large blocks presents an element of danger; sorting of waste cannot be done in the stopes, nor can small blocks of waste be left in place.

Sublevel caving may have been developed from top slicing. In this method every other slice is mined by driving crosscuts (slice drifts). The ore between the crosscuts as well as that in the slice above is then mined and the overlying material caved. The method is applicable when the ore is weak and moderately soft and the capping material is hard enough to cave with a minimum of fine material. The method is used in irregular and inclined ore bodies that would not be suitable for working by top slicing. Because the ore must be removed before the capping caves, the capping must cave more slowly than is the case with top slicing.

The interval between levels ranges from 45 to 60 metres (150 to 200 feet). Sublevels are also driven from the tops of raises (vertical mine passages between levels) from the main and transverse haulage drifts. Raises are commonly spaced about 15 metres (50 feet) apart but may be farther apart in narrow ore bodies. Broken ore is pulled to the raises by scrapers. When all the ore is removed from a set and the capping is caved, the next set is caved, and so on, back to the raise at the sublevel.

The advantages of sublevel caving are that cost of mining heavy ground is comparatively low; cost of timber is lower and less development is required than for top slicing; ore may be mined more rapidly; the percentage extraction of ore may be high with little dilution; and the method can be applied to wet and sticky ore not suitable for block caving.

Disadvantages of this method include: there may be more dilution of the ore than in top slicing and square-set stoping; there is little sorting of ore in the stopes; stopes are difficult to ventilate; there is danger of fire from accumulation of timber; and low-grade ore in the capping or near the boundary of the deposits cannot be mined.

Block caving. The block caving method evolved from the methods used to mine pillars in Lake Superior iron mines. It is particularly applicable to very large ore bodies that have cappings that are sufficiently weak to be caved. Preliminary work consists of a series of evenly spaced crosscuts below the level of the ore body, from which a series of raises, known as main, branch, and finger raises, are extended up into the ore. The mass of the ore is then weakened by undercutting, and side cutting is sometimes necessary. The ore is further weakened, usually by blasting, just above the point where the ore is drawn off. The result is that the weight of the ore in addition to the weight of the capping crushes the ore and forces it down through the raises to the haulage way. The movement of the upper mass often affects the strength and stability of the lower mass.

Porphyry-copper deposits where both the ore and capping are weak lend themselves particularly to the block caving method. Large deposits of relatively strong ore are also mined by this method, using appropriate procedures.

Advantages of block caving

The cost of mining by block caving is low, approaching that for open-pit mining, and a high rate of production is attained once production is started. Mining conditions can be standardized, making for safe, efficient operation, and the accident rate is fairly low. On the other hand, capital expense is large, with a long period of time before a mine produces ore. Extracted ore is diluted with waste, and there is some loss of ore; drawing of the ore must be carefully controlled; and, finally, low-grade ore in the capping and at the margin of the ore body is lost.

SURFACE MINING

Basic open-pit mining techniques. Surface mining, or open-pit mining, is utilized to produce coal, copper, iron ores and nonmetallic mineral products such as clay, gypsum, phosphate rock, several types of stone, sand, and gravel. In the United States, more than three-quarters of copper ore and iron ore produced has been mined by open pit, whereas these values are about one-half for other countries.

Just as in underground mining, mineral deposits mined by surface methods are found by prospecting, evaluated by exploration, then developed, and finally placed in production.

For coal and metallic ores, the ratio of overburden to ore that can be moved is greater than for nonmetallics. Some of the factors that are important in excavation are drilling and blasting, cost of explosives, improvement of efficiency and size of excavation equipment, and methods of haulage.

Prospecting for deposits that may be mined by open-pit methods is conducted in much the same manner as for deeper deposits mined by underground methods. Surface geologic structures are mapped and studied. Drilling, geophysical methods, or both may be employed to determine details of subsurface geology, to ascertain the presence of valuable minerals, and to make preliminary estimates as a basis for the exploration phase that follows.

When a new mineral deposit is found, it must be explored so that its size, value, shape, overburden, and other related factors can be determined. This information is utilized to ascertain whether it is economical to mine the deposit, to plan the method of operation, to predict profits, and to determine economical limits of mining and the other essential elements. Methods of drilling and sampling are adjusted to the type of deposit, and the results of sampling are evaluated by statistical methods to yield reliable evaluation. It is estimated that only one in 100 mining ventures develops into a successful operation. The exploration of a deposit, if properly done, will yield most of the information necessary to determine whether it may be mined at a profit.

Drilling programs

Drilling programs must be carefully planned; the type of drill chosen depends upon the material to be drilled. In reasonably intact rock and ore, diamond drilling yields more detailed geologic information as well as mineral value. Unconsolidated materials must be drilled by other methods. The first step of a drilling program is further verification of prospecting information; the second is outline drilling, a method of determining the approximate dimensions of the deposit; and the third step is sampling drilling, to yield enough detailed information about the deposit for economic appraisal. Drillings from all three phases are sampled and tested for valuable contents. Drill holes must be properly spaced and values of samples correctly utilized to evaluate the deposit. Total ore reserves determine the life of the mine.

Factors that must also be considered in the overall operating feasibility of a prospect are geography, land and mineral rights, sociological and political factors, geology, mining conditions, ore-treatment requirements, and economic analysis.

Ownership of mineral land must be proven and stable before a deposit can be explored. Political and sociological factors must be conducive to profitable mining. Geological factors include the occurrence, size, mineral content, depth, and value of a deposit and must be evaluated sufficiently to justify expenditure of capital for development. Mining conditions refer to labour, working conditions, elevation, and related factors. Ore-treatment requirements are critical and include answers to questions of whether the ore must be concentrated before it is smelted or, if nonmetallic, whether treatment will be required before a marketable product can be produced. Finally, an economic evaluation of all of the above must be made, including investment costs, again taking into account that a mineral deposit is a wasting asset.

After a deposit has been explored and its grade, total tonnage, depth of overburden, size, shape, and other factors determined, pit layout must be planned and correlated with all mining operations. In addition to the factors described above, types of ore, cutoff (the lowest grade of ore that is profitable to mine), rate of production, pit slope, bench height (a bench is a level surface worked separately in an excavation), road grades, and related items must

be considered. Careful sampling is required to fix the boundaries between capping and ore and to define waste inclusions. The ratio between worthless overburden and valuable ore must be calculated in order to ensure that a profit can be made. Flatbedded deposits are much easier to evaluate than irregular massive deposits. Pit slopes must be gentle so that rock and ore will not slide, and material near geologic fault planes sloping with the pit must be carefully excavated.

The industrial use of mined material determines some mining procedures. Concentration of low-grade metallic ores requires that dilution be minimized. Nonmetallics that require purification are in the same category. Samples of the mined material must be representative for proper design of treatment processes. Some mining companies employ computers to calculate pit dimensions and other layout factors.

For stripping overburden and mining, ore equipment is purchased to suit operating conditions. Types of excavators include power shovels, draglines, scrapers, and bucket-wheel excavators. Haulage equipment includes bulldozers, scrapers, trucks, trains, and conveyors of various types. The combination of shovels and trucks is commonly selected.

The total production system usually includes drilling, blasting, loading, and haulage. Modern procedures include analyses of such cycles by advanced statistical techniques that require the use of computers. Successful open-pit mining thus requires modern management techniques and application of good industrial-engineering principles, particularly in large operations. Small or relatively simple operations do not require such complex analysis.

Production drilling is required to make holes for explosive in rock that is too hard to excavate without blasting. The two common methods of rock drilling are mechanical and thermal. Other novel methods, such as sonic, chemical, laser, electron beam, and water jets, are currently in the conceptual or research stage.

In surface mines, blasting rounds (groups of holes) are spaced according to the size and shape of the benches, rock resistance, joint and fracture systems, and similar factors. Blasting is conducted to give the best breakage with a minimum displacement of the broken material. Large boulders require costly secondary breakage, and a pile of broken rock close to the new bench face is less expensive to load. If blasting is near populated areas, ground vibrations and air blast must be minimized.

Blasting in an open-pit mine

Blasting with nuclear explosives is being considered for several purposes but mostly for large-scale underground mining or for excavation of canals and harbours. Proper burial of the explosive device minimizes atmospheric radiation and the unit explosives costs decrease with increase of the size of the charge for nuclear explosives.

Pit operation is performed in cycles of drilling, loading explosives, blasting, and loading and haulage. In addition to power shovels, draglines and clamshells are employed in many countries, as are continuous excavators, consisting of a bucket wheel on the end of a long boom, which excavates loose material and dumps it in a conveyor system, which in turn conveys ore to haulage units and waste to a spoil bank.

Large mines are organized and equipped for maintenance of all types of mining equipment, while, in small mining operations, maintenance may be performed by operating crews.

Removal of water and flood control are important in many mines, depending upon rainfall and the presence of water-bearing rock. Slope stability in large mines may present problems. Research in rock mechanics and geology is often essential in preventing slides and predicting rock mass movement. In some mines, seismic geophones have been installed in the rock at critical points to detect minute failures as evidenced by microseisms (tiny earth tremors). Their intensity and magnitude are indications of rock failure and possible slides.

Some valuable minerals, such as gold and cassiterite (tin mineral), are found in gravel deposits, where they have been deposited after being eroded from their place in primary deposits. Deposits of this type (placer deposits) are often exploited by dredges or hydraulic mining, as is overburden over iron-ore deposits in some countries.

Dredging

A dredge is a floating barge with a boom at one or both ends and engines and a concentrating plant if the latter is required in the barge or hull proper. A bucket-line dredge has one boom with a bucket line that may be lowered into the water in which the dredge floats. The dredge is pivoted about the far end on a spud (a heavy, pointed column driven into the floor of the water body), the bucket line continuously excavating material as it is swung by means of tension lines. The excavated material is conveyed through the concentrating plant if it contains valuable mineral; waste is conveyed up the rear boom to the top of the spoil bank. Dredges for moving waste material have no concentration plants on board. They are operated either in a pond in a gravel deposit (California, Malaysia) or offshore (Indonesia and the Philippines). Mined materials are screened and the valuable minerals recovered by washing the finer gravel over riffles, grooved sluices, where the heavy minerals settle out by gravity.

Hydraulic mining consists of washing the material from its position by a large stream of water under high pressure from a large nozzle (known as a hydraulic giant). It is used for mining terrace placer deposits that cannot be mined by dredging or other means. A large supply of water is necessary. Finer materials are transported to a sluice where valuable (heavy) minerals are recovered in riffles. Debris from the operation is accumulated downstream, preferably in ponds behind dams that will not be washed away by floods.

Recently, in the alluvial tin mines in Malaysia, mining practice has evolved almost entirely to an adoption of a system of breaking ground with water jets and removing the loosened material with gravel (suction) pumps.

Offshore operations include mining of gold in Alaska, tin-bearing sands off the coast of Thailand, tin off the shores of the Indonesian island of Sumatra, iron sands off the southwestern tip of Kyushu Island (Japan), cement sands (shell) west of Akranes (Iceland), and gravel from the English Channel.

Quarrying. One meaning of the word quarry is "squared stone," hence, quarrying was initially applied to the mining of dimension (cut) stone. More recently, quarrying has been used to refer to the mining of several types of nonmetallic rock from shallow open-pit mines.

All three types of rock are quarried—igneous, metamorphic, and sedimentary. The most common igneous rock is granite; several sedimentary rocks are quarried. Marble, quartzite, and slate are common metamorphic types that are quarried. The two products of quarrying are dimension stone and crushed stone. The former is employed for buildings, monuments, and decorative purposes. Such rock usually must be of uniform grain size and have desirable coloration, as well as lasting qualities related to its intended use. Sandstone is used for building, granite and marble for memorials, slate for roofing, and quartzite for flagging.

When usable rock is found, the first part of mining consists of stripping the overburden and weathered rock. Large blocks are then separated from the parent mass, sometimes with the aid of low-powered explosives. High-energy explosives would cause the rock to shatter and cannot be used.

Quarrying dimension stone

For most dimension-stone quarrying, the success of the art depends largely upon taking advantage of joints and cleavage planes in the rock. Granite has a "rift" along which it may be split with comparative ease. The planes perpendicular to the rift are the "head" and the "grain," along which it splits less easily. The rift of sedimentary rock is the bedding plane (*i.e.,* parallel to its layering), while some rocks, such as diabase, diorite, and porphyry (trap rock), have no rift. Joints in granite may be at obtuse angles, but, where they are perpendicular, the excavation is termed a block quarry.

A usual method of splitting granite by hand is to drill a row of holes about two centimetres (0.75 inch) in diameter, eight to 13 centimetres (three to five inches) deep and 15 to 20 centimetres (six to eight inches) apart. In each of the holes are placed two "feathers" and a plug. The

feathers are half-round shims or wedges between which is placed a wedge plug. The plugs are successively driven with light blows of a hammer, creating a line of tensile stress between the holes that causes the rock to split. Granite and some limestone split with shallow holes, while sandstone and marble require deeper holes. Holes may be drilled with hand tools or with pneumatic drills. In softer rocks the holes may be drilled farther apart.

Special quarrying methods

Special methods of quarrying include machine channelling, broaching, the Knox system, and flame-jet channelling. Machine channel bits are made of hard steel mounted to cut a channel in softer rock (that is, other than granite or hard metamorphic rocks). Broaching consists of drilling a row of holes and cutting the rock out between them. In the Knox system deeper holes are drilled, grooved on opposite sides, and then loaded with black powder, which is fired to break the rock. For some types of rock containing quartz, flame-jet channelling has been employed. The rock must have the proper spallation (thermal breakage) characteristics for this method to be applied. Wire saws are also utilized; these employ a wire that runs as a belt over the rock, cutting by abrasion as sand or hard grains and water are fed onto it. Blocks are cut to the size desired, hoisted by derricks or other devices, and transported away from the quarry for further cutting or sale.

Crushed stone is mined by the methods used in similarly shaped deposits of metallic ores; that is, the overburden is removed, and then the rock is drilled and blasted. The broken material is loaded and hauled to a crusher, where it is crushed, screened, and sized. It is used for such purposes as road building, railroad ballast, and concrete aggregate. Some limestones have special uses, such as chemical applications, making portland cement, and for flux in blast furnaces.

THE MINING INDUSTRY

Organization and management. The organizational structure of mining companies varies, but similar types are found in coal mining, open-pit mining, and underground mining. A large company usually has a corporate structure, with departments or divisions for operations, engineering, sales, exploration, land, purchasing, and financial and legal aspects. All of these are administered by a centralized executive office headed by the president of the board and the president of the company.

A typical organization at a mine location consists of a superintendent, an assistant superintendent, master mechanics, a chief electrician, pit or mine foremen, shift bosses, blasting foremen, engineers, clerical staff, and other assistants.

Strong support of research in various phases of mining has proved to be important to progressive management and successful operation of most mines. Most large mining companies throughout the world have research departments whose responsibility is to perform investigations that will result in the solution of problems in health and safety, rock mechanics, operations efficiency, and other related matters.

Conservation, ecology, and mining. The conservation of mineral resources, reduction and prevention of environmental pollution, and the preservation of wilderness areas are interrelated in varying degrees. Heavily populated countries are particularly concerned about the problems involved.

Inasmuch as mineral deposits are wasting assets that cannot be replaced, it behooves every nation to conserve them and use them wisely. An available quantity of minerals and mineral products is a vital factor in maintaining high national economic standards, in political power among the nations, and in the ability of a nation to develop its self-defense. Iron, coal, steel, copper, aluminum, and steel-alloying metals are basic necessities, with others of slightly less importance. During wartime many of the ores yielding these metals are classified as strategic minerals.

Effects of conservation regulations

In the United States many mining and smelting operations have contributed to pollution of streams and the atmosphere. Legislation, both national and state, has been passed to improve these conditions, and many mining companies have taken steps to correct pollution problems. The Bureau of Mines is supporting research and development efforts to assist the mineral industries in the reduction of pollution due to mining and processing of minerals.

Because of the nature of financial risks involved in the exploration and development of mineral deposits, many mining companies have found that regulations increase mineral-production costs; for example, regulations may prohibit temporary or permanent roads, aircraft landing strips or heliports, or the use of motorboats, motor vehicles, or other motorized equipment. An aircraft may not be allowed to drop equipment, supplies, or personnel. A permit is required for all types of prospecting except walking over the land. The validity of an unpatented claim may be contested, even though the valuable mineral may be far below the surface. Permits may be refused on the ground that the use of equipment is not consistent with desires to preserve various wilderness areas. Hearings may be called by interested groups to prevent the issuance of a permit, and repeated reconsideration of decisions may hold up prospecting indefinitely. (G.B.Cl.)

Metallurgy

Metallurgy is the science of extracting metals from their ores, refining, purifying, and working them, mechanically or otherwise, to adapt them to use. It is also concerned with the chemical and physical properties of metals, their atomic and crystalline structure, the principles of combining them to form alloys, the means for improving or enhancing their properties for particular applications, and the relations between properties, structures, and uses. Further, it includes the thermal and mechanical processing of metals as materials of manufacture.

Metallurgy as a science is relatively new, metals having been subjected to scientific study for less than a century. Since the 19th century, more and more attention has been devoted to a scientific understanding of the properties and structure of metals.

This section deals primarily with technology. It covers the following major areas: history of metallurgy; mineral dressing; process metallurgy; and metal processing.

DEVELOPMENT OF METALLURGY

The present-day use of metals is the culmination of a long path of development extending over approximately 6,000 years. It is generally agreed that the first metals known to man were gold, silver, and copper, which occurred in the native or metallic state, of which the earliest were in all probability nuggets of gold found in the sands and gravels of riverbeds. Such native metals became known and were appreciated for their ornamental and utilitarian values during the latter part of the Stone Age.

An essential step toward the Metal Age was the discovery that metals could be melted and cast to shape in molds; another was the discovery that metals could be recovered from metal-bearing minerals. There was a close relationship between these metallurgical developments and the growth of civilization. Man needed a stronger metal than native copper. Possibly more by accident than intent, he discovered that certain copper ores containing the tin-bearing mineral cassiterite could be treated in a charcoal fire at a red heat to produce an alloy of copper and tin known as bronze. This discovery occurred between 4000 and 1400 BC. Bronze is harder than copper, has a lower melting temperature, and casts more readily, producing sounder castings. There is definite evidence that the advantages of these bronzes were soon appreciated. The earliest copper ores used in smelting (extraction of a metal from its ores by heat) were surface outcrops that had been weathered until many of their impurities were removed.

Discovery of bronze

Emergence of bronze, iron, and brass. As the depth of ore bodies below the surface increased, the ore not only tended to become harder, it tended to change radically in composition. A fundamental step was taken when an early genius realized that copper also occurred in unweathered ores; processes that involved quite advanced metallurgical knowledge were gradually devised for working these ores.

As they were less pure than the weathered surface material, the copper produced was correspondingly impure. By the latter half of the 3rd millennium BC, copper was being produced with its impurity content rising to 2 to 3 percent. The alloys containing several percent of arsenic, antimony, iron, or lead, forerunners of the true Bronze Age, at least in central Europe, possessed marked advantages when compared with copper itself; but they tended to be brittle, and the superiority of tin as an alloying material was soon appreciated. Starting with a tin content of 5 percent or less, these bronzes quickly settled down to tin contents of about 10 to 12 percent. The Bronze Age was heralded by these discoveries.

Although bronze was known in Sumer as early as 3000 BC, tin ores were scarce and not as widely distributed as copper ores; hence sources of tin ores were sought outside the Middle East. At about 2000 BC the people of Hissarlik, the site of the city of Troy, learned the art of bronze making using ores obtained from central Europe. Migration of peoples from the Middle East to Egypt, Europe, India, and, possibly, China, brought on by wars and drought, was largely responsible for the spread of the bronze culture into these areas.

Lost-wax process

During the Bronze Age the Egyptians introduced the lost-wax process, a method known today as the investment method of casting. In this method an exact model or pattern of the article to be cast is made in wax, then covered with a claylike material to form the mold. The composite is heated to harden the mold and to melt the wax, which drains off, leaving an exact negative impression, which is then filled with molten metal and allowed to solidify. The statue of the pharaoh Pepi I and his son, dating from about 2600 BC, is one of the earliest examples known of the lost-wax art.

By 1400 BC iron was assuming increasing importance, and it was smelted and worked simultaneously with copper and bronze, as indicated in Table 1.

Iron does not occur on Earth except in ores, but meteoric iron, while rare, has been known from a very early date. A characteristic of meteoric iron is a high nickel content, and iron of this type has been found in ancient iron articles. By about 1200 BC, iron had attained sufficient importance for this date to be used as marking the beginning of the Iron Age.

Smelting of iron from ore involves a much more elaborate technique than does smelting of copper, and iron smelting was correspondingly slow in coming. The earliest record of iron smelting dates from about 1300 BC. Evidence from excavations indicates that the art of iron making originated in the mountainous country to the south of the Black Sea, an area dominated by the Hittites. Later the art apparently spread to the Palestinians, for crude furnaces dating from 1200 BC have been unearthed at Gerar, together with a number of iron objects.

Smelting of iron oxide with charcoal demanded a high temperature, and the product even then was merely a spongy mass of pasty globules of metal intermingled with a semiliquid slag. Repeated reheating and hammering

Production of wrought iron

eliminated much of the slag, creating wrought iron, a much better product than the iron sponge. This wrought iron was low in carbon content, hence soft, and its use for weapons and tools was limited. It was discovered that continued heating to a red heat in a mass of glowing charcoal (during which carbon was absorbed) made the soft metal much harder. As early as 1200 BC iron with 1.5 percent carbon at the surface, corresponding somewhat to modern case-hardened iron, was being processed. The carbon-containing iron had the further great advantage that, unlike carbon-free iron, it could be made still harder by quenching—*i.e.*, rapid cooling by immersion in water.

By 1000 BC iron was beginning to be known in central Europe. Its use spread slowly westward; iron making was fairly widespread in Great Britain at the time of the Roman invasion in 55 BC. In Asia iron was also known in ancient times, in China by about 700 BC.

Brass, an alloy of copper and zinc, made its appearance during the period from 1600 BC to 600 BC. In the beginning the alloy probably was obtained when mixtures of copper and zinc ores were unintentionally smelted. In any event, its value was not appreciated, and the metal was employed sparingly until it was used by the Roman Empire in the form of currency. The general establishment of a brass industry was one of the important metallurgical contributions made by the Romans. Bronze, iron, and brass were, then, the metallic materials on which successive peoples built their civilizations and of which they made their implements both for war and peace.

Table 1: Prehistoric and Classical Metallurgy

period	smelting and refining	alloying and working
Prior to 4000 BC	native gold and copper and meteoritic iron used occasionally without melting	hammering to shape and to harden copper annealed about 4000 BC
4000–3000 BC	reduction of oxidized ores of copper and lead native silver used copper–arsenic alloys by smelting natural mixed ores, followed by bronzes from intentionally mixed ores of copper and tin (*c.* 3500 BC)	melting and casting nearly pure copper, followed by copper–arsenic and copper–tin alloys (made from ores) permanent molds of stone and metal lost-wax process for complicated castings natural gold–silver alloys soldering with copper–gold and lead–tin alloys
3000–2000 BC	roasting and smelting of sulfide ores of copper experimental production of sponge iron tin smelted (perhaps earlier) and becomes important trade item cupellation of lead to extract silver, and of gold and silver for purification	most jewelry techniques known before 2500 BC—inlay, stamping, repoussé, raising, soldering, riveting, granulation, surface colouring, etc. gold leaf wire made by cutting sheet
2000–1000 BC	gold purified by cementation bellows used in furnaces by 1800 BC wrought iron becomes important by 1600 BC (directly reduced from ore as sponge and forged without melting)	bronze made from metallic tin and copper (perhaps much earlier), high-tin bronze (Speculum metal) for mirrors, leaded bronzes for statuary niello and enameling on gold silver–copper alloys blanched brass from copper and zinc oxide ores (not important before 200 BC) steel regularly made by carburization in hearth, perhaps hardened by quenching, 1200 BC
1000–0 BC	cast iron (China only) cast steel ("Wootz": India only) gold purified by sulfide process mercury distilled from ores amalgamation of gold ores	vast expansion of production of iron, common nonferrous metals, and gold and silver iron and steel welded into composite tools and weapons stamping of coins, 700 BC gilding of bronze and silver with gold amalgam assay by cupellation, cementation, touchstone and (?) by density

Source: Cyril Stanley Smith, The University of Chicago.

Developments from 500 BC to AD 1500. By 500 BC considerable technical progress had been made in the mining and recovery of ores, progress that was definitely linked with the beginning of scientific thought in metallurgy. By this time rich lead-bearing silver mines had opened in Greece and had reached depths of several hundred feet. These mines were vented by drafts provided by fires lit at the bottom of the shafts. Ores were hand-sorted, crushed, and washed with streams of water to separate valuable minerals from the barren, lighter materials. Because these minerals were principally sulfides, they were roasted (heated in air without fusion) to form oxides that were then smelted to recover a lead-silver alloy.

Lead was removed from the silver by cupellation, a process of great antiquity in which the alloy was melted in a shallow porous clay or bone-ash receptacle called a cupel. A stream of air over the molten mass preferentially oxidized the lead. Its oxide was removed partially by skimming the molten surface; the remainder was absorbed into the porous cupel. Silver metal was retained in the cupel. The lead from the skimmings and discarded cupels was recovered as metal upon heating with charcoal.

Native gold-silver alloys

Native gold itself often contained quite considerable quantities of silver. These silver-gold alloys, known as electrum, may be separated in a number of ways, but presumably the earliest was by heating in a crucible with common salt. In time and with repetitive treatments, the

silver was converted into silver chloride, which passed into the molten slag, leaving a purified gold. Cupellation was also employed to remove from the gold such contaminates as copper, tin, and lead. Gold, silver, and lead were used for artistic and religious purposes, personal adornment, household utensils, and equipment for the chase.

In the thousand years between 500 BC and AD 500 man made a vast number of discoveries that were significant in the growth of metallurgy. The Greek mathematician and inventor Archimedes, for example, demonstrated that the purity of gold could be measured by determining its weight and the quantity of water displaced upon immersion—that is, by determining its density. In the pre-Christian portion of the period, the first important steel production was started in India, using a process already known to ancient Egyptians. Wootz steel, as it was called, was prepared as sponge (porous) iron in a unit not unlike a miniature blast furnace. The product, an unmelted sponge iron, was hammered while hot to expel slag, broken up, then sealed with wood chips in clay containers and heated until the pieces of iron absorbed carbon, converting it to steel. Steel is defined as iron that contains carbon as an essential constituent in any amount up to about 1.5 percent. The steel pieces could then be heated and forged to bars for later use in fashioning articles, such as the famous Damascus swords made by medieval Arab armourers.

Arsenic, zinc, antimony, and nickel may well have been known from an early date but only in the alloy state. By 100 BC mercury was known and was produced by heating the sulfide mineral cinnabar and condensing the vapours. Its property of amalgamating (mixing or alloying) with various metals was employed for their recovery and refining. Lead was beaten into sheets and pipes, the pipes being used in early water systems. The metal tin was available and Romans had learned to use it to line food containers. Although the Romans made no extraordinary metallurgical discoveries, they were responsible, in addition to the establishment of the brass industry, for contributing toward improved organization and efficient administration in mining.

During the early centuries of the Christian Era, many scientific philosophers, known as alchemists, believed that it was possible to transmute one element into another. Although this was not achieved, the discoveries of the alchemists contributed to a better understanding of metals and their compounds and laid the foundations of modern chemistry.

Beginning about the 6th century, and for the next thousand years, the most meaningful developments in metallurgy centred around iron making. Great Britain, where iron ore was plentiful, was an important iron-making region. Iron weapons, agricultural implements, domestic articles, and even personal adornments were made. Fine quality cutlery was made near Sheffield. Monasteries were often centres of learning of the arts of metalworking. Monks became well known for their iron making and bell founding, the products made being utilized either in the monasteries, disposed of locally, or sold to merchants for shipment to more distant markets.

Fine Sheffield cutlery

In Spain, another iron-making region, the Catalan forge was invented, and its use later spread to other areas. A hearth type of furnace, it was built of stone and was charged with iron ore, flux (substance added to help fusion), and charcoal. The charcoal was kept ignited with air from a bellows blown through a bottom nozzle (tuyere). The sponge iron that slowly collected at the bottom was removed and upon frequent reheating and forging was hammered into useful shapes.

By the 14th century the furnace was greatly enlarged in height and capacity. Because the iron was in contact with hot charcoal for longer periods of time, it absorbed more and more carbon, which lowered its melting temperature sufficiently for it to exist in the molten state.

The first purposeful use of the iron from medieval furnaces was not for steelmaking but for producing cast-iron castings. Cast iron is a variety of iron containing about 2 to 4.5 percent carbon, poured molten into a mold to solidify in a desired shape. Chemically it is the same as pig iron, the crude iron from a furnace; normally it is brittle and cannot be hammer welded, forged, or rolled and is more easily fused than steel.

Developments after 1500. During the 16th century, metallurgical knowledge was recorded and made available. Two books were especially influential. One, by the Italian Vannoccio Biringuccio, was entitled *De la pirotechnia* (Eng. trans., *The Pirotechnia of Vannoccio Biringuccio,* 1943). The other, by the German Georgius Agricola, was entitled *De re metallica.* Biringuccio was essentially a metalworker and reported methodically and with unusual completeness the knowledge he had acquired. His book dealt with smelting, refining, and assay methods (methods for determining the metal content of ores) and covered metal casting, molding, core making, and the production of such commodities as cannon and cast-iron cannonballs. His was the first methodical description of foundry practice.

Metallurgical writings

Agricola, on the other hand, was a miner and an extractive metallurgist; his book considered prospecting and surveying in addition to smelting, refining, and assay methods. He also described the processes used for crushing and concentrating the ore and then, in some detail, the methods of assaying to determine whether ores were worth mining and extracting. Some of the metallurgical practices he described are retained in principle today.

From 1500 to the middle of the 19th century, metallurgical development was largely concerned with improved technology in the manufacture of iron and steel. In England, the gradual exhaustion of timber led first to prohibitions on cutting of wood for charcoal and eventually to the introduction of coke, derived from coal, as a more efficient fuel. Thereafter the iron industry expanded rapidly in Great Britain, which became the greatest iron producer in the world. The crucible process for making steel, introduced in England in 1740, by which the materials were placed in clay crucibles heated by coke fires, resulted in the first reliable steel made by a melting process.

Introduction of coke

Closely associated with steelmaking was the making of tinplate, an industry that began to flourish early in the 17th century. The process was said to have originated in the metal-rich region of Bohemia (Czechoslovakia). The canned food industry that developed late in the 18th century stimulated a large expansion in consumption of tinplate.

The earliest methods of steelmaking suffered the serious inherent disadvantage that the iron, normally in the form of pig iron, remained in intimate contact with charcoal, which made it difficult to reduce the carbon content to a point at which the metal became ductile. This difficulty was overcome by the puddling process invented in England in 1784. In it, melting was accomplished by drawing hot gases over a charge of pig iron and iron ore held on the furnace hearth. During its manufacture the product was stirred with iron rabbles (rakes), and, as it became pasty with loss of carbon, it was worked into balls, which were subsequently forged or rolled to a useful shape. The product, which came to be known as wrought iron, was low in elements that contributed to the brittleness of pig iron and contained enmeshed slag particles that became elongated fibres when the metal was forged. Later, the use of a rolling mill equipped with grooved rolls to make wrought-iron bars was introduced.

These developments in metal recovery and metallurgical technology in the early 17th century and thereafter ran parallel with and played an important part in the Industrial Revolution sweeping England and the Continent. In England, especially, metallurgical skills were well developed in extractive metallurgy, exploitation of mineral wealth, manufacture of iron and steel, and in the growth of industries that depended upon basic metal-making processes.

Metallurgy developed from an art dependent on trial, experience, and rule of thumb into a combination of art and science involving knowledge of the principles underlying the behaviour of metals—principles grounded largely in chemistry and physics. It would be unprofitable, therefore, to continue the history of metallurgy, especially during the 19th and 20th centuries, along purely chronological lines. The reader will often find, however, that historical facts pertaining to certain significant processes and events

in metallurgical technology will be given in the account of processes and events.

MINERAL DRESSING

Before the Metal Age began, man had to discover that metals could be extracted from mineral-bearing rocks. Through ingenuity and, doubtless, accident, this came about; thus, metallurgy, from its beginning, was concerned with the problem of treating crude ore to segregate valuable minerals and to extract metals from them. This activity is embraced in extractive metallurgy, which consists of two phases: mineral dressing and process metallurgy.

The beneficiation (concentration) of ore has undoubtedly been practiced in one way or another since man first produced copper as far back as 3,000–4,000 BC. Treatments of ore involve, first, a separation of particles of different minerals with little or no change in their chemical composition and, second, a breaking down of the valuable minerals to isolate the desired metallic elements. The mechanical operations necessary to separate mineral from gangue are called mineral dressing. Modern processes in mineral dressing have become increasingly directed toward using fundamental principles governed by laws of physics, chemistry, and electricity. The practice of mineral dressing constitutes only a part of the total effort involved in the recovery of metals from ores. Because most ores as mined contain more than one species of mineral, the first step in the general scheme of ore treatment is to free them from an interlocked state physically by comminution—a process that includes the breaking down of the ore to fine powders, their ultimate fineness depending upon the fineness of dissemination of the desired mineral particles to permit their separation from waste. As a rule, comminution begins by dry crushing the ore to below a certain size and finishes by wet or dry grinding to liberate the minerals from the unwanted gangue.

Types of treatments

Crushing and grinding. In primitive times, crushers were either small hand-operated pestles and mortars or millstones that could be turned by men or horses or operated by waterpower. Today, crushing and grinding for liberation before beneficiation is generally done in stages, with the possibility of some beneficiation between stages. Generalization is difficult because each ore has individual characteristics that influence the crushing and grinding procedures.

Some ores in nature occur as mixtures of discrete mineral particles, such as gold in gravel beds in streams and diamonds in mines. These mixtures require little or no crushing, as the valuables are recoverable using other techniques. Most ores, however, are made up of hard, tough rock masses that must be crushed before the valuable minerals can be released. Coarse, or primary, crushing today is most often done by squeezing until fracture occurs using large crushers capable of taking crude ore in sizes slightly greater than two metres (about seven feet) in its largest dimensions and reducing it to lumps of from 25 to 150 millimetres (one to six inches) in size. The product from primary crushing may then pass to one of several types of secondary crushers, where it is further reduced in size to a fraction of a centimetre. The main task of primary-, secondary-, and, sometimes, tertiary-stage crushing is to prepare the ore for fine grinding by either wet or dry methods. The minimum size of crushing consistent with economic considerations is in the order of six to 13 millimetres (0.25 to 0.5 inch).

Tumbling mills, horizontal cylinders containing grinding media of various kinds, such as rods, pebbles, or steel balls, are almost universally used for fine grinding to secure further comminution. Screens woven of wire to a given mesh size customarily are used to determine and to control particle size. Mesh is defined as the number of openings per linear unit; in the widely used Tyler Series of screens, it ranges from one to 128 per centimetre (three to 325 per inch).

Concentration or mineral preparation. Hand sorting is the simplest form of concentration. The Chinese in 500 BC are reported to have processed marginal ores by hand sorting. Occasionally, hand sorting is still used to remove contaminating and waste products as well as to single out especially rich materials as ore passes along moving tables or conveyor belts. Minerals to be sorted in this fashion must have sufficiently distinctive colour, lustre, texture, or general appearance for the picker to distinguish the wanted from the unwanted materials. Another simple form of concentration is the art of panning, sometimes used in gold recovery; it concentrates heavy particles at the bottom of the pan while the lighter waste material is washed off on top. Still another simple concentrating device is an ordinary screen where the valuable mineral is either finer or coarser than the bulk of the ore. Washing coarse ore with a stream of water to remove fine sand and clay is often used.

Gold recovery by panning

In large-scale operations, where crushed and ground ores are to be concentrated, advantage is taken of the more common physical, chemical, magnetic, and electrical characteristics of the minerals. The basis of an economically practical process is that the particles of concentrate and gangue must not respond alike in any important degree to the specific separating force employed.

Gravity concentration. Gravity methods used to beneficiate ores are dependent on the combined effect of mass and shape of the particles in determining their movement relative to flowing water. In one method the water flows vertically upward either continuously, as in hydraulic classification, or with an oscillating motion. To achieve separation by hydraulic classification, the rate of settling between the heavy and light mineral particles in the continuously flowing current of water must be reasonably different. By adjusting the rate of upward flow of a continuous stream of water, a classification between coarse and fine or between heavy and light particles is possible.

In using oscillating motion of water, a process known as jigging, the size variable is first reduced by screening, after which the differential settling rate of the particles is exploited. Jigging is one of the oldest methods for concentrating ores and is still widely used in some areas. In jigging, water is pulsed strongly upward and downward through a bed of particles. After sufficient pulsation, the top portion of the bed becomes an accumulation of the lighter gangue, which can be rejected, while the lower part is the heavier enriched concentrate.

Another method makes use of relatively undisturbed water that has been mixed with slow-settling heavy minerals (finely ground ferrosilicon, for example) to provide a liquid that can be maintained at a high density. When ore is fed into the liquid, particles of lower density than the liquid float while particles of higher density sink. This is called sink-and-float, or heavy media, concentration.

In a third method of gravity concentration, the ground ore is carried either horizontally or down an inclined plane in a flowing stream of water. Separation depends on forces acting upon the particles as they move in a layer of water, on the rate of fall of the particles, and on their resistance to lateral displacement once they contact the floor of the concentrator. Spirals, a type of gravity separator using this principle, have been found especially satisfactory for treating ores of the finer sizes. The device is a spiral channel with a modified semicircular cross section into which the pulp (ground ore and water) is fed from the top. As the pulp flows down, the heavier particles concentrate on the inner side of the stream and are removed at conveniently located ports.

Use of spiral separators

For these gravity methods to provide satisfactory separation, a marked difference in density of the wanted minerals and gangue must exist. Even when such a difference exists, these methods can become cumbersome and inefficient if the average particle is so small that its surface friction becomes the controlling property determining its movement through the surrounding fluid.

Separation by flotation. The process of concentrating ores by flotation was developed on a commercial scale early in the 20th century to remove very fine mineral particles that formerly had gone to waste in gravity concentration plants. Flotation has now become the most widely used process for extracting many minerals from their ores. It is based on the fact that the wettability of the different minerals, upon treatment, may be altered so that certain mineral particles remain unwetted and adhere to

Extensive use of flotation

air bubbles. Through the use of relatively small amounts of chemicals or oils, particle surfaces may be altered to an air-avid or water-avid condition—that is, the surfaces prefer either water or air.

Thus, in a pulp of mineral particles and water through which air bubbles are passing, the air-avid particles attach themselves to the bubbles and are carried to the upper surface of the pulp, where they enter the froth; the froth containing these particles can then be removed. The water-avid particles remain in the pulp.

It is possible not only to float various minerals from gangue but also to separate certain valuable minerals from others selectively. Many complex ore mixtures formerly of little value have come to be among the principal sources of certain metals by means of the flotation process.

Magnetic separation. Magnetic separation is the most used and probably the most efficient method of beneficiating iron ores, the principal mineral of which is the iron oxide known as magnetite (Fe_3O_4). The technique is based largely on the differing degrees of attraction exerted on various minerals when influenced by magnetic fields. It is applicable to the treatment of both coarse and fine particles, either wet or dry. In concentrating magnetic iron ore, low-intensity wet magnetic separators are most commonly used.

When ores are conveyed through a magnetic field of the proper strength, the magnetic portion is separated from the nonmagnetic portion. High-intensity magnetic separators have been developed for the separation of weaker magnetic iron minerals, thus permitting beneficiation of such minerals as hematite (Fe_2O_3), limonite ($2Fe_2 \cdot 3H_2O$), siderite ($FeCO_3$), franklinite ($ZnFe_2O_4$), ilmenite ($FeTiO_3$), and wolframite (Fe, Mn) WO_4.

Electrostatic separation. This is a concentration method between minerals of different electrical properties. It is based on the phenomenon of attraction between unlike electrical charges and repulsion between like charges.

The process has proved quite successful in the treatment of heavy beach sands from which the valuable minerals ilmenite ($FeTiO_3$), rutile (TiO_2), or zircon ($ZrSiO_4$) may be recovered. Its application to iron-ore beneficiation has also been considerable.

Amalgamation. Amalgamation is an ancient process still used to recover free gold. The process depends on the fact that when clean surfaces of gold or silver come in contact with mercury they readily form an alloy called an amalgam. The gold (or silver) is recovered by heating the amalgam in retorts until the mercury is distilled off.

Fluidized-bed technique. The fluidized-bed technique has been adopted in a number of processes to increase the economic recovery of metals from ores and concentrates. In this process, the material to be treated is kept in motion in a chamber by an upward current of gas that maintains the individual particles in suspension. Through precise adjustment of the velocity of the gas to conform with the grain size and specific gravity of the materials to be treated, a floating mixture of gas and solids is generated that behaves almost like a liquid. Increasing attention has been given to studies of high-temperature reactions in fluidized-bed systems, "high-temperature" implying the region of 1,370°–1,540° C (2,500°–2,800° F), the practical limit of fuel-combustion heating.

Agglomeration. For some time, the metallurgical industry has been confronted with the task of agglomerating ore fines (fine particles) and concentrates. Although a number of processes have been studied intensively, they have been unable to compete economically, except in special instances, with sintering and pelletizing.

In sintering, a mixture of fine material, fuel, and water is fed onto a grate where its surface is ignited. Air is drawn through the bed to promote combustion. The passage of a high-temperature flame front through the bed causes incipient fusion with the formation of a porous, coherent mass called sinter. Although sinter is too friable (subject to crumbling) to resist excessive breakage during shipping and must, therefore, be made near the site of its use, its performance in metallurgical processes related to control of the size of the lumps has been excellent, creating an increasing demand for it throughout the world.

In pelletizing, moistened, finely divided materials, which often contain binders to enhance balling and limestone to promote high-temperature strength and to serve as flux in subsequent use, are fed into balling pans that deliver roughly spherical aggregates or pellets that must be dried and fired at a suitable temperature to harden. Pellets are usually less than 13 millimetres (0.5 inch) in size and sufficiently strong after firing to resist breakage; hence they have no serious limitation from the standpoint of shipping. The pelletizing process is particularly suitable in agglomerating very fine concentrates and ores.

Sampling and analysis. Routine sampling is undertaken to provide information needed for technical control and economic appraisal of ores and concentrates. By sampling, it is possible to direct attention to adjustments that may be needed during the process of mineral dressing. Sampling and sample-preparation methods frequently are major factors in determining the speed, reproducibility, and accuracy of analyses obtained with the use of modern instruments.

Even before the 16th century, comprehensive schemes of assaying (measuring the value of) gold, silver, lead, copper, tin, mercury, and bismuth ores were known, using procedures that do not differ materially from those used in modern times. Although conventional methods of chemical analysis are used today to detect and estimate quantities of elements in ores and minerals, they are slow and not sufficiently accurate, particularly at low concentrations, to be entirely suitable for industrial control. As a consequence, to achieve greater efficiency, sophisticated analytical instrumentation is being used to an increasing extent as an integral part of overall process-control schemes in mineral dressing.

Use of instrumentation

Emission spectroscopy and X-ray fluorescence spectroscopy, for example, are widely used for metallurgical and ore analysis. In emission spectroscopy, an electric discharge is established between a pair of electrodes, one of which is made of the material being analyzed. The electric discharge, which may be a spark or a sustained arc, vaporizes a portion of the sample and excites the elements in the sample to emit characteristic spectra. Detection and measurement of the wavelengths and intensities of the emission spectra reveal the identities and concentrations of the elements in the sample.

In the X-ray-fluorescence spectroscope method of analysis, a sample is bombarded with X-rays, whereupon elements in the sample give off fluorescent X-radiation of wavelengths characteristic of these elements. The amount of emitted X-radiation, measured with a detector, is related to the concentration of individual elements in the sample. The sensitivity and precision of this method are poor for elements of low atomic number (*i.e.,* few protons in the nucleus, such as boron and beryllium), but for slags, ores, sinters, and pellets where the majority of the elements are in the higher atomic number range, as in the case of gold and lead, the method has been generally suitable.

The trend toward achieving automatic control in metallurgical operations demands development of new and improved methods and devices to gain a better understanding of the complex physical and chemical nature of the materials being handled. Great strides in this direction have been made through the use of instrumentation, more sophisticated control systems, and the electronic digital computer for processing of data and process control.

PROCESS METALLURGY

Process metallurgy comprises a series of steps or processes by which the impure minerals that have been isolated either by mining or mineral dressing are reduced to metal, refined, alloyed, and made available according to specifications. An old art involving chemical changes that break down, or decompose, the impure minerals into their metallic and nonmetallic components, it has become, with the aid of chemical science, the chief element in modern metallurgy. The steps followed to recover metals from their minerals are normally selected on the basis of several considerations, among them the chemical nature of the mineral itself, which may be a sulfide, oxide, carbonate, silicate, or other species. There are three classes of meth-

Classes of recovery methods

ods: (1) pyrometallurgy, the use of heat; (2) electrometallurgy, the use of electricity; and (3) hydrometallurgy, the use of water.

Pyrometallurgy. Pyrometallurgy is any metallurgical extractive process involving high temperatures that depends on heat action. Normally, the heat is supplied either by combustion of such fuels as coke, oil, or gas, or by electric power.

Gas–solid reactions. Ores in the form of oxides may be treated directly, but sulfide ores first must be roasted. Roasting is heating in air without fusion, its chief purpose being to transform sulfide minerals into an oxide. The sulfur escapes as a gas in the form of sulfur dioxide. Roasting is often employed to yield a product that is chemically and physically more amenable to some later process. It is commonly applied to such minerals as galena (PbS), sphalerite (ZnS), and chalcopyrite ($CuFeS_2$) to render them in suitable state for conversion to the metal (reduction) by carbon.

Many roasting reactions are exothermic—that is, once the reaction has started on contact with an air blast, the heat liberated enables it to continue spontaneously. Careful control during the roasting is required because overheating may ensue, producing fusion and the cessation of oxidation. The product of roasting, often purposely brought to an agglomerated state, is known as calcine.

A development for processing ore using gas-solid reactions to extract metals is chlorination, in which the oxides of some metals are converted to chlorides by the reaction of chlorine only, and the oxides of other metals are converted by the additional help of carbon. Chlorination is especially applicable in the processing of complex ores where the conventional metallurgical methods are inadequate. Low working temperatures, high yield, and sharp separation are the essential characteristics of chlorination. The chlorinating agents have a high reactivity, enabling the whole reaction to be carried out at much lower temperatures than those employed in usual smelting techniques. The advantages of chlorination are illustrated in the processing of titanium, for its oxide is technically impossible to reduce because of the high temperatures involved. Chlorine in the presence of carbon decomposes titanium dioxide at a relatively moderate temperature, forming titanium tetrachloride, from which the metal can be obtained by reduction with magnesium.

Chlorination in titanium processing

In the processes so far described, the products obtained are in a fairly divided condition. Although this result is suitable for some later smelting processes, in the production of lead and iron in the blast furnace a coarse product is essential. In the preliminary preparation of these finely divided materials for smelting, agglomeration is frequently important and may be achieved by sintering or pelletizing as described earlier.

Distillation. Metals obtained during the initial stages of processing usually contain impurities derived from the raw materials. To get a higher order of purity, further treatments and refining procedures, such as distillation, are used. Such volatile metals as zinc, mercury, and cadmium may be readily purified by distillation.

In zinc production, carbon reduction of the oxide in retorts is the major industrial method employed. The ore is mixed with pulverized coal, and the mixture is subjected to temperatures sufficiently high to reduce the oxide, generating a gas of zinc vapour, which is then liquefied in a condenser without contact with oxidizing gases.

Efforts to develop a blast-furnace process for zinc reduction have been successful. The furnace follows the standard blast-furnace design except that volatilization of zinc prevents it from being tapped from the furnace bottom as other metals are. Instead, the zinc vapours pass upward and are rapidly condensed with a shower of molten lead in an adjoining condenser. Impurities are rejected from the furnace as slag composed primarily of the oxides of iron, calcium, silicon, and aluminum. Subsequently, the metal is refined by redistillation to remove the chief impurities, consisting of lead, iron, and cadmium. Because lead and iron are less volatile than zinc and cadmium is more volatile, the operation is divided into two stages. In the first, zinc and cadmium are distilled from the lead and iron; in the second, cadmium is separated from zinc by fractional distillation, known as refluxing, followed by condensation (see below *Zinc production*).

High volatility is a characteristic common to some of the unique compounds of carbon monoxide with the heavy metals. In general, these compounds, called carbonyls, are produced by direct action of carbon monoxide on the finely divided metal. The first to be discovered, late in the 19th century, was nickel carbonyl, $Ni(CO)_4$; since then, carbonyls of iron, cobalt, chromium, molybdenum, tungsten, and manganese, as well as of some of the less common metals, have been produced. One of the important uses of metal carbonyls is in the preparation of metal of exceptionally high purity.

The carbonyls of nickel, iron, cobalt, and molybdenum can be prepared by direct synthesis; *i.e.,* by treatment of the finely divided metal with carbon monoxide. The carbonyls of the other transition elements, such as silver and copper, are prepared from the halides, compounds formed with one of the halogen elements—fluorine, chlorine, bromine, and iodine.

Direct synthesis of carbonyls

At one time, nickel carbonyl, $Ni(CO)_4$, was employed in a commercial process for the recovery of nickel from its ore. The gaseous carbonyl was heated and passed into vertical columns filled with pure nickel shot (small spheres), whereupon the carbonyl gas decomposed and deposited metallic nickel on the surface of the shot.

Iron pentacarbonyl, $Fe(CO)_5$, is commercially available and is used in industry in the preparation of iron powder (carbonyl iron) obtained by thermal decomposition of the carbonyl. Because of its high purity and desirable magnetic properties, carbonyl iron is used by the electronics industry for the production of magnetic devices. Carbonyl iron also finds application in powder metallurgy.

Ferrous and nonferrous metallurgical processes. Iron- and steel-manufacturing techniques are described in detail in the sections *Iron production* and *Steelmaking.*

Production techniques for aluminum, copper, lead, zinc, and other nonferrous metals of economic importance also receive attention in subsequent sections.

Melting. The preparation of metal in the liquid state is a distinct stage in most metallurgical processes. Melting is only the beginning of a series of operations by which the metal is brought into its finished form. Although the procedures by which the more common ferrous and nonferrous metals are brought to the molten state do not differ greatly among the metals, there are reactive metals, such as titanium, that, because of their affinity for the gases nitrogen, oxygen, and hydrogen, must be melted and cast under vacuum or in the presence of an inert gas.

Special melting conditions

Several processes of melting, described below, promise to become increasingly significant in upgrading the quality of metals, particularly of steels. These processes are limited in production capacity, but they are valuable because of the high grade of metal they produce by melting or remelting.

In vacuum-induction melting, the charge, which can be either solid or molten metal, is induction-heated (*i.e.,* heated by a current produced in the material by an external coil) under vacuum. In the consumable-electrode vacuum-arc process, by contrast, the charge serves also as the electrode. As the electrode melts, the vacuum removes the gases; water cooling limits contamination of the melt by the furnace walls (usually copper). The process is useful especially for consolidating reactive and refractory metals and for producing high-temperature, high-strength alloys. When metal of particularly high quality is needed, the electrodes are prepared from ingots that previously were vacuum-induction melted.

Electroslag refining was developed and is applied widely in the U.S.S.R. In this process also, the electrode (sometimes two) is the charge. Using no vacuum, the furnace is charged with a slag that covers the melt and eventually the ingot. The process is said to produce metal equal in quality to vacuum-arc remelted steel.

Alloying. A basic condition of metallurgic technology is that it must constantly meet increasingly exacting demands from the metal-using industries. Exact types of metal are specified, including the precise components they should contain so that specific mechanical and other crit-

ical properties can be forecast. The deliberate addition of various chemical elements to the molten metal, a process known as alloying, has gone far to make it possible to attain consistently good performance of finished parts in the enormously varied conditions of modern metal use.

The chemical elements to be incorporated by alloying to liquid steel were often added as ferroalloys rich in the desired element. They include alloys of iron with boron, calcium, chromium, niobium, manganese, molybdenum, nitrogen, phosphorus, selenium, silicon, tantalum, titanium, tungsten, vanadium, and zirconium. With the production of alloys for steelmaking purposes now being carried out in electric-reduction and other types of furnaces, however, a number of the alloy additions contain relatively little iron. Elements such as nickel, copper, lead, cobalt, and aluminum are usually added as virgin metal or scrap. The term addition agent, instead of ferroalloy, is preferred to describe any of the materials added to molten steel for altering composition and properties. In preparing nonferrous alloys, the elements added in any quantity are usually in the form of virgin metal.

Alloying additions may be added at various times in the melting process, such as with the charge, in the molten batch near the end of the melting period, in the ladle or in the mold. Timing of such additions is determined by their effect on the temperature of the molten metal, the ease with which the addition goes into solution, the susceptibility of an addition to oxidation, and the formation and elimination of reaction products. The principal purposes of alloying are to create materials from base metals that have increased strength and toughness, improved response to heat treatment, less susceptibility to embrittlement at low temperatures, or increased resistance to heat, corrosion, or wear, or a host of other properties. Alloys composed of a complex group of elements are continually being tailored to satisfy specific needs.

Casting. The operations of casting metals in the liquid state vary widely according to the form in which the metal is finally to be used. Broadly speaking, metallic objects tend to be divided according to their mode of production into two types: cast and wrought. So far as metallurgical operations are concerned, cast materials are brought to final form by permitting molten metal to fill and solidify in molds of desired shape. On the other hand, wrought materials generally begin their career by solidifying in a metal mold of simple shape and being brought to desired form by working, either cold or at elevated temperature, using practices such as rolling, pressing, forging, stamping, drawing, and extrusion. (For descriptions of these practices, see below *Principal areas of metal processing.*) Casting is a process practiced by foundries all over the world as a basic method for the production of shapes, using in one form or another almost all of the metals known to man.

Types of casting processes

Most castings are made by pouring metal into sand molds prepared from mixtures of moistened sand and clay packed over the faces of the pattern. Recently, however, specialized processes have been developed to produce castings having greater dimensional accuracy, smoother surfaces, and more finely reproduced detail than is obtainable in sand molds. Important processes among these are plastic-mold, composite-mold, investment, permanent-mold, and die casting.

The principal mold ingredient in plastic-mold casting is calcium sulfate, to which small amounts of two or more other substances are added to attain specific properties. The molds are used to produce nonferrous castings of superior surface and dimensional accuracy.

Composite molds are assembled from several components, among which at least one varies from the others in the molding materials used. The goal is to derive the specific advantages of each molding material without being penalized by its disadvantages. Composite molds are mainly for alloyed aluminum castings classified as premium-quality castings or engineered castings.

Investment casting (also known as the lost-wax and precision-casting process) employs molds that are produced by surrounding an expendable pattern of wax or plastic with a refractory slurry that sets at room temperature. The wax or plastic is then melted out and the metal poured into the hot cavity that remains. In one variation, the patterns are of frozen mercury, which is melted out, leaving the impression to be cast. The investment-casting process is used particularly on alloys that cannot be softened or made plastic for working by heat but must be shaped to exact dimensions.

Permanent-mold castings are produced in metal molds that can be used over and over again. The molds consist of two or more parts so that they can be readily disassembled to eject the casting and then reassembled to be ready for the next casting. The process is particularly suitable for high-volume production of simple castings, such as gray iron pipe having a fairly uniform wall thickness and no undercuts or intricate internal coring. Permanent mold casting is used successfully for casting gray iron and alloys of aluminum, magnesium, copper, or zinc.

The process of die casting has found extensive application for the lower melting nonferrous metals because of the accuracy of dimensions attainable and especially because of the speed with which castings are made. The process consists generally of forcing molten metal into metal dies or molds under pressure; the die is then opened by a machine and the casting ejected. It is widely used in making castings of low-melting alloys of zinc, tin, aluminum, magnesium, and, to a limited extent, for certain types of brass and bronze. The difficulties of the process increase rapidly as the melting temperatures of the metal rise. The process in Europe is termed pressure die casting.

Ingot solidification. By far the largest proportion of liquid metal produced throughout the world is cast in molds for ingots, which are then used for future mechanical working such as in rolling and forging. Steel ingots can range in size from small blocks weighing a few pounds to huge, tapered, octagonal masses weighing more than 100 tons.

In pure metal free from gas, the solidification process is comparatively simple, but it is by no means simple in complex alloys such as steel and the many nonferrous alloys that often contain dissolved gases as well as suspended oxides, sulfides, hydrides, nitrides, and silicates. The principal factors that play a part in ingot solidification are: contraction, which comes about because the metal after solidification generally occupies a smaller volume than in the liquid state; gas liberation, which comes about because most liquid metals dissolve large volumes of gas, especially oxygen and hydrogen, that are driven off during solidification as a result of their lower solubility in the solid state; segregation, resulting from impurities and various constituents that may undergo nonuniform separation upon solidification; and contraction stresses, resulting from the fact that the outer portions of an ingot necessarily solidify while the inner portions are still liquid.

The amount of gas, chiefly oxygen, dissolved in liquid steel and the amount of gases released during solidification determine the type of ingots; these are termed killed, semikilled, capped, or rimmed in ascending order of amount of gas retained.

Nonferrous metal and alloy ingots and slabs, although subject to the same general laws and phenomena as steel ingots and slabs, are rarely as large, so that many of the problems arising from large size, very slow cooling, and extensive segregation common to large steel ingots are not encountered in the nonferrous industries. On the other hand, blowholes (captured gas bubbles) lying just under the ingot skin, which often are regarded as normal in steel and weld during subsequent working, are not tolerable in nonferrous metals. They do not weld but tend to open out during the annealing process following cold rolling, causing defects.

The casting in simple shapes of both ferrous and nonferrous metals is accomplished by continuous casting processes. Many continuous casting techniques had been used for years to cast shapes of nonferrous metals, particularly wire rod. Adapting these techniques to ferrous products has been difficult, largely because of the much higher casting temperature. The advantages promised by continuous casting, however, have led to persistent experimentation and to step-by-step improvements that have resolved most

Continuous casting processes

of the problems for steel in the period since 1950. The big economic gain from continuous casting is the high yield of sound steel, more than 94 percent in contrast to about 80 percent for conventional ingot practices.

Electrometallurgy. Electrometallurgy is the science and technology of metallurgical operations conducted by electrochemical means—specifically, the use of electrolytic processes in the extraction and refining of metals and in electroplating. It is concerned with two branches: the conversion of electrical energy into chemical energy (as in electrowinning, electrorefining, and electroplating, described below); and the utilization of thermal energy derived from electrical energy, which is mainly concerned with the electric furnace and its applications. The scope of both branches is very broad.

Electrowinning. Electrowinning (winning or recovering the metal by an electrolytic process) can be carried out on the molten metal or compound, or on a solution of a compound of the metal. In the first process a raw ore is fused (melted) or converted into an easily fusible compound, and the resulting current-conducting solution (electrolyte) is electrolyzed (decomposed by an electric current), producing a relatively pure metallic deposit on one electrode. The second process involves leaching the raw ore (see below *Hydrometallurgy*) with an aqueous solution; the resulting solution, or electrolyte, rich in the metal, then is recovered and electrolyzed, again producing a relatively pure metal deposit on one electrode.

Production of aluminum and magnesium

The production of aluminum and magnesium are typical electrowinning operations. Unlike the other common industrial metals, pure aluminum is not produced by the direct smelting of its ores. It is produced electrolytically by a process that was discovered almost simultaneously in the United States and France in 1886 and is the basis of the modern aluminum industry. Purified alumina (Al_2O_3) is dissolved in molten cryolite (Na_3AlF_6) and electrolyzed with direct current. The mixture is contained in a carbon-lined steel box that serves as one electrode while carbon rods or blocks dipping into the molten bath form the other electrode. Under the influence of the current, the oxygen of the Al_2O_3 is deposited on the carbon, which is thereby slowly burned, while the molten aluminum is deposited on the carbon lining of the cell, collecting in the bottom. Additional alumina is stirred into the electrolyte from time to time and the molten metal is removed. The alumina required is somewhat more than 0.7 kilogram per kilogram (1.9 pounds per pound) of metal.

The English scientist Michael Faraday was the first to succeed (1833) in producing metallic magnesium by electrolysis of molten magnesium chloride. The electrolytic process was further developed and was used exclusively until about 1941, when other techniques, employing a thermal process, came into use for a small proportion of magnesium production. In the U.S., magnesium is produced by electrolysis of fused magnesium chloride, the pure chloride coming largely from seawater. The electrolytic cells are steel shells or pots containing suspended carbon rods for one electrode and iron or carbon for the other electrode.

Electrowinning is also employed in producing an important percentage of zinc and cadmium. In the extraction of both metals by electrolysis, the solution is the sulfate of the metal, which is electrolyzed in cells using aluminum for one electrode and lead for the other.

Electrorefining. In electrorefining, the crude metal to be refined forms one electrode and is suspended in a suitable aqueous electrolyte. Under electric current action this electrode dissolves, and pure metal is deposited on another electrode, usually composed of the same metal as that deposited. Electrorefining processes are employed commercially in the extraction of such metals as copper, nickel, and lead and the precious metals gold and silver. The production of electrolytic copper and nickel are typical electrorefining processes.

More than three-quarters of the copper produced in the U.S. is electrolytically refined. Slab electrodes of fire-refined copper are further refined by immersing them in a solution of copper sulfate in an electrolytic cell with the other electrode being of copper. In the multiple system, separate anodes (positive electrodes) and cathodes (negative electrodes) are used, the cathodes consisting of thin sheets of high-purity copper, known as starting sheets. In an individual cell all the anodes are connected in one parallel group and all the cathodes in another. In the series system no starting sheets are used, the electrodes of impure copper being in series in a single cell and serving as both anode and cathode. This arrangement allows the impure copper to be corroded from the face of an electrode while pure copper is deposited on the back. Of the two systems, the multiple system is most widely used. Its advantages lie chiefly in the ability to treat copper of lesser purity and to operate at a lower voltage. Gold, silver, and other valuable insoluble impurities are recovered as by-products of electrorefining of copper.

In the electrorefining process for nickel, the impure nickel is pressed into anode plates. The electrolyte used is composed of a nickel sulfate-chloride solution containing some boric acid, sodium sulfate, and about 40 grams of nickel per litre (2.5 pounds per cubic foot). Nickel starting sheets are used as cathodes. Elaborate precautions, such as the use of canvas diaphragms, are taken to separate the anodes and cathodes, to prevent the deposition of copper and iron, as well as nickel.

Electroplating. Electroplating is the production of a thin coating of one metal on another by electrodeposition. Practically any metal can be electroplated under properly controlled conditions. The process is used extensively for coating various base metals with zinc, cadmium, nickel, chromium, tin, aluminum, and many other metals. The deposited metal is obtained from a solution of its salts by the passage of an electric current through the solution. It is a complex process, and careful control of the current and of the chemical composition, temperature, and cleanliness of the bath are required for good deposits.

Hydrometallurgy. Hydrometallurgy is the science and technology of extracting and recovering metals chemically from ores by the use of aqueous solutions. The principles involved in hydrometallurgy are those of physical chemistry, inorganic chemistry, electrochemistry, and analytical chemistry. The production of metals by hydrometallurgy can be broadly described as extensions of these principles on an industrial scale.

Leaching operation

The operations usually involved are: dissolving a metal or metal compound by a suitable aqueous solvent (water alone or with added reagents), usually termed leaching; the separation of waste and purification of the leach solution; and the precipitation of metal from the leach solution by chemical or electrolytic means. Although the technology of hydrometallurgy was developed at the turn of the 20th century for recovering gold from low-grade ores, its features were practiced as early as the 16th century.

An important historical prerequisite to the development of hydrometallurgical procedures was the advent of commercial direct-current generators to provide an efficient low-cost technique for precipitating metals by electrolysis from aqueous solutions. After the beginning of the 20th century methods were improved and were augmented by the addition of still other techniques. Consequently, there evolved the great copper-leaching plants in South America, the U.S., and Africa; vast aluminum-extraction plants throughout the world; the electrolytic zinc industry; plants producing nickel and cobalt from sulfide ores and from the lateritic deposits in Cuba and other tropical areas; the gold-producing plants of South Africa, Canada, Australia, and the U.S.; and the gigantic plants in many parts of the world that employ hydrometallurgical techniques to extract uranium from ores that contain from a fraction of a pound to a few pounds of uranium in a ton of ore.

The feasibility and successful conduct of a hydrometallurgical operation depend on a variety of factors, any or all of which may be critical. An adequate supply of water is, of course, indispensable and, because metals are present as ions in aqueous solutions, a source of energy is required to convert the ions to metallic form. Beyond these essential features, several practical physical and chemical factors must be considered. The metals or minerals must be soluble in the leaching solution; the solvent must usually be such that it will dissolve selectively the desired metals or

Physical and chemical factors

minerals; the solution, when it contains valuable metals, should be such that the metals present can be readily separated from one another; and the solution must also be such that the desired metals can be precipitated from it by one means or another, producing either pure metal, a mixture of metals, or a metal salt. For economic reasons, it is usually necessary that the leach solution either be capable of being regenerated and used again or that it be susceptible of treatment to recover by-products.

Aspects of hydrometallurgical processes

The various aspects of hydrometallurgical processes include the preparation of the ores, concentrates, and other materials that are to be leached; the leaching itself, which generally applies to the selective dissolution of metal compounds and minerals from ore mixtures; the solution purification, which depends largely on the chemistry of the positive and negative ions in the solution; the precipitation of the metal to be extracted by electrolysis with direct current or by treating the solution with a reducing gas such as hydrogen; and the regeneration of the solution for its reuse.

PRINCIPAL AREAS OF METAL PROCESSING

The principal areas of metal processing covered below are cold and hot working, foundry processes, surface treatments, powder metallurgy, nuclear metallurgy, and heat treatment.

Cold and hot working. A great part of the world's metal production is cast into ingots that must be shaped to finished products by mechanical forces. Mechanical working not only brings the metal to desired shape but also changes its structure and properties.

When metal solidifies from a melt, especially if it does so in large masses and therefore slowly, it has a crystalline structure that is always relatively coarse and sometimes very coarse. Mechanical working breaks down this original cast structure and replaces it with a much finer and more satisfactory structure from the standpoint of physical properties. It is possible to produce a refined structure, comparable in scale with that of worked metal, by heat treatment alone; but this method is practical only with pieces of moderate size because larger masses cannot be cooled rapidly enough. Mechanical working also causes the redistribution of the nonmetallic constituents (inclusions), making them less harmful and, particularly for steel, serves to close up and even bring about welding of cavities.

Cold working consists in the plastic deformation (working) of a metal at such temperatures and at such rates that no recrystallization occurs during the process. Under such conditions, the metal is strain hardened—that is, increased in hardness and strength as a result of plastic deformation at temperatures lower than the recrystallization range. In general, these limiting temperatures are high in metals that melt at high temperatures and low in metals that melt at low temperatures. Cold deformation of metal thus produces fundamental structural changes accompanied by an increase in properties related to strength and a decrease in properties related to ductility.

Only when the working temperatures are above the recrystallization temperature, hence when softening occurs during the working process, can the term hot working be applied. During hot working, metal is deformed by mechanical working above the recrystallization temperature so that the coarse crystallinity of the cast state becomes progressively fragmented and rearranged to bring about the reconstruction and refinement of the crystalline structure. By proper selection of temperature and by mechanical working with dies or rolls, for example, various properties of the metal can be improved.

Forging. In forging, articles of metal are shaped by hammering or pressing metal blanks between pairs of forging dies. Hammering was the first method employed by man in shaping metal. In most forging processes, an upper die is attached to the ram of the forging hammer or press, so that it may be raised and dropped or forced under pressure against the rigidly supported, stationary lower die. Forgings are made in many metals and alloys, including copper, brass, bronze, nickel, aluminum, magnesium, steel, titanium, and various high-temperature alloys.

As the metal is processed from the ingot stage to various shapes (blooms, billets, slabs, and bars) for forging, its crystals are elongated, producing fibre-like flow lines of grain structure. In closed-die forging, these flow lines of grain structure are worked in such positions as to impart maximum strength relative to the stresses expected in the forged article. For maximum strength, care must be taken that the grain runs parallel to the axis of greatest bending stress.

Drop forging employs a drop hammer, so designed that the falling weight of the ram and upper die, often augmented with force from a steam or air piston, forces the heated metal into cavities cut into the face of each die. The metal being worked in a forging hammer is commonly at a sufficient heat to bring it to a plastic state.

Much progress has been made in forging metals cold—that is, at temperatures below the point at which the grain structure begins to recrystallize. Cold forming of heads has been used for many years in the manufacture of bolts and similar parts. More recently, however, cold forming has been replacing conventional hot forging, especially in forming of larger parts.

The formation of parts by forging, called chipless machining, is popular, for example, in the forge rolling of gears from solid blanks. Many other parts that can be made this way will become candidates for the process because of the savings in materials and labour.

Metal rolling. Of the many methods of shaping metals, rolling is most widely used. The rolling mill has passed through a succession of developments, not only in size, power, and productive capacity but also in design and in ability to produce varieties of shapes of sections with grooved rolls. The process of shaping metal by rolling consists essentially of passing the material between rolls revolving at the same peripheral speed and in opposite directions and spaced so that the opening between them is somewhat less than the depth of the material entering. Metals are rolled both hot and cold.

Hot rolling

Except in the few instances when ingots are forged, hot working by hot rolling is almost universally used to reduce ingots of metals by mechanical force from their original cross section through the intermediate stages to bar, structural shape, plate, sheet, and strip, the more common hot-finished forms that are produced. Hot rolling is done by passing the heated metal through pairs of heavy rollers, which are either flat or grooved, so that each successive pass lessens the cross section and brings that material more nearly to the required final shape. The hot-working temperatures are different for the different metals, but the highest are always those at which the lowest melting constituent in the structure starts to melt. The temperature range of hot working is limited by this temperature and by the temperature at which the metal remains sufficiently plastic at the lower end of the range, so that it will not rupture on working. The temperature at which the mechanical working is finished, known as the hot-working finishing temperature, is important in governing the final treatments necessary and the structural and other properties obtained.

Hot rolling accomplishes three main purposes: (1) reduction of cross-sectional area, (2) refinement of structure, and (3) shaping of the metal into desired form. Although the products of a rolling mill vary widely in weight and shape, a common characteristic is that each piece is of uniform cross section from end to end and is of a length limited only by the weight of the mass of metal being rolled. All the shaping is done in the intermediate passes of the mill. The flow of metal in the rolling operation is continuous and almost entirely in the longitudinal direction. With steel, nearly all hot rolling is done at full red heat, at which it is quite plastic.

The details of rolling practice vary widely according to the metal being rolled and the purpose for which the product is required. Heavy plates and extremely thin sheets naturally demand entirely different treatment and procedures. The rolling of broad hot strip or sheet in coils has been in continuous development since 1926, when the first mill capable of rolling strip more than 61 centimetres (24 inches) wide was built. In these mills, which are continuous and used mainly for steel, the hot metal is fed

into the first set of rolls rotating at a moderate speed. The material emerging from the first rolls is fed directly into a second set, which, since the length is increased by the passage through the first rolls, must run at considerably higher speed. A number of sets of rolls is used in series in this way, the finished product leaving the last rolls at a very high speed. Other types of rolling result in the production of sections having the form, for example, of H and I beams, channels, angles, railway rails, and other products. These are made by the action of suitably shaped grooves cut in the rolls. Successive passes in such rolls must provide for the easy flow of the metal from one shape to the next.

Cold rolling

Although steel is hot when it is reduced in size from its ingot and billet form, some of the softer nonferrous metals and alloys can be treated cold. The essential nature of cold rolling does not differ from hot rolling. The effect of cold rolling, however, is to produce hardening so that, after a time, the metal has to be annealed (heated and then cooled slowly) to soften it before further cold rolling. Cold rolling is used to reduce the thickness of section, produce a smoother, more dense surface, develop controlled mechanical properties, or obtain any one or combination of these three effects.

Drawing. Drawing is a process for forming sheet metal, tubing, and wire to obtain certain shapes and dimensions. Flat sheet metal is drawn to produce both shallow and deep shells or cups with straight walls. The distinction between shallow-drawing and deep-drawing is somewhat arbitrary. Shallow-drawing generally refers to the forming of a cup no deeper than half its diameter.

Wiredrawing is one of the oldest industrial arts and is employed to provide a slender rod or bar of metal of uniform cross-sectional shape. The shape can be round, square, hexagonal, octagonal, oval, half oval, half round, rectangular, or flat. The limits for drawn wire range from 0.025 millimetre (0.001 inch) to approximately 25 millimetres (one inch) in maximum dimension. The rod or bar of metal is pulled through a series of dies of decreasing diameter until the final shape is obtained.

Tubes are sometimes drawn through a die to reduce the inside and outside diameters, to make various shapes, and to increase the strength and hardness of the metal.

Extrusion. Extrusion is a process of shaping metal into a continuous form by forcing it through a die of appropriate shape. In forward extrusion the metal is held in a container having an opening (die) of the desired finished section at one end. The metal is forced through the die under great pressure by a ram. In backward extrusion the ram is hollow and contains the die. The container has a closed end, and pressure of the ram forces the metal in the opposite direction. The peak pressures exerted by the ram may be as low as 351 kilograms per square centimetre (5,000 pounds per square inch) for soft metals such as lead, or as high as 31,640 kilograms per square centimetre (450,000 pounds per square inch) for alloy steels. Collapsible container tubes are formed by striking a cold slug of metal, held in a container and die, a high-speed impact blow from a punch. The impact causes the metal to fill the die cavity and to flow or squirt upward out of the die around the punch. The process is called impact extrusion.

Spinning. Spinning is another process for forming sheet metal or tubes into hollow cylinders, cones, and other circular shapes by a combination of rotation and force. Tools, usually consisting of rollers mounted in a lathe, are pressed against the spinning metal with just sufficient pressure to deform the sheet into and against a mandrel (a metal bar serving as a core), made to the shape of the completed workpiece. Normally, spinning is done without applying heat.

Foundry processes. Earlier, under *Casting* it was noted that foundries throughout the world practice the art of casting metals in molds with cavities made to conform to the object wanted. The process is known as founding. Molds are generally prepared in two halves, each formed by firmly packing a mixture of sand and clay moistened with water against a wood or metal pattern that conforms in shape to the object to be cast. Although many castings are solid, there is likely to be a hollow space within the casting. To form the hollow, a core, generally of sand held to shape with a binder, is set in the lower portion, or drag, of the mold. The upper portion of the mold is known as the cope.

Cast iron

The general term cast iron includes a number of different irons, known as gray irons, pig irons (the product of the blast furnace), white cast irons, chilled cast irons, malleable cast irons, and ductile cast irons. Cast irons cover a wide range of iron-carbon-silicon alloys usually containing more than 1.7 percent and less than 4.5 percent of carbon along with varying amounts of silicon, manganese, sulfur, and phosphorus. They may have an alloy content of less than one percent and up to several percent. Special types of corrosion-resistant iron are made with greater alloy content; for example, with 11 percent to 17 percent silicon, 25 percent to 30 percent chromium, or 4 percent to 30 percent nickel.

Gray iron is by far the most widely used of the cast metals (see also below *Iron production*). The essential characteristic of gray iron is that much of the carbon is present as flake graphite, which gives it a gray-coloured fracture when broken. Its attributes have led to its widespread use for many industrial purposes, including automobile engine blocks, agricultural and machinery parts, stove and furnace parts, hollow ware, and ingot molds.

Cast steel

When steel is cast into suitable molds having cavities conforming to the shape of desired objects, it is known as cast steel and the objects as steel castings. Steel for steel castings is produced commercially by almost all the melting processes used commercially for making steel.

Five classes of commercial steel castings are produced; namely, low-carbon steels with carbon content less than 0.20 percent, medium-carbon steels with carbon content between 0.20 percent and 0.50 percent, high-carbon steels with carbon content above 0.50 percent, low-alloy steels with total alloy content less than 8 percent, and high-alloy steels with total alloy content above 8 percent. Steel castings are most often heat-treated, using the various methods that apply to wrought steel, to develop the necessary characteristics.

Nonferrous castings

The sand-casting process for production of nonferrous castings differs in no essential way from that used for production of castings made from ferrous metals. It permits the use of a large number of nonferrous alloys, providing a variety of physical and mechanical properties in both the as-cast and heat-treated conditions. In addition to casting in sand, many nonferrous castings are produced in permanent metal and in plastic molds, by die casting and by investment casting processes described earlier.

Surface treatments. Metals are frequently given surface treatments to produce a protective layer that has corrosion resistance and other important properties differing from those of the base material. Such surface layers, applied to metals of all types, often also serve to change or improve appearance. They may be applied by hot dipping, electroplating, cementation, and several other ways.

Hot dipping. Hot-dipped coatings are generally of the lower melting point metals, such as zinc, tin, lead, tin-lead alloys (called terne), and aluminum. To apply the hot-dipped coatings, careful attention must be given to preparation of the metal surface to be coated and to the duration and temperature of immersion. The molten metal must wet the base metal and must alloy with it.

Galvanizing of iron and steel

Zinc coating by hot dip (one method of galvanizing) provides an adherent, protective coating on the surfaces of iron and steel products that is effective against corrosion from exposure to the atmosphere. Zinc is more active than iron, hence tends to corrode instead of the iron; by this "sacrificial" protection, it prevents iron from corroding at minor discontinuities in the zinc coating.

Tin coatings produced by hot dipping are excellent in appearance, resist corrosion, give a surface that assists in soldering, and provide a nontoxic, protective coating that is particularly useful where the metal comes in contact with foods and beverages. Hot-dipped tin coatings are applied mainly to steel, cast iron, copper, and copper alloys.

Lead coatings, which are applied to ferrous articles by hot dipping, contain from 2 percent to 10 percent tin to assist in bonding.

Hot dipping is one of the various methods used for applying aluminum coatings to ferrous articles. Close control of liquid-metal temperature and time of immersion is important as these factors influence the thickness of the unwanted brittle intermetallic layer between the base metal and the pure aluminum overlay.

Cementation. Another process for introducing elements into, as well as onto, the outer surface of metal objects is called cementation. Essentially, the part to be cemented is heated to an elevated temperature in contact with a solid, liquid, or gaseous medium containing the cementing material. The cemented material penetrates inward from the exposed surface, forming an alloy on the surface decreasing in concentration as the distance inward increases. Among the various cementation processes used by industry are: the carburizing (impregnating with carbon) of steel; calorizing (heating the part in contact with aluminum powder) of both ferrous and nonferrous metals; and ceramic coating to protect nickel-base, cobalt-base, and vanadium-base alloys and refractory metals against oxidation at very high temperatures.

Other processes for coating metals. Metal spraying, metal cladding (covering by bonding), protective film formations, painting, and vitreous enamelling are among the other processes used in coating metals. Sprayed metal coatings are produced by use of an air blast to spray heated metal particles against the object to be coated. Such coatings are often used to resist corrosion, reclaim undersized and worn parts, and to provide hard, wear-resistant surfaces. In metal cladding, surfaces of slabs of two metals are brought carefully together for the purpose of developing a bond, sometimes by heat and pressure alone.

Anodizing of aluminum and alloys

The formation of surface films on metals is accomplished by various methods, including oxide coating of aluminum and its alloys (the so-called anodizing treatment) in a bath of sulfuric acid, or forming a phosphate protective coating on iron, steel, zinc, cadmium, aluminum, and their alloys by treating them in a dilute acid phosphate solution to convert the metal surface to an insoluble crystalline phosphate film. Vitreous (glassy) or porcelain enamels on metals are used widely in the chemical, pharmaceutical, and food industries because of their excellent resistance to chemical attack.

Powder metallurgy. The methods of powder metallurgy afford a unique way of producing and shaping metals and alloys. In its simplest form, powder metallurgy consists of: (1) pressing metal powder in a die to produce a compressed shape called a green compact, and (2) sintering (diffusion bonding by heating) the compact at an elevated temperature in a protective atmosphere. During sintering, the compact becomes consolidated and strengthened. Powder-metallurgy methods are competitive under certain conditions with other metal-forming techniques.

Powder metallurgy is usually employed for two reasons: first, because it may be the only known means of producing certain metals, alloys, or mixtures of two or more metals having certain desirable physical or mechanical properties; and second, because powder metallurgy may be the cheapest way to manufacture certain small parts, often a result of the fact that parts can be formed directly to shape, with little or no machining.

Powder manufacture. Metal powder may be manufactured by mechanical or chemical methods. The most widely used mechanical method, starting with solids, is milling by power-driven hammers (hammer milling) or by balls in a rotating container (ball milling). In starting with liquid metals, various mechanical processes that break up the metal into tiny droplets before they solidify can be employed. This method is called atomization.

Chemical methods for producing metal powders provide either for the reduction of compounds of the desired metals to powder using various agents or for the electrolysis of a liquid solution (aqueous or molten) that contains the desired metal. The latter process is similar to electroplating, but the conditions of current and temperature are adjusted to produce a porous, flaky, or brittle deposit of metal that is scraped off the cathode and broken up to obtain the powder. Nickel, copper, and iron powders are often made by this process from aqueous solutions.

To obtain specific properties in the finished product, powders are often mixed and blended before use. Excessive blending, however, especially in an overloaded blender, may work harden the powder, making it less compressible and thus more difficult to compact. To bond the powder particles into a mass solid enough to withstand the stresses in service, the powder must be heated or sintered after compressing. Sintering is done in protective atmospheres or in vacuum at temperatures below the melting points of the metals. Sometimes a liquid phase forms and assists in consolidation. In all sintering processes the dimensions of the mass of powder being sintered decrease, the amount of shrinkage depending on the composition of the powder, the amount of compression to which the powder has been subjected before sintering, and the time and temperature of sintering. This shrinkage must be allowed for in the design of parts that are to be made directly to shape. When the highest possible density is required, the compacts (parts) may be repressed and then resintered.

Shrinkage in sintering

Processes and products. Three general classes of processes are used in powder metallurgy, depending on the type of metal used in each. One class involves metals of relatively low melting point that are ductile, such as silver, copper, and iron. Another class involves the refractory metals; *e.g.*, metals with melting points usually above 1,600° C (2,912° F), which normally are brittle, such as tungsten, molybdenum, tantalum, and titanium. The third class concerns the production of materials known as cemented carbides, which contain tungsten carbide.

In preparing compacts from ductile metal powders, the pressing operation forms the metal part to the desired shape at room temperature. Pressures range from 14,000 to 70,000 kilograms per square millimetre (10 to 50 tons per square inch), compacting the powder to a density of about 60 percent to 80 percent of that of a similar mass of solid metal. The ductile metal compacts show shrinkages of about 1 percent to 5 percent in dimensions during sintering.

Powder-metallurgy procedures of the refractory metals are used chiefly to produce a bar or ingot of the refractory metal, which is then further processed by mechanical working to make wire, rod, tube, or sheet. Because most refractory metals are either too brittle or too hard to be compacted in the same manner as the ductile metals, they are compacted at room temperature with a binder, such as paraffin wax, that holds the compressed particles together until they can be presintered to drive off volatile matter and establish a weak bond. Sintering in a satisfactory atmosphere at a proper temperature then follows to develop the needed strength.

The cemented-carbide materials made by powder metallurgy are cemented together by a metal, usually cobalt, forming a product of great hardness and high compressive strength. The principal constituent is tungsten carbide powder, although carbides of other refractory metals, notably titanium and tantalum, are often used along with the tungsten carbide. The carbide powders, blended with cobalt powder ranging in amounts from 3 percent to 25 percent, are mixed with a binder such as paraffin wax and pressed to shape, followed by presintering at about 900° C (1,652° F) to evaporate the binder. The presintered material can then be machined or ground to provide the shape needed for the finished part before it is finally sintered. Certain advantages of cemented carbide as a cutting-tool material stem from its ability to cut while red-hot.

Tungsten carbide powder

Powder metallurgy is applied to iron-base compacts to produce an essentially pore-free part having increased strength and density and improved machinability. This result is accomplished by infiltrating molten metals, commonly copper and copper-base alloys, into the pores of the compact by capillary action.

Nuclear metallurgy. Nuclear metallurgy includes the production, fabrication, and application of metals of interest in nuclear engineering and the study of their properties. The study of atomic fission in 1939 led to the emergence of several relatively unknown metals and paved the way for the practical use of atomic energy. The phenomenon of fission concerns the splitting of an atomic nucleus into two or more parts by the addition of a neutron to that nu-

cleus. Fissionable materials include plutonium and an isotope of uranium, uranium-235; thorium can be converted into another fissionable isotope of uranium, uranium-233 (see ATOMS: *Nuclear fission*).

These materials are referred to as nuclear fuels. Because of their affinity for neutrons, the metals boron, hafnium, cadmium, or gadolinium are used to control the number of neutrons striking the nuclear fuel and thus shut down a reactor or control its power level of operation.

Because uranium metal is highly reactive toward water or air at elevated temperatures, protective cladding of the fuel elements is necessary. Aluminum and zirconium are commonly used for this purpose. Beryllium is desirable for use as a moderator—that is, to slow down high-energy neutrons produced by the fission process so that they become more effective in producing additional fission reactions.

Low-melting sodium and sodium-potassium alloys are used as coolants in several reactors. The liquid metal is circulated over the hot fuel elements and then to a steam generator where the heat in the liquid metal is recovered.

In addition to the usual problems associated with metals at high temperatures, there are some unique problems in nuclear metallurgy. One such is radiation damage, the failure of a material due to the intense radiation in a reactor. Uranium undergoes embrittlement and drastic changes in dimensions as a result of irradiation. Another serious problem that eventually results in a shutdown of the reactor is accumulation of fission products in the nuclear fuel. Direct removal of the fission products and economic reprocessing of the fuel constitute a major challenge to the nuclear metallurgist (see also ENERGY CONVERSION: *Nuclear reactor*).

Heat treatment. Many articles of iron, steel, aluminum, copper, magnesium, titanium, nickel, and their alloys intended for important engineering uses are subjected to some form of heat treatment for the purpose of obtaining certain desirable conditions or properties. Heat treatments are applicable to all metals, both ferrous and nonferrous. Changes in metals afforded by heat treatment involve changes in the nature, form, size, or distribution of the structural constituents, although frequently metals are heat-treated merely to decrease or relieve internal stresses.

The time–temperature cycle

Heat treatment consists of subjecting a metal to a definite time–temperature cycle. Basically, this cycle can be divided into three parts: (1) heating, (2) holding, and (3) cooling. The heating medium also is important, since it affects the rate of heating and causes or prevents chemical changes, such as oxidation, that proceed inward from the surface of the metal. Although in some instances, such as in the case of pure metal or certain alloys, the rate of cooling is likely to be unimportant, when the structure of the material being heat-treated is different at the maximum temperature from that at low temperatures, the rate of cooling is of great importance. Iron–carbon alloys that change from one crystal structure to another (gamma to alpha iron) on cooling from above the transformation range are particularly sensitive to cooling rate.

Aging, or precipitation, reaction. In heat treatment, three types of reactions are of importance industrially and form the basis for most heat-treating operations: the aging, or precipitation reaction, the allotropic transformation reaction, and the decomposition reaction. The aging reaction occurs when the solubility of a constituent in the basic material (matrix) is greater at high temperature than at low temperature. In this process, the alloys are heated to a temperature sufficiently high to cause a constituent that is more soluble in the matrix at high than low temperature to dissolve. The alloys are then quenched to room temperature at a rate sufficiently rapid to prevent apparent change in structure by precipitation of the constituent. When the quenched materials are reheated within the proper temperature range, however, precipitation does take place. By controlling the conditions for precipitation, the precipitated particles can be maintained within a critical size, and considerable variation in mechanical and other properties of the alloys can be achieved. Maximum property changes are achieved with a particular combination of time and temperature for the precipitating reaction, which determines the size and distribution of the precipitated particles, and for the compositions of the alloys. Certain copper, aluminum, and magnesium alloys are among the engineering metals whose mechanical properties are vastly improved through the mechanism of precipitation hardening.

Allotropic transformation reaction. Allotropy is the capacity of a material to exist in more than one crystallographic structure. Materials that have this capacity are termed allotropic. Allotropic transformation occurs in some metals upon heating and cooling. It serves as the basis for most heat-treating changes that occur in steel. Steel at high temperatures, for example, is composed primarily of a solid solution of carbon having a particular crystal structure known as gamma iron. The solid solution is called austenite. Gamma iron is capable of taking into solution fairly large quantities of carbon. The solid solution on cooling converts to a mechanical mixture of alpha iron (which has another crystal structure), called ferrite, and iron carbide, called cementite. Alpha iron has a crystal structure in which carbon has an extremely low solubility. As a consequence of this low solubility, the carbon in the steel separates as iron carbide.

The rate of cooling affects the transformation of austenite and largely determines the final structure of a given steel and hence its physical and mechanical properties.

Decomposition reaction. This reaction involves the decomposition of certain constituents into one or more other constituents that possess more desirable properties. Its importance is illustrated by the production of malleable cast iron from the hard, brittle cast iron known as white cast iron. The iron carbide or cementite in the white cast iron is decomposed to ferrite and a spheroidal form of graphite after a prolonged heating in the temperature range of 750° C (1,380° F) and 955° C (1,750° F). This graphitized material is considerably more ductile and more easily machined than the parent material and, hence, much more useful.

Production of malleable cast iron

(C.H.L./Ed.)

Iron production

Iron is the most useful of the metallic elements and the second most abundant in the Earth's crust, after aluminum. In its elemental form or as steel, iron has supplied civilization with most of its tools and machinery, many of its products, and the bulk of its structural elements in large-scale construction. It weighs 7,865 kilograms per cubic metre (491 pounds per cubic foot) at 20° C (68° F) and is a solid at temperatures below 1,535° C (2,795° F). Pure iron has a silvery lustre and takes a high polish. It is highly ductile at ordinary temperatures, becoming brittle only at the temperature of liquid air (−150° C [−238° F]) or at white heat just below its melting point. Except in meteorites, it is rarely found in its elemental form but occurs widely in chemical compounds in the Earth's crust.

HISTORY

Iron was probably discovered accidentally late in the Bronze Age when, after fires were built on outcroppings of red iron ore (iron oxide), metallic iron was found in the ashes of these fires. With the useful properties of iron thus realized and the red ore, known as paint rock, identified as the source, iron production began. Fires were built against banks of ore exposed to prevailing winds; later, crude rock furnaces were built, and their fires were fanned by bellows to reduce iron oxide to metallic iron. Iron tools and weapons ushered in a new age in both the Near East and the Far East. As recorded in Genesis (4:22), Tubal-cain, the eighth-generation descendant of Adam, was a forger of instruments of bronze and iron. Iron has been found at many sites in Egypt, including two pieces in the Great Pyramid of Giza, built about 2900 BC. Hittite artifacts, from 1900 to 1200 BC, include a cast-iron ring. Several iron-smelting furnaces dating from about 1300 BC have been uncovered in Israel.

Iron ore was discovered independently in China at an unknown date; before the beginning of the Christian Era the Chinese had also independently invented the blast

Table 2: Typical Percentage Composition of Iron Ores

	Swedish magnetite	red hematite (U.K.)	brown hematite		German spathic iron ore	clay ironstone	
			U.K.	Spanish		American	Yorks
Ferrous oxide (FeO)	28.42	—	—	—	55.64	36.14	39.92
Ferric oxide (Fe_2O_3)	62.06	86.50	90.05	78.80	—	1.45	3.60
Manganous oxide (MnO)	—	0.21	0.08	0.65	2.80	1.38	0.95
Alumina (Al_2O_3)	—	—	—	3.50	—	6.74	7.86
Calcium oxide (CaO)	—	2.77	0.06	—	0.92	2.70	7.44
Magnesium oxide (MgO)	1.44	1.46	0.20	—	1.77	2.17	3.82
Carbon dioxide (CO_2)	—	2.96	—	—	38.35	26.57	22.85
Ferrous sulfide (FeS)	0.07	—	—	—	—	0.10	0.11
Phosphoric anhydride (P_2O_5)	—	—	0.09	0.07	—	0.34	1.86
Silica (SiO_2)	7.60	6.55	1.07	5.55	—	17.37	7.12
Water	—	—	9.22	11.65	—	1.77	2.97
Organic matter	—	—	—	—	—	2.40	1.64

furnace and cast iron. In the West the Romans inherited ironworking techniques from the Greeks, carried them to a high degree of perfection, and spread their knowledge to northern Europe. Roman castings have been found in Czechoslovakia, France, and Great Britain. Pliny the Elder (AD 23–79) devotes two chapters of his Natural History to an account of iron production and use.

The Iron Age

Though the Iron Age is usually dated by archaeologists as beginning about 1200 BC, it was not until the European Middle Ages, with the introduction of iron cannon, that iron actually overtook copper and bronze to achieve first place among metals in use. Iron remained unknown in the Americas until the arrival of Columbus.

Before the 18th century charcoal was universally used as fuel for blast furnaces, but about 1708 Abraham Darby of Coalbrookdale, England, submitted coal to the same process used on wood to make charcoal and produced the first coke. Substitution of coke for charcoal took place quickly in coal-rich England but much more slowly in areas where forests abounded, such as North America.

The first American ironworks was built in Virginia in 1621 but was destroyed by an Indian raid a year later. The Massachusetts Bay Colony imported an English ironmaster and built an ironworks on the Saugus River, 16 kilometres (10 miles) north of Boston, in the 1640s. The operation, fed with bog iron (a hydrated iron oxide), was highly successful. Some direct casting of implements and utensils was done, but the furnace primarily turned out cast-iron bars to be worked by local forges.

In the 19th century the invention of the Bessemer and Siemens-Martin processes, making possible the large-scale production of steel, especially for structural purposes, created an enormously expanded demand for iron (see below *Steelmaking*). This demand was met by more intensive ore extraction in Great Britain, continental Europe, and elsewhere, and the discovery and exploitation of new ore fields, most notably the great Mesabi Range in the north central United States.

Principal 20th-century developments have centred on improved mining machinery and larger capacity blast furnaces; one constructed in Japan in the 1960s is capable of producing 7,000 tons of iron in a 24-hour day.

MINING AND PREPARATION FOR PROCESSING

Common ores. The most common compounds from which iron is produced are oxides, carbonates, and sulfides.

Hematite. Red hematite (Fe_2O_3), when pure, contains 70 percent iron. When mined, the content may be as low as 25 percent. It is found mainly in Sweden, Belgium, Western Australia, and the Alabama and Lake Superior regions of the United States. It may readily be distinguished from magnetic and titanium-bearing iron ores by its red streak. Hematite is sometimes mixed with sufficient magnetite to cause it to adhere to a magnet.

Limonite or brown hematite. The hydrated variety of hematite ore is called limonite, or brown hematite ($2Fe_2O_3 \cdot 3H_2O$), and contains about 60 percent iron, when pure. An important ore, it is yellowish to brown in colour. The chief constituent of bog iron ore, it is found in East and West Germany, France, Spain, Western Australia, the Soviet Union, and the United States.

Bog iron ore

Magnetite. Magnetite, a magnetic iron ore, is a natural black oxide of iron (Fe_3O_4); it is often called black sand. Magnetite is found in Sweden, Norway, the Soviet Union, and the United States.

Carbonate. Carbonate ores include siderite and spathic iron ore ($FeCO_3$) and contain 48 percent iron when pure. These ores are mostly clay ironstone and blackband ironstone found in the United Kingdom.

Taconite. This ore consists of an iron-bearing chert derived from ferrous silicate that encloses the Mesabi ores in the United States. In U.S. practice, taconite is any grade of extremely hard, lean iron ore that has its iron either in banded or in well-disseminated form, and which may be hematite or magnetite, or a combination of the two within the same ore body. When leached by hydrothermal methods, it yields a satisfactory ore, the iron remaining as hematite. Leaching is the recovery of usable constituents from heterogeneous material by dissolving in water, dilute acid, or other chemical solution.

Mining techniques. A substantial amount of iron ore is recovered by surface methods, usually known in metal mining as open-pit or opencut. For deeper lying deposits in which the ore and wall rocks are firm, open stopes—a series of descending steps—are cut; in cases of small deposits, the entire ore body may be removed from wall to wall without leaving any pillars. Where the ore body lies buried in rock, deep underground, shaft mining is employed; the vertical shaft may go down several thousand feet. Horizontal tunnels from the shaft follow the ore deposits as they are mined, with locomotives and cars on tracks used to move the ore to the shaft elevators.

Ores with a low sulfur content are more suitable for smelting. Formerly, only ores containing more than 30 percent iron could be smelted profitably, but because of various upgrading processes lower grade ores can now be used. The value of an ore deposit depends on geographic location and accessibility.

Upgrading ores (beneficiation). As the high-grade deposits became more inaccessible or exhausted and as shipping costs increased, it became necessary to separate and discard unusable materials from the iron ores at, or near, the mines. Processes broadly termed beneficiation were developed to upgrade the ore before shipment. Concentration or other preparation of ores is accomplished by leaching and drying, flotation, agglomeration, or magnetic separation. Flotation is an ore-dressing process by which finely pulverized ore is agitated in a mixture of oil and water. Constituent minerals are separated from one another by virtue of their respective abilities to be wetted by water and by their specific gravities. Agglomeration is a concentration process based on making large particles out of smaller particles. Magnetic separation, widely used to separate magnetic material from nonmagnetic, is especially effective in the beneficiation of iron ores.

Flotation

Sintering. In order to prepare or pre-reduce ore further, sintering and pelletizing were developed. The sintering process is used mainly with fine iron ore that would not be suitable for blast furnace reduction because it would pack too closely to allow counterflow of gases. The fine grains are mixed with coke and ignited, producing semifused lumps that are satisfactory for smelting. Most sintering is done at blast-furnace plants.

Pelletizing. Pelletizing on a commercial scale was begun

in the mid-1950s. In this process, iron ore concentrates are ground to powder, mixed with a binder such as bentonite and a small amount of water, and fed into a balling mill that forms the mixture into small biscuit-shaped pieces. These are baked to expel the moisture and to harden sufficiently to withstand shipment and subsequent handling up to the blast furnace. The pelletizing process is now completely automated in many places.

Again, because of the transportation costs, it has been found economical to locate the pelletizing plants at or near the mines. Self-fluxing pellets have been developed by mixing limestone or dolomite with the ore concentrate, an action that minimizes the amount of flux that must be added to the blast-furnace charge and lessens the amount of slag produced. In turn, the tonnage of iron produced by a blast furnace in a given period is increased.

An example of the progress that has been made in the beneficiation of low-grade ores located in inaccessible regions is the Savage River project in Tasmania. The iron-ore deposits there are in a rugged, almost impenetrable region in the northwestern corner of the island. The Savage River deposits, known by 1877, were reconnoitered many times but they defied economical exploitation. The ore is low grade, averaging 38 percent iron; it contains troublesome impurities, particularly titanium, and is locked in bands of rock in a remote, dense forest where average rainfall is more than 2,540 millimetres (100 inches) per year.

Finally, a study developed the feasibility of overcoming the titanium impurity by magnetic separation and by upgrading the low iron content to 67 percent for pelletizing. The remaining transportation problem, with the nearest accessible coast 85 kilometres (53 miles) away and with mountains and gorges between, was solved by a slurry pipeline. At Savage River a plant was built to crush the ore to powder, magnetically separate the iron-rich particles, and mix them into a slurry with a 40 percent water content that could be pumped to the coast through a 3.5-centimetre (nine-inch) pipeline. Another plant was built on the coast to dry the ore concentrate and bake it into pellets. Finally, a jetty 1.6 kilometres (one mile) in length and conveyor were constructed to load ships. The operation produced a high of about 2,000,000 tons of pellets during the 1970s.

Pipeline slurry

ORE PROCESSING

Blast-furnace construction and operation. *Structural details.* A blast furnace is constructed from a steel shell lined with refractory brick that resists high temperatures. Spaced around the lower part of the furnace are water-cooled tuyeres (nozzles), usually copper castings, to introduce the air blast. The charge, called the burden, which is dumped into the top of the furnace, consists of alternate layers of charcoal or coke, raw ore, sinter or pellets, and flux. Iron-making in a blast furnace is a continuous process, furnaces usually being shut down only for major repairs, such as relining with refractory. The modern furnace is some 24 to 30 metres (80 to 100 feet) high, with a hearth diameter of about 9 metres (30 feet).

At the bottom of the furnace is the hearth, built of graphite or refractory brick, which collects the molten iron as it drips down from the reduced ore. Progressing upward are the tuyere zone; the bosh zone (the region where iron and slag begin to melt), extending upward to the plane of maximum diameter; the mantle, or region immediately above, where iron and slag first begin to melt; and, finally, the stack. The stack houses the bells (cone-shaped inserts) that regulate the downward flow of the burden. From this section, the blast-furnace gas, produced by the partial combustion of the coke during the ore reduction, is piped off. Such gas is combustible, though of comparatively low heat value, and is used as a fuel for preheating the air blast.

To prevent spalling (chipping), the refractory brick furnace lining is cooled by a flow of cold water through hollow copper castings inserted at regular intervals.

Stoves

Adjacent to the furnace are stoves to preheat the air blast; these have steel shells, lined with refractory brickwork, and are fired with the blast-furnace gas and coke-oven gas from the coke plant. The hot blast is piped into the furnace through the tuyeres.

Coke. A coke-making plant, usually included as part of a blast-furnace plant, produces coke strong enough to withstand the burden in the furnace and of sufficient size and porosity to leave openings for the upward passage of the gases. Because of the scarcity of bituminous coal suitable for making good coke, various methods have been tried for reducing the tonnage of coke required per ton of iron. The most common is to introduce a supplementary fuel through the tuyeres, which may be gas, fuel oil, or finely pulverized coal. In addition, the hot-air blast may be enriched by the addition of oxygen.

Operation. The molten iron is tapped from the blast furnace at regular intervals and the accumulation of slag is drawn off. The iron is either sent in large ladles directly to steelmaking furnaces or is pigged (cast into ingots) and stored for later shipment to foundries.

Direct-reduction processes. Because of the high capital investment necessary for blast-furnace construction, many direct-reduction iron-making processes have been tried. None has been totally successful in producing economically the tonnage levels of a modern blast furnace, but research is continuing under the added stimulus of problems of air and water pollution and the shortage of metallurgical coke for furnace operation. Two basic approaches offer promise of delivery to the foundry of hot iron of the proper composition and temperature for direct casting of gray, ductile, and malleable iron: these are the rotary kiln and the travelling grate.

Rotary kiln. A rotary kiln and solid-carbon fuel can prereduce ores that are in lump or pellet form. Rich ore lumps, 5–20 millimetres (0.2–0.8 inch) in size, or iron pellets 10–15 millimetres (0.4–0.6 inch) are fed into the rotary kiln, along with pieces of lump coal and flux about 10 millimetres in size. The temperature of the kiln is maintained at 1,050° to 1,100° C (1,922° to 2,012° F). Since the ore is prereduced, iron particles are produced that are approximately the same size as the charged ore or pellet particles. The product of this method is a molten high-carbon metal.

One major disadvantage of this process is that fine ores cannot be used. Other problems include the frequent shutdowns needed for maintenance, the degradation of the charge in the kiln, and the difficulty in maintaining carbon-to-oxygen ratio control.

Problem of ore size

Travelling grate. In the travelling-grate machine used for prereduction of the iron ore, carefully sized raw materials are not required. Fine ore, coal, and flux are proportioned and ground together. After filtering to remove water, the filter cake is formed into pellets about 25 millimetres (one inch) in diameter, which are fed continuously onto the travelling grate machine. After drying, igniting, and carbonizing, the pellets are discharged at about 982° C (1,800° F) into a closed submerged electric-arc furnace and smelted to a controlled grade of high-carbon foundry iron.

In the commercial plant the composite pellets are ignited and carbonized, prereduced, and heated on the travelling grate in less than 20 minutes. Metallurgical control is obtained primarily by varying the proportioning of the raw materials. A major advantage of this process is its ability to use ore of any size and type. Another advantage is that the operation of the travelling-grate machine can be interrupted, whereas a kiln, to be economical, must be kept in constant operation.

Foundries. Throughout history, iron castings have been made by direct casting from smelting furnaces. During the past century or two, the direct casting of iron has given way to indirect methods in order to achieve more precise control.

Foundries remelt the iron either in cupolas or electric furnaces for the manufacture of castings. The most common product of the cupola is gray iron (an iron-carbon alloy containing two to four percent carbon), though the carbon content can be regulated and various metals, such as chromium, nickel, or magnesium, can be added in small quantities to yield mechanical and physical properties needed for specific uses.

The jobbing foundry. In a jobbing foundry that makes small numbers of castings of many different designs, pat-

Table 3: Simple Visual Tests for Identifying Irons

	white cast iron	gray cast iron	malleable iron	wrought iron
Fracture	very fine, silvery white, silky crystalline formation	dark gray	dark gray	bright gray
Unfinished surface	evidence of mold; dull gray	very dull gray	dull gray	light gray, smooth
Newly machined surface	rarely machined	fairly smooth, light gray	smooth surface, light gray	very smooth surface, light gray
Appearance of chips	small broken fragments	partially broken chips but smooth groove	chips do not break short	smooth edges where cut
Size of chips	—	⅛ in.	¼–⅜ in. (0.6–1.0 cm)	continuous
Speed of melting	moderate	moderate	moderate	fast
Colour change while heating	dull red before melting	dull red before melting	red before melting	bright red before melting
Appearance of molten metal	fluid and watery; reddish white	fluid and watery; reddish white	fluid and watery; straw colour	liquid; straw colour

Casting molds

terns are usually made of wood. These are placed in boxes called flasks, and damp sand is rammed around them to make the molds. The surface must be such that the pattern can be withdrawn without damage to the mold. An opening, or gate, is cut in the mold for entry of the molten iron. Because of the shrinkage of the iron as it solidifies, other openings called risers often are left. These fill with molten iron that feeds down into the casting and prevents voids. When the castings are cooled, they are shaken out of the molds, and the gates and risers are cut off. The molding process, long a handcraft, is becoming more and more mechanized.

The production foundry. Production foundries are those that make hundreds of thousands of castings from the same pattern. Today these are often completely automated, and many of the processes such as molding-sand preparation, cupola charging, and iron pouring can be computer-controlled. Many castings are now made with permanent molds of iron or steel that can be used repeatedly, whereas a sand mold is destroyed when the casting is withdrawn from it.

Modern versions of the lost-wax (cire perdue) process, which consisted of making the patterns of wax, instead of wood or metal, have been developed. As noted earlier, this process, now known as investment casting, appears to have been discovered by the ancient Egyptians. Many fine metalworks made by this process still survive. The wax patterns do not have to be withdrawn from the molds, since they melt and vaporize on contact with the molten metal. Hence, many complicated shapes can be cast. In addition to wax, the patterns are today being made from various volatilizable plastics.

THE METAL AND ITS ALLOYS

Wrought iron. Wrought iron is a form of iron containing less than 0.3 percent and usually less than 0.1 percent carbon, and carrying also 1 or 2 percent of slag mechanically mixed with it. Wrought iron finds product applications because of its inherent properties of ductility, ease of working, and welding, and superior resistance to service conditions involving corrosion, shock, and fatigue.

Catalan forge. The Catalan forge, which originated about 1300 in Catalonia, Spain, represented a major advance in the manufacture of wrought iron directly from ore. This forge has a siliceous bottom lined with charcoal, with a tuyere inclining downward. The front of the forge is piled with ore and the back with charcoal, the whole being covered with fine mixed ore and charcoal dust, moistened with water (Figure 10).

Puddling. In the 18th century, Henry Cort, in England, invented what is known as the puddling process, still in use. Pig iron is melted in a furnace with iron ore added. The operation is carried out in an oxidizing atmosphere in which silicon, manganese, sulfur, and part of the phosphorus are oxidized and removed in the slag. The temperature is then raised, promoting interaction between the iron oxide and carbon. Carbon monoxide is evolved, causing violent boiling of the charge, which is then worked with a rabble, or rake. As the iron becomes purer, it no longer remains liquid at the working temperature and the entire mass becomes pasty and is separated into portions and hammered and rolled into shapes.

Working with rabble

Cast iron. Cast iron is a generic term describing a family of iron alloys containing 1.8 to 4.5 percent carbon. These alloys are usually made into specified shapes for direct use or for further processing by machining or heat treating.

Gray iron. Gray iron covers a wide range of basically iron–carbon alloys containing 2 to 4 percent carbon, in which nearly all the carbon not included in pearlite (an aggregate of ferrite, which is a salt of a ferric hydroxide acting in its capacity of an acid, and cementite, which is iron carbide, Fe_3C) is present as graphitic carbon. Gray iron is produced from pig iron and scrap, or from scrap iron alone. In modern practice, it is customarily melted in a hot-blast cupola, which is supplied with heated air blast as in blast-furnace practice. Various additives to enhance its properties may be introduced into the cupola, after the iron is molten, or into the ladle, before pouring it into molds.

Because of the presence of graphite, the fracture has a gray colour. Gray irons are the most widely used of the cast irons because of their relatively low cost, ease of melting, and excellent casting characteristics.

The engineering properties of gray cast iron has taken on even greater significance since the early 1960s. From simply being a material that was inexpensive, easy to cast, and readily machinable, gray iron has become an important engineering material whose composition, structure, and physical and mechanical properties can be carefully controlled to obtain the best combination of strength and service. While certain mechanical properties, such as tensile strength, improve to a moderate extent by natural aging, heat treatment—followed by quenching and tempering—accomplishes much better results in a few hours. For applications involving wear, such as in machine beds, the wearing surface is commonly hardened by flame or induction heating.

From H.E. McGannon (ed.), *Making, Shaping and Treating of Steel*, 9th ed., copyright 1971 by United States Steel Corporation

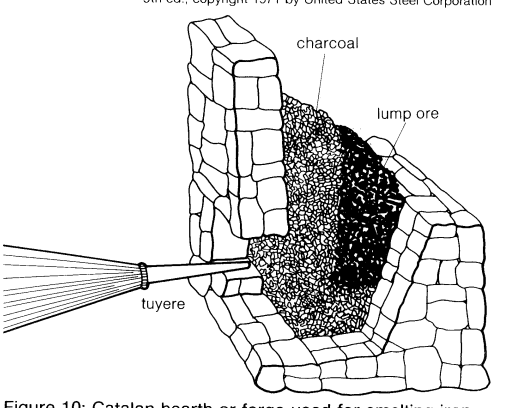


Figure 10: Catalan hearth or forge used for smelting iron ore until relatively recent times. The method of charging fuel and ore and the approximate position of the nozzle supplied with air by a bellows are shown (see text).

Technical developments in founding have permitted considerable reduction in machining allowances in finished castings, and they are produced with a much greater prospect of complete soundness.

White iron. White iron, also named for its fracture colour, contains practically all its carbon in combined forms as cementite (iron carbide, Fe_3C) and pearlite. Low silicon content is usually required, and such stabilizing elements as chromium are added when white iron is to be used in the "as cast" condition. It is frequently alloyed with nickel and chromium for applications involving wear resistance. It is extremely hard, difficult to machine, and not malleable at any temperature.

Ductile, nodular, or spheroidal graphite cast irons. In these variously named cast irons the graphite is transformed from the long, thin flakes typical of gray iron into discrete spheroids. Spheroidal graphite structure can be produced by adding either cerium or magnesium. In both cases, the addition of the nodularizing agent is followed by the addition of ferrosilicon. Spheroidal graphite iron possesses the advantages of superior strength and impact resistance.

Malleable cast iron. Malleable cast iron is a cast iron made by a prolonged anneal of white cast iron in which decarburization (removal of carbon) or graphitization (conversion into graphite), or both, takes place to eliminate some, or all, of the cementite. The graphite is in the form of temper carbon (clusters of finely divided graphite). If decarburization is the predominant reaction, the product, when fractured, shows a light colour and is called whiteheart malleable; otherwise, the fracture is dark, and the iron is called blackheart malleable.

Alloys. Iron alloys can be grouped by their commercial applications as magnetic alloys, electrical-resistance alloys, heat-resisting alloys, corrosion-resisting alloys, and thermal-expansion alloys.

Magnetic alloys. Magnetic alloys may be classified in accordance with their magnetic properties as soft, or nonretentive of magnetism, and hard, or retentive. The soft alloys range from cast iron, the most retentive, to the iron-nickel alloys, the least retentive. These alloys find widest application in communications and electric-power equipment. Some possess certain properties that make them peculiarly suited for specific applications. The magnetic properties of the alloy containing about 70 percent iron and 30 percent nickel, for example, vary with the temperature. This alloy is used in shunts in electrical and other instruments to compensate for ambient temperature variations. Although the soft alloys generally are used as solid metals, they are sometimes employed in a finely divided condition for inductance cores in telephone and radio-frequency circuits where they minimize eddy-current losses.

Hard alloys

The hard, or retentive, alloys, which remain magnetized after application of a magnetic field, are used in applications in which a constant magnetic field is required but in which it is impractical to establish an electromagnetic field, as in electrical measuring instruments. Hard iron alloys may be made from cast or wrought iron or prepared from powder by sintering.

Electrical-resistance alloys. The electrical-resistance iron alloys are best known for their use as heating elements in radiant heaters, electric ranges, and toasters. Of a number of iron-containing alloys that find wide application in this field, the most popular contain, in addition to the iron, 60 percent nickel and 16 percent chromium; 38 percent nickel and 19 percent chromium; and 22 percent chromium, 5 percent aluminum, and 1 percent cobalt.

High-temperature alloys. In the choice of iron alloys for high-temperature service, the material not only must retain substantially its original identity in the atmosphere and at the temperature of operation, but it must also possess sufficient strength and shock resistance to meet service requirements. The high-temperature alloys are broadly classified as iron-base, cobalt-base, or nickel-base, depending upon the predominant element, but all contain some iron. They are mainly used for turbine blades and components in exhaust superchargers in aircraft engines.

Corrosion-resisting alloys. The term corrosion resisting, as generally used for alloys, means the resistance to corrosion attack by liquids and gases other than air or oxygen. The intensity of such attack is dependent on several factors, such as the alloy concerned, the particular corrosion medium under consideration, its concentration, and the oxidizing or reducing conditions. A large number of the alloys considered corrosion resistant contain iron in amounts varying from 2 percent to more than 50 percent. Nickel and chromium are common constituents, together with additions of molybdenum, cobalt, copper, tungsten, and silicon, or combinations of these.

Because of the several factors involved, no single alloy so far has been found that resists adequately all forms of corrosion. An iron alloy with 15 percent silicon, for example, is widely used in vessels for handling many alkalies and acids but is quite unsatisfactory in resisting hydrochloric acid. With the addition of about 4 percent molybdenum, however, the same alloy resists this acid. Many corrosion-resistant alloys are also heat resistant, and the borderline between the two groups is not clearly drawn.

Thermal-expansion alloys. A group of iron-containing alloys has uniquely low thermal-expansion properties. Usually containing a high percentage of nickel, these alloys are used in rods and tapes for geodetic work, compensating pendulums, balance wheels in clocks and watches, and in delicate instruments.

Bimetals

In these so-called bimetals, temperature change causes movement, which is the basis of applications in thermal switches and temperature regulators. In one example, the low-expansion alloy contains approximately 45 percent iron, 52 percent nickel, with additives of titanium and aluminum. It is paired with a high-expansion alloy that contains 49.5 percent iron, 42 percent nickel, 5.5 percent chromium, and again additives of titanium and aluminum.

Thermal-expansion alloys are used in the manufacture of various types of vacuum tubes in which seals must be made between glass and metal. An alloy for this application must have an expansion curve similar to that of glass and must exhibit a high degree of resistance to thermal shock. An alloy possessing these properties is 54 percent iron, 29 percent nickel, and 17 percent cobalt.

IRON POWDER AND COMPOUNDS

Powdered iron. In powder metallurgy, iron powder, pulverized to talcum-powder fineness, is used pure or mixed with other metals or nonmetals to form alloys. The process permits alloy proportions that would not be possible using conventional means.

The iron or iron mixture is combined with a binder and pressed in dies or metal molds into desired shapes and sintered. The resulting alloy can be rolled into sheet or drawn into wire. Most importantly, complicated shapes can be formed to very accurate dimensions. Such shapes have good surface characteristics and require little, if any, machining. The physical and mechanical properties of powdered-iron products, in most cases, equal those of castings.

Iron compounds. In addition to its alloys, iron forms many useful compounds. Some of these are among the naturally occurring compounds, such as ferrous phosphate, which occurs as the ore known as vivianite, and ferrous carbonate, which occurs naturally as spathic iron. The former is one of a number of compounds useful medically. Ferric ammonium citrate is used as an appetite stimulator. Ferrous gluconate, ferrous sulfate, and ferric pyrophosphate are among compounds used to treat anemia. Ferric salts, which act as coagulants, are applied to wounds to promote healing.

Many iron compounds are used industrially; for example, in abrasives, pigments, and magnetic tapes. Ferric carbide and ferrous carbonate have been widely used in the steel industry; ferrous carbonate was for some time a principal constituent of tool steels.

Iron compounds are also widely used in agriculture. For example, ferrous sulfate is applied in spray to acid-loving plants, and various other compounds are used as fungicides.

(H.A.W.)

Steelmaking

As noted above, steel is iron that contains carbon as an essential constituent. The carbon content can range from approximately 0.03 percent to 1.5 percent and in a few cases may go as high as 2.25 percent. Broadly, there are three commercial classes of steels: carbon steels, low-alloy steels, and high-alloy steels. In addition to carbon, the carbon steels contain up to 1.65 percent of manganese and possibly small quantities of silicon, aluminum, copper, and other elements. The low-alloy steels also include a total of from 1 to 5 percent of common alloying elements such as nickel, chromium, molybdenum, tungsten, titanium, niobium (columbium), and often other elements. The high-alloy steels contain over 5 percent of at least one alloying element, principally chromium, nickel, manganese, molybdenum, or tungsten. The stainless steels constitute a major class of the high-alloy steels. Usually a steel with more than 5 percent chromium is called a stainless steel. By common practice, an alloy containing less than 50 percent iron is not considered to be a steel. Carbon steels are used extensively for automobile bodies, appliances, machinery, ships, containers, and the structure of buildings. The low-alloy steels are used most often in applications in which special properties are needed, but cost is of considerable importance. Examples are forged products for highly stressed parts of machines, aircraft landing gear, shafts, hand tools, gears, and high-strength structural steels. The high-alloy steels are used for their unusual properties: typical are the stainless steels, used when resistance to corrosion and oxidation and good appearance are required; *i.e.,* as parts of jet engines, chemical equipment, tableware and cooking utensils, and cutting tools.

Three classes of steels

Great quantities of iron ore, coal, fuel oil, electricity, natural gas, limestone, oxygen, and air are needed as raw materials by the industry. In the United States an average of 794 kilograms (1,750 pounds) of iron ore, 476 kilograms (1,050 pounds) of steel scrap, 726 kilograms (1,600 pounds) of coal, 363 kilograms (800 pounds) of limestone, 150 kilowatt-hours of electrical energy, and 60,565 litres (16,000 gallons) of water are required to produce one ton of raw steel.

The tremendous variety possible in the properties of steels arises principally because of the effects of atoms of elements such as manganese, chromium, nickel, molybdenum, and vanadium substituting for atoms of iron in the crystal lattice and by elements such as carbon and nitrogen, which can lodge interstitially in holes between the atoms of iron. The importance of these interstitial elements can be seen from the fact that in many commercial steels the carbon content must be controlled within 0.01 percent and nitrogen within 0.001 percent.

The general course of the chemical changes that iron undergoes in the conventional conversion of iron ore to steel is as follows. Oxygen is removed from the ore by reactions with carbon, carbon monoxide, and hydrogen. The resulting metallic iron is melted, and in this state it will dissolve carbon and other impurities. Most of these impurities are removed from the liquid iron in the steelmaking furnace by reacting the metal with oxygen; the process is controlled, however, to prevent the iron from being oxidized again to the oxide. After the liquid metal has been refined, the necessary alloying elements are added, and the steel is solidified. In the subsequent rolling, forging, pressing, and processing of the solid steel, care is again taken to prevent undue oxidation of the steel and loss of carbon from the surface.

MASS PRODUCTION OF STEEL

Late 19th and early 20th centuries. The rapidly rising demand for iron and steel in the middle of the 19th century resulted from the use of these materials in the construction of railroads, ships, buildings, and machinery, required by the rapid industrialization of western Europe, Great Britain, and the United States. War and advancing military technology also demanded steel.

Improvements in machinery for working and rolling wrought iron and steel made it possible to produce a wide variety of useful products. The steam hammer that came into general use between 1850 and 1860 had a steam-driven piston that actuated the hammer directly, permitting rapid operation with precise control. Two rolling achievements of note were the production of an iron rail 19 metres (62 feet) long in 1851 at Port Talbot, Wales, and the installation of a mill at Dowlais, Wales, capable of rolling 46-centimetre- (18-inch-) deep H beams 15 metres (50 feet) long and rails up to 37 metres (120 feet) long. The universal mill with both vertical and horizontal rolls, which was invented in England in 1829, was first used in 1848. The first I beams were rolled in France in 1840. The so-called three-high mill, which first appeared in England in the early 1820s and came into general use after 1850, when it became possible to build bearings, gears, and drive systems of satisfactory quality and precision, has three rolls placed one above the other, and a bar or sheet can be passed back and forth through the mill. The rolling can be accomplished in both directions without the need for reversing the mill. The advantage of this mill over the two-high mill (see Figure 11) was that the piece did not have to be passed back over the mill by hand after each rolling operation.

From A.K. Osborne, *An Encyclopaedia of the Iron and Steel Industry;* Technical Press, London

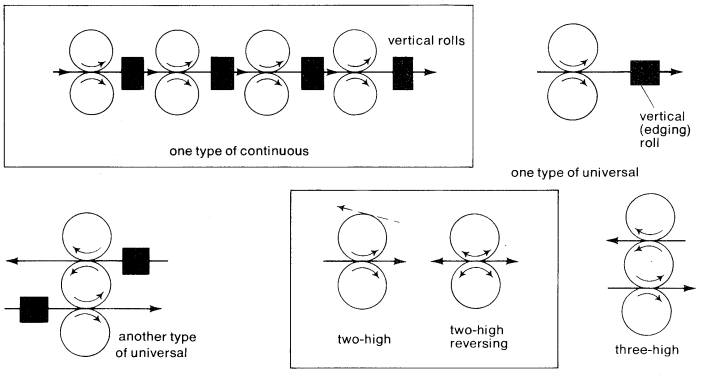


Figure 11: Representative roll arrangements.

The continuous mill for rolling rods for wire first appeared in England in 1861. It has several consecutive two-high stands of rolls that are alternately vertical and horizontal. A bar passing through the mill is in several stands at once, and its speed increases as its cross section is reduced. The multiple-stand continuous rolling mill for producing large coils of steel strip from hot slabs of low-carbon steel was first used in 1923 in the United States. The multiple-stand continuous mill for rolling these coils cold at high speed into thin sheets began to replace old hand-operated mills in the 1930s.

Continuous rolling mill

To reduce waste as well as the amount of rolling required and the size of mill needed, equipment and methods have been developed for casting steel continuously into bars, rods, and slabs. Employed successfully since 1950, these strand-casting methods have been used to make lengths of steel with round or square cross sections called blooms.

First techniques of mass production. By 1850, relatively large quantities of liquid pig iron were being produced by the blast furnace, but the existing processes for making wrought iron and steel, principally the puddling furnace and the crucible process, could produce metal in batches of not more than a few hundred kilograms each. In addition, these operations involved heavy manual labour and the exposure of workers to high temperatures. This situation and the growing need for large tonnage of wrought iron and steel stimulated the invention of methods for the mass production of steel. Work on a pneumatic means for producing steel from liquid impure iron by blowing a blast of cold air up through the molten mass was begun by William Kelly of Kentucky, in 1846. In England, Henry Bessemer's efforts along the same lines led to the patenting of the Bessemer process in 1856. At first, Bessemer's and Kelly's process could not be used to make good steel from common English and American liquid crude iron. It was soon discovered, however, that the addition of manganese to the blown metal from the converter would make the steel ingots dense and easily workable by

rolling and forging without cracking. In 1856 William and Friedrich Siemens, German-born scientists who also lived in England, invented a regenerative system for preheating the air supplied to a furnace so that flames of very high temperature could be produced. A chamber filled with open brickwork is heated for a time by the hot gases from the furnace. Then, for a period subsequently, the air supply for combustion is heated by passing it through this chamber. In this way the chamber is alternately heated and cooled in cycles; normally, a furnace is equipped with two regenerator chambers. In 1858 a patent was issued to the Siemens brothers for the use of regenerators with a shallow-hearthed furnace for melting and refining steel. The Siemens furnace used a grate for burning solid fuel, however, and the ductwork necessary to convey the gases from one end of the furnace to the regenerator chambers and the hot air to the grate at the other end was complex and inefficient. Pierre and Émile Martin in France in 1864 built a furnace that was fired by gas and placed a set of two Siemens regenerator chambers at each end of the furnace; one heated air, and the other heated the fuel gas. By firing the furnace with the gas alternately at one end and then the other, each set of chambers alternately was heated by waste gases and then heated the incoming air and fuel gas. As a consequence, it was possible to operate with a flame whose temperature was sufficiently hot to melt steel scrap. This furnace became known as the Siemens–Martin furnace, or, more commonly, as the open-hearth furnace. The regenerative system was also applied to the crucible process.

Open-hearth furnace

Native ores in Europe and in England contain considerable phosphorus, and pig iron produced from these ores, upward of 1 percent phosphorus, is unsuitable for steelmaking by Bessemer's process or the Siemens–Martin furnace. In 1876 S.G. Thomas and Percy Gilchrist, also in England, invented a modification of Bessemer's process by which satisfactory steel could be made from high-phosphorus pig iron. Instead of silica, a lining of lime was used; this permitted operation with a slag high in lime that could absorb phosphorus from the metal. Very shortly after this discovery, open-hearth furnaces were also built with their hearths lined with lime or magnesia instead of silica. Thus, the basic open-hearth furnace came into being.

The new methods for making steel were utilized rapidly in the commercial production of steel on the European continent, in Great Britain, and in the United States. The first commercial production of Bessemer steel in the United States was in Wyandotte, Michigan, in 1864, and in 1868 the first commercial open-hearth operation was started in Trenton, New Jersey. These processes produced steel in batches, or heats, that ranged in size from one or two to 15 tons or more. Thus, steel products in the form of rails, structural members, plates, and castings could be made cheaply and in relatively large quantities. The expansion of railroads in the United States following the Civil War depended largely on the availability of steel for rails. In the 1880s and 1890s buildings were built using steel members as the skeletons. Though the electric motor came into use for many purposes between 1880 and 1890, it did not completely replace the steam engine in the large rolling mills for many decades.

Use of the electric dynamo in steelmaking

As soon as the dynamo became a practical reality, chemists and metallurgists put electricity to work as an energy source for steelmaking. William Siemens is credited with having first used electric power to make steel (1878). The heat source was the electric arc. P.-L.-T. Héroult in France in 1899 operated the first commercial electric furnace for the production of steel. His design was the forerunner of the modern high-powered arc furnace. This type of furnace was enlarged in an installation in the United States in 1904 with a capacity of 3,600 kilograms (four tons) per charge. The electric-arc furnace is well suited to the production of high-quality special and alloy steels, and it rapidly replaced the crucible process for making these steels.

Development of fuels. As has been noted, fuels have been of great importance in the production of wrought iron, pig iron, and steel. Charcoal was virtually the only fuel employed from antiquity until early in the 16th century. Coal appears to have been used only occasionally before 1000, and heap coking of coals was practiced in the Middle Ages. The beehive coke oven was first used in 1620, but heap coking of coal for metallurgical purposes persisted. Attempts to develop ovens from which by-products could be obtained from the materials volatilized from the coal began in 1781. The modern retort oven was developed in 1861 and has been the chief means for producing coke for metallurgical use since about 1910. The beehive ovens continued to be operated for many years, particularly in times of high demand for coke. An important step in the development of fuels for steelmaking involved the use of gas as a fuel in the open hearth. This permitted all of the combustion to occur over the charge, a condition that could not be achieved with the combustion of solid fuel in a grate. Between 1880 and 1900, heavy oils, atomized with steam or air under pressure to form a long, nonturbulent flame over the hearth, also came to be used as fuels in the open-hearth furnace. The Bessemer converters and the basic oxygen process are said to be autogenous, because all the heat required is supplied by oxidation of elements in the charge, and no fuel is required. An important advantage of the electric furnace is that there is no possible contamination of the charge by impurities from the source of the energy at the generating station, such as can result from fossil fuels or nuclear power. Supplies of low-sulfur coal and fuel oils are being depleted rapidly, however, and the steel industry will face greater problems of contamination of steel with sulfur from fuels in the 1980s and beyond.

Use of gas in open hearth

New steels. The growth of steel production in the world was rapid between 1870 and 1950. By 1910 the use of the acid Bessemer process had begun to decline in the face of competition of the basic open-hearth furnace, and the acid open hearth came to be used for the production of special steels for such items as forgings and bearings. The basic Bessemer process was used extensively in Europe for refining pig iron produced from certain low-grade ores. The electric-arc furnace became the principal method for producing alloy steels by 1920, and the use of the crucible process declined rapidly thereafter. Wrought iron continued to be produced in appreciable quantities, but there was a gradual decline in output in the face of competition from steel. The Aston process was developed in the United States for producing a "synthetic" wrought iron starting with low-carbon iron from an acid Bessemer converter and was used commercially until 1971.

The principal changes in the methods of producing steel in the early part of the 20th century were the rapid increase in the capacity and number of producing units of the iron blast furnace and of the basic open-hearth furnace. In 1950 a large blast furnace produced 1,500 to 1,800 tons of iron a day; and a large open hearth, 400 tons of steel in a heat time of eight to 10 hours. The types of steels being produced expanded greatly, and the highly alloyed steels, such as the stainless steels (1915), came into use.

Post-World War II period. The enormous demand for steel created by World War II, the destruction of steel-producing facilities in a number of countries as a result of the conflict, and the economic changes in the immediate postwar era had a profound effect on the steel industry in most countries. The mines that supplied high-grade iron ores that could be used in the blast furnace with little or no processing were seriously depleted and could no longer supply the rapidly rising requirements for iron and steel in the postwar years. In consequence, it became necessary to process low-grade ores to obtain high-grade concentrates. Because these concentrates were very fine, they were agglomerated to obtain a granular product of uniform size.

The development of supplies of iron ores of carefully controlled size and composition has resulted in a very large increase in the output of a blast furnace since 1950. A furnace that produced 1,500 to 1,700 tons of iron a day in 1950 using an unprepared burden, for example, is expected to produce 3,500 to 4,000 tons a day today with an ore supply of good quality. Other improvements have aided in making this increase in productivity possible: increased pressure and temperature of the air blast, injec-

tion of fuel with the air, enrichment of the air blast with oxygen. Many mechanical improvements in equipment have also helped to increase the output. The availability of high-grade raw materials have made it practical to build very large furnaces with capacities of 10,000 tons of iron a day. Several of these furnaces have been constructed in Japan since 1970.

Use of pure oxygen. Metallurgists have dreamed since Bessemer's time of using pure oxygen instead of air to refine liquid pig iron into steel. Although the fractionation of air to produce pure oxygen was developed in 1889, oxygen remained too costly for many years for it to be used as a primary oxidant in the manufacture of iron and steel. The first large-scale use of oxygen occurred in Germany immediately prior to World War II, following the development of a process for extracting oxygen from air in 1937. Immediately before and after the war, experimentation of the use of oxygen in steelmaking began in Germany, Switzerland, England, Sweden, and the United States. In 1949 a new process came into being as a result of experimental work in Austria. The furnace for the process is a roughly pear-shaped steel vessel lined with a basic refractory (Figure 12).

"Oxygen and Steelmaking Processes; A Comparison of Open Hearth, LD, Kaldo and Rotor Processes," by E.F. Kurzinski, *AISE Yearly Proceedings, 1960*, p. 99

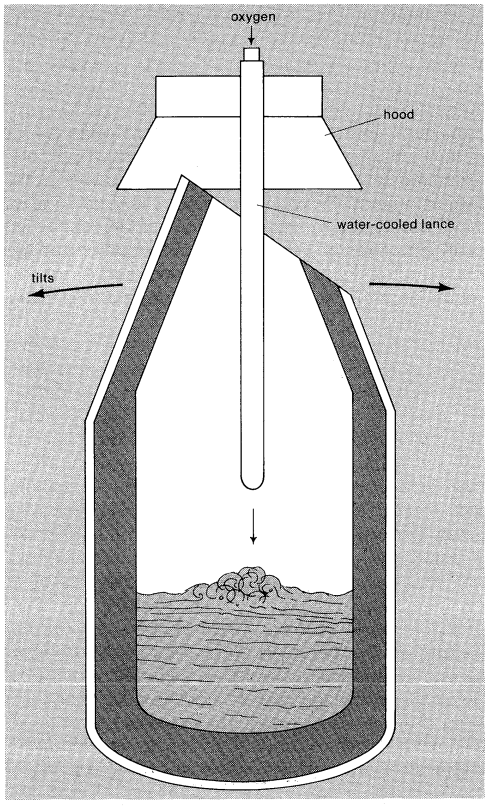


Figure 12: Elements in the L–D, or basic oxygen, furnace (see text).

The mouth of the vessel is open and faces upward. The liquid metallic charge placed in the vessel is oxidized by a supersonic jet of oxygen from a water-cooled lance that directs the oxygen vertically downward onto the charge of scrap and liquid hot metal. Depending on where it is located, this process is known as the L–D process (from the towns of Linz and Donawitz, in Austria), the basic oxygen furnace, the basic oxygen process, or the basic oxygen steel process. The first commercial oxygen furnaces had a capacity of 38 tons, but successive installations became larger, and furnaces with capacities of over 275 tons per heat have been in operation in Italy, the United States, and Germany since 1964. A plant with one furnace of this size in regular operation can produce more than 2,500,000 tons of steel a year.

L–D process

The basic oxygen furnace produces carbon steels at least the equivalent in quality to basic open-hearth steel. Many new installations of this process have been made throughout the world, and it is virtually the sole method used in making carbon steels in Japan. In 1970 it surpassed the basic open hearth in the production of raw steel in the United States, and it has been installed in a number of the major steel plants in the Soviet Union. It is also being operated extensively in Europe as a result of the greater use of low-phosphorus ores from overseas. Several modifications of the basic oxygen process are used to refine high-phosphorus pig iron on the European continent and in the United Kingdom.

Direct reduction. For several hundred years, metallurgists have argued the relative merits of the indirect and the direct processes for producing steel. The indirect process consists of producing a crude, relatively impure iron in the blast furnace and then refining it in the liquid form in one of the steelmaking processes, such as Bessemer, basic open hearth, or basic oxygen. The indirect process is commonly employed in virtually all modern steel plants throughout the world. Its detractors point, however, to the complicated processing sequence and the enormous capital investment for the facilities. In the direct process, relatively pure iron ore is reduced to the solid-metallic state by hydrogen, carbon, or carbon monoxide. The solid metal is then melted, alloyed, and cast into ingots or strand-cast bars and slabs for subsequent processing into finished products. If the reduced solid iron is pure enough, it can be compacted and formed into a bar of iron without the metal having been melted. The problems, however, of finding large quantities of iron ore of very high purity and of controlling the composition of the product have prevented large-scale use of this type of process. In the past century a large number of schemes have been proposed and patented for the direct production of steel. Many of them have been sound technically, but, except in unusual circumstances, none has been able to compete economically with the indirect route of the blast furnace and steelmaking furnace.

Interest in direct methods revived between 1965 and 1970 for a number of reasons. Relatively large supplies of high-quality ores suitable for direct processing have become available commercially, and the modern high-powered electric-arc furnace is well suited to economical melting of solid reduced ore. Moreover, direct processing is suited to small-scale production of steel in the range of 500 to 2,000 tons a day away from the large-scale producing centres. It remains to be seen whether the improved technology for direct reduction will alter the century-old economic advantage of the indirect processing.

Since World War II the production of steel has spread to many countries of the world. The changing technology by which steel can be produced economically from scrap on a relatively modest scale of a few hundred tons a year has favoured this expansion. Also, a country often desires a steel industry as a mark of its industrialization and as a means of reducing its expenditures abroad.

Alloy steels. The use of alloying elements to obtain special properties in steels began on a commercial basis in 1861, when a patent was issued to cover the use of tungsten in tool steels. Chromium-bearing steels began to appear in the United States, Britain, and France in the 1870s. Chromium was used in tool steels; its first employment in constructional steels was in the Eads Bridge built in St. Louis, Missouri, 1867–74. This was the beginning of widespread use of the high-strength, low-alloy steels for constructional purposes. Steel containing 12.5 percent manganese and 1.2 percent carbon was found to be exceptionally tough and resistant to impact and wear. Invented by Sir Robert Abbott Hadfield in England in 1883, this steel is used for railway tires, railway crossings, and the tips and cutting edges of excavating equipment. Hadfield also assisted in the development of silicon-bearing steels for electrical machinery. These steels came into use just after 1900. Late in the 1880s there was a rapid development of the use of nickel as the alloying element in steels used in guns, armour plate, and projectiles in Britain, France, and the United States. Improved armour plate was obtained with the addition of chromium to nickel steels by the French in 1891. Soon after, the use of nickel-chromium

Steel used in construction

steels spread to all types of forgings and machine parts. High-speed tool steels that retain a cutting edge at high temperatures and permit high cutting speeds were first developed in the United States about 1900. These steels contain tungsten, chromium, and vanadium. The appearance of the motorcar in considerable numbers about that time also accelerated the use of alloy steels for crankshafts, gears, springs, and shafts. Vanadium is used extensively in these steels because it imparts resistance to shock and fatigue. The appearance of the electric furnace not only led to improved methods for making alloy steels but also resulted in the production of numerous kinds of ferro-alloys required in the manufacture of these steels. Shortly after 1900, ferro-alloys containing silicon, chromium, tungsten, titanium, molybdenum, and vanadium became readily available to the alloy steelmakers. Stainless steels, which contain over 5 percent chromium, were first developed in Great Britain (1912). These steels have superior resistance to corrosion and were first used in cutlery. Between 1909 and 1912 the more familiar chromium-nickel stainless steels resulted from research work in Germany. The best known composition contains 18 percent chromium and 8 percent nickel; hence, it is known as 18–8 stainless.

Although many alloy steels were developed for munitions and military purposes, during wartime the shortage of materials made it necessary to economize on the use of alloying elements. In World War I a number of special alloy steels were developed that required reduced levels of alloying elements to obtain the desired properties. This was also true during World War II, when additional economies in the use of alloying elements were achieved by adding boron to steels. Boron is a potent agent, and 0.0005 percent of it permits the reduction of other alloying elements by half without loss of strength, ductility, and hardness. Subsequent to World War II, metallurgists throughout the world have been seeking ways of making and solidifying steels and processing them to reduce the amounts of expensive alloying elements that have been used in the past to impart special properties. To a large measure, the efforts have been successful, and in many places low-alloy steels or coated carbon steels are being used in place of more highly alloyed steels.

Protecting steel

Steel can be protected from corrosion and rusting by coating it with tin, zinc, chromium, lead, or copper. Tinning of iron was first practiced in Bohemia in the 17th century. The iron was cleaned in acid and then was dipped into a pot of molten tin to form a thin, adherent layer of tin. This method of coating iron and steel with tin, zinc, or lead containing some tin has been followed for centuries. Galvanized steel has a coating of zinc, and so-called terne plate has a lead-tin coating. To conserve tin, which was scarce in World War II, the electrolytic method was developed for making tinplate. A very thin coating of tin is applied continuously to the surface of a strip of sheet steel by the electrolysis of tin from an aqueous solution. The layer is then melted quickly and cooled to form a bright, adherent coating. The production of galvanized steel was made continuous just after World War II. A strip of sheet steel moves continuously through a furnace, which prepares the surface, and then the strip is passed through a pot of zinc alloy. Subsequently, the strip is cooled to solidify the zinc. Since 1960 a wide variety of coatings have been developed for steel sheets; aluminum coatings by hot dipping or deposited from the vapour and a large variety of organic coatings are examples.

MODERN STEEL TECHNOLOGY

Raw materials. *Pig iron.* Virtually all large steel plants in the world have blast furnaces for producing liquid pig iron, which usually constitutes more than 50 percent of the metal-bearing part of the charge for the steelmaking furnace. In these large plants, a small quantity of cold pig iron may also be included in the charge for steelmaking, but hot metal is preferred, because no extra heat is required to melt it. Usually the cold pig iron is obtained from discarded ingot molds and the excess hot metal, which is cast into pigs from time to time. Pig iron may constitute up to 15 percent of the metallic charge of electric and open-hearth furnaces that operate on a completely cold charge, as it is an essential source of carbon for the system in these cases. Hot metal is transported from the blast furnace to the steel-producing shop in ladles of various types that are carried on special railway trucks. Common forms for these ladles are either a simple refractory-lined bucket (having a heat-resistant lining) with an open top or the submarine-shaped refractory-lined tank having an opening in the top central portion. There is also equipment for tilting the ladle so that metal can be poured from it and slag and refuse can be dumped out. The capacities of submarine ladles range from 100 to 300 tons, and iron will remain molten in them for periods as long as 24 hours. At the steelmaking shop, hot metal is poured from these ladles into an open-top ladle, which is used to transfer the metal to the furnace. In some plants, especially the older open-hearth shops and Bessemer shops, the metal will first go to a large refractory-lined tank or mixer that serves as a reservoir. A mixer usually has a capacity of 1,000 to 1,500 tons of hot metal. Liquid metal is taken from the mixer as needed by the steel furnaces.

Steel scrap. Steel scrap is the second most important source of iron in the charge to the steel furnace. Generally speaking, scrap has constituted between 48 and 50 percent of the total tonnage of metallic raw materials used in steel production throughout the world in recent years. There are three primary sources of scrap: revert scrap is the material discarded within the steel plant in the production of steel products such as plates, sheets, tubing, and wire from the raw cast steel; industrial scrap is the discarded steel from the plants and factories that process steel into useful objects, such as automobile bodies, machinery, and appliances; and reclaimed scrap is the ferrous materials obtained from refuse, worn-out objects, old buildings, and other discarded objects. Revert scrap varies greatly in size and thickness, because it consists of discarded ingots, crops removed from the ends of bars, billets and blooms during rolling, and clippings and trimmings from all kinds of products as they are cut to the required dimensions. It includes rejected products that may be found unsatisfactory at any point in the production process. Approximately 30 percent of the raw steel produced by the steel furnaces is recycled for remelting within the steel-producing plant. Thus, it is imperative for economic reasons that the steel-melting facilities be able to consume this scrap. Where possible, the steel plant segregates scrap from special alloy steels to use when making new steel having a similar composition.

Sources of scrap

Industrial scrap from factories and plants is generally of very good quality. Usually it is clean, its composition is known, and it is baled, or compacted into bundles of a size suitable for charging to the steelmaking furnace. The amount of this scrap produced varies greatly, depending upon the nature of the operation in the factory. Many new manufacturing techniques have been developed, however, that tend to reduce the amount of steel lost as scrap. In consequence, the fraction of the finished steel produced by the steel industry that returns as industrial scrap to the steel plant also will probably be reduced in the years ahead. The physical condition and composition of reclaimed scrap varies widely. It may be contaminated with undesirable substances (*e.g.*, paints, coatings, cement, and dirt) containing zinc, copper, tin, sulfur, arsenic, or silica that cause difficulties in steelmaking and may be deleterious to many grades of steel. An element such as copper or nickel, for example, may be valuable in certain kinds of steel but quite undesirable in others.

Metallized ore. Metallized ore, the solid metallic product obtained by the direct reduction of iron ore, is becoming an important raw material, particularly for the electric-arc furnace. It is of great value in the charge for high-grade steels because of its freedom from contaminants such as copper and tin. It has the disadvantage that the worthless rock (gangue) it carries from the ore increases the slag volume of the melting furnace. The reduced product must be processed and handled so it does not reoxidize extensively between the time it leaves the reduction system and the time it is melted in the steel furnace. In some cases the spongy metallic product is compacted under high pressure to form dense briquettes. These briquettes can be han-

dled and stored under cover without reoxidation being a serious problem. The usual practice has been to transfer the reduced product, either hot or cold, directly to the furnace for melting with a minimum of delay. Provisions for a cover of an inert gas may be included in the handling facilities and storage bins to prevent contact of the metallized ore with air and moisture. Metallized ore may constitute part or all of the initial charge to a furnace, and it may also be fed continuously as the furnace operates. In the latter case, special handling facilities are required.

Fluxes. Limestone, lime, and fluorspar (CaF_2) are important as fluxes for forming a slag of the correct composition and consistency in the steelmaking furnace. The amounts used vary with the different steelmaking methods and will be considered later when each process is described. Lime and limestone must have low levels of sulfur, in the range of 0.04 percent, and silica of under approximately 2 percent to be satisfactory. They must also be crushed and sized by screening to eliminate dust and large pieces. The most desirable burnt lime is as completely calcined as possible without becoming hard and dense due to its being overburned. Similarly, fluorspar should be coarsely granular to prevent loss by dusting when it is shipped, transported within the steel plant, and added to the furnace. Relatively large quantities of fluorspar are required in oxygen steelmaking to assist in the formation of the slag.

Preparation of fluxes

Liquid iron. The cylindrical-shaped cupola furnace has been used since Bessemer's time to melt pig iron and scrap to provide liquid iron high in carbon for steelmaking. High-quality coke is used as a fuel. During the early years of the Bessemer process, all of the iron from the blast furnace was pigged, and this cold iron remelted in a cupola to provide liquid iron for steelmaking. Several decades elapsed before liquid iron directly from the blast furnace could be used. Since World War II the hot-blast cupola has been used to melt scrap to provide a supply of liquid pig iron for oxygen steelmaking in steel plants where the demand for steel is insufficient to support a blast furnace.

Oxygen. Enormous quantities of gaseous oxygen are used in steelmaking throughout the world. Produced by the fractionation of air, the gas is usually 99.5 percent pure to avoid contaminating the steel bath with nitrogen. In most cases the gas is transmitted by pipeline from the producing plant to the steel furnace. The basic oxygen furnace consumes more oxygen per ton of steel made than does the basic open hearth, and, as the former has replaced the latter, the consumption of oxygen in steelmaking has risen rapidly.

Refractory materials. Large quantities of costly high-grade refractory materials are consumed in steelmaking. Every steel plant must have facilities for storing and handling these refractories and for relining ladles and converters and for rebuilding furnaces. The consumption of refractories by the basic open-hearth process is much greater than it is for the basic oxygen process, because the furnace structure is more complex. As the basic oxygen process has replaced the basic open-hearth process, the requirement of basic refractories in steelmaking has declined markedly.

The basic steelmaking process. Virtually all the world's annual production of 670,000,000 to more than 700,000,000 tons of steel is made in a process that begins with the production of an impure liquid bath from a charge that is principally steel scrap, hot metal, pig iron, and, perhaps, a small amount of ore. The bath is then refined to obtain the desired carbon content and to remove impurities. Finally, alloying elements are added to the refined bath to adjust its oxygen content and to bring its composition to the formulation required for a given steel product. The major impurities in the bath removed during refining are carbon, silicon, manganese, phosphorus, and sulfur. These are oxidized to form a gas containing carbon monoxide and carbon dioxide from the carbon and a liquid slag containing the oxides, silica (SiO_2), manganese oxide (MnO), and ferrous oxide (FeO), which react with lime (CaO) in the charge. Undesirable materials introduced with the charge that also enter the slag are sand, dirt, bits of cement, small quantities of blast-furnace slag, and gangue in the ore. Materials such as lime, limestone, and fluorspar (CaF_2) may be added to assist in the formation of the slag and to give it the desired composition.

Impurities in the liquid bath

Types of furnaces. Contemporary steelmaking furnaces can be classified broadly in terms of their shape and the way in which heat is provided. The open-hearth furnace has a broad, shallow hearth to contain the pool of metal and slag. Heat is supplied from a large, luminous flame that plays over the surface of the slag. The electric-arc furnace also has a shallow hearth; heat is supplied from arcs struck between graphite electrodes and the metal bath. The converter is a pear-shaped vessel into which an oxidizing gas is injected. Heat to support the process is provided by reactions between the gas and materials in the charge. An induction furnace is also used on a small scale: the scrap charge is contained in a cup-shaped crucible, and heat for melting and refining is supplied from electric currents induced in the metallic charge from an external coil in which an electric current is circulated.

Slags and refractories. Steelmaking processes may also be differentiated on the basis of the type of slag necessary for refining the steel and the type of refractory used in the furnace. Broadly, there are two types of oxides used in slags: acids, such as silica (SiO_2) and phosphorus pentoxide (P_2O_5), and bases, such as lime (CaO) and magnesia (MgO). Ferrous oxide (FeO) and alumina (Al_2O_3) are essentially neutral in steelmaking slags, and manganous oxide (MnO) tends to be basic. Acidic steelmaking slags contain an appreciable concentration of silica with very little lime, and basic slags contain an appreciable concentration of lime and up to about 20 percent silica. The decision whether to use an acid or basic slag depends on the levels of sulfur and phosphorus to be contained in the charge to the furnace. If the amount of either is greater than that which can be tolerated in the final steel product, it is necessary to use a basic slag, because it is able to absorb both of these elements. The nature of the slag also determines the type of refractory materials that are used to contain the slag and metal. An acid slag will attack and rapidly dissolve basic refractories, while a basic slag will destroy acid refractories. Silica is the principal refractory used in acid steelmaking furnaces. Basic refractories are composed chiefly of pure magnesia (MgO), calcined dolomite ($CaO \cdot MgO$), or lime. In some cases, chromite ($FeO \cdot Cr_2O_3$) may be used as a hearth lining if steels high in chromium are to be made. It is usual to choose a refractory with a low solubility in the slag. Only a few percent of magnesia, for example, will dissolve in basic steelmaking slags. The pure oxides from which refractories are made have melting points well above the operating temperatures of the steelmaking system. In most processes, the composition of the slag must be adjusted so that it is completely liquid at the operating temperature of the furnace.

Acidic and basic slags

Amount of nitrogen. The amount of nitrogen introduced into the steel during processing also plays an important part in the choice of a steelmaking process. Actual levels are very low; 0.005 percent or less is desired in sheet steel, and values above 0.009 percent are desired only in selected types of steel. (The effect of nitrogen on the properties of steel is described below.) Nitrogen is introduced into liquid steel during melting and refining if the liquid metal is exposed to air or other gas that contains nitrogen. The rate of absorption rises rapidly as the temperature of the liquid metal increases and is particularly rapid when the metal is exposed to an electric arc operating in air or when a jet of air or of oxygen carrying small concentrations of nitrogen impinges on the metal surface. The practical steps necessary to minimize the uptake of nitrogen by the metal are to keep the metal covered with a slag that is impervious to nitrogen, to exclude nitrogen from the furnace atmosphere and be certain that the oxygen used in refining has a very low nitrogen content, and to obtain a vigorous boiling action of the bath by the formation of carbon monoxide as bubbles of these gases carry nitrogen out of the system.

Refining reactions. Liquid impure iron and steel can be refined rapidly by reactions with oxygen from an oxygen jet, from iron oxides such as ore added to the furnace, and from iron oxide dissolved in the slag. The reactions

between ferrous oxide in the slag and silicon, manganese, and phosphorus in the crude iron can be written:

$$2FeO\ (slag) + Si\ (metal) = SiO_2\ (slag) + 2Fe\ (iron)$$
$$FeO\ (slag) + Mn\ (metal) = MnO\ (slag) + Fe$$
$$5FeO\ (slag) + 2P\ (metal) = P_2O_5\ (slag) + 5Fe.$$

The first two reactions occur very early in the refining process, and usually all but minute quantities of silicon and manganese go into the slag, particularly if it contains more than about 10 percent ferrous oxide. The oxidation of carbon by iron oxide in the slag is written:

$$FeO(slag) + C(metal) = CO(gas) + Fe$$
$$2FeO(slag) + C(metal) = CO_2(gas) + 2Fe.$$

The gas formed will contain only a few percent carbon dioxide. Pure liquid iron at 1,600° C (2,900° F) in the presence of pure liquid ferrous oxide will dissolve 0.24 percent oxygen. Since solubility rises very rapidly as the temperature increases, the steel bath will absorb oxygen from ferrous oxide in the slag and from a jet of oxygen gas impinging on the surface. As oxygen is added to the bath, it will react with carbon to form carbon monoxide and carbon dioxide. A steel bath at 1,600° C containing 1 percent carbon will have to contain approximately 0.002 percent oxygen to be able to eliminate carbon as carbon monoxide in a bubble of gas passing through the metal. As the carbon content of the bath is reduced during refining, the oxygen content of the bath necessary to drive carbon out into the gas phase rises rapidly. At 0.1 percent carbon, a minimum of 0.02 percent oxygen in the steel is needed, and, at 0.05 percent carbon, 0.04 percent oxygen is required. During refining of the steel bath the actual oxygen level is about twice the minimum. Since the solubility of oxygen in solid steel is very low, the oxygen in the liquid steel evolves very rapidly, chiefly as carbon monoxide, during solidification. Liquid steel containing 0.2 percent carbon and 0.02 percent oxygen may evolve over 100 times its own volume of hot carbon monoxide as it solidifies. Unless the release of gas is controlled, sound ingots and castings of the steel cannot be obtained. Control is effected after refining by adding deoxidizing agents such as aluminum, silicon, titanium, and manganese to the liquid; they combine with most or all of the oxygen and produce small particles of the oxides of the elements. Some will float out of the steel, but many will remain in the solidified steel. If a metal serves as an alloying agent and will also deoxidize the steel, only a small excess need be added for deoxidation, because the amount consumed by the reaction with oxygen is usually only a few hundredths of a percent of the weight of the steel.

Removing sulfur from the bath

Sulfur can be removed from the steel bath only if a basic slag is present. Very little sulfur is carried off in the waste gases. The chief requirements for control of sulfur are that the ratio of lime to silica in the slag be greater than 2.8, that the slag be fluid and hot, and that the slag and metal be stirred together actively. Strongly reducing conditions in the furnace also favour desulfurization. The concentration of sulfur in a basic slag ranges between 0.05 and 0.20 percent in normal commercial oxidizing steelmaking operations such as the open hearth and the basic oxygen furnace. Since the final steel should contain from about 0.01 to 0.025 percent sulfur, control of this element is a major problem of the steelmaker.

Types of steelmaking processes. *Acid Bessemer process.* The converter developed by Bessemer and Kelly is lined with a siliceous refractory material and operated with an acidic slag and, therefore, cannot be used if phosphorus must be removed from the charge. Its use grew rapidly in the United States, where pig iron low in phosphorus was produced from plentiful supplies of low-phosphorus ores. On the other hand, it found little use in Europe and Great Britain, where most steelworks produced hot metal high in phosphorus from local ores high in phosphorus. Even in the United States, however, the open-hearth process enjoyed advantages because of superior control of sulfur and the avoidance of nitrogen in the steel from the air blast.

The acid Bessemer converter is either a pear-shaped or a cylindrical vessel with a slightly constricted nose that has an opening at the top (see Figure 13). Waste gases

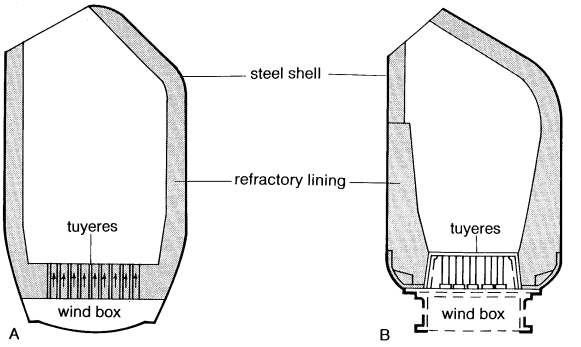


Figure 13: *Bessemer vessels.*
(A) Bottom-blown acid converter; (B) bottom-blown basic converter.
By courtesy of United States Steel Corporation

leave through this opening. The vessel also has a perforated bottom for admission of the air blast. The furnace is mounted in a steel ring around its middle, and this ring is supported by two horizontal shafts, or trunnions, fixed at a right angle to the longitudinal axis of the vessel so it can be tipped for charging and tapping. All charge materials enter through the mouth of the vessel, and finished steel is poured from the mouth as well. Compressed air for the blast is led through a tube in one trunnion and then into a steel box on the bottom of the vessel, from which it passes into the vessel through holes (tuyeres) in the bottom refractory lining.

The refractory lining of the vessel walls, usually composed of sandstone or micaceous schist, will usually last over a thousand heats. The bottom refractory, however, is subject to severe wear and survives only 25 to 35 heats. Accordingly, the bottom of the vessel is constructed for ready removal to be relined and replaced. Spare bottoms are kept on hand and can be installed in 30 minutes.

In the United States a Bessemer shop was usually equipped with two or three vessels that stood in the open. One vessel was available for relining and repair while the other one or two were in use. A vessel was usually three to four metres in diameter by six metres high (approximately 11 by 20 feet) and produced from 25 to 30 tons of steel each blow. Ladles of hot metal and chutes for charging scrap were carried by overhead cranes or on railcars. Ladles for the finished steel and pots for the slag were mounted on railcars while being filled. Usually the ladle of steel was transferred by an overhead travelling crane to the teeming platform, where the steel was poured into ingot molds. The operation was controlled by an operator in a pulpit so positioned as to permit an unobstructed view of the vessel. Hot metal was stored in a large mixer with a capacity of between 800 and 1,200 tons. Such a reservoir was needed because hot metal came from the blast furnace infrequently and in large batches but was used in small quantities and at much more frequent intervals. The mixer also provided a supply of hot metal that was relatively uniform in composition and temperature. The shop included facilities for scrap preparation, relining of the vessel and bottoms, storage and weighing and ferroalloy additions, and a compressor to provide air blasts at a pressure of 1.4 to 1.75 kilograms per square centimetre (20 to 25 pounds per square inch).

The hot-metal reservoir

The usual hot metal for acid Bessemer operations contains 1.1 to 1.5 percent silicon, 0.4 to 0.7 percent manganese, about 4 percent carbon, 0.09 percent maximum phosphorus, and 0.03 percent maximum sulfur. A typical charge is approximately 90 percent hot metal and 10 percent scrap. Sometimes small quantities of high-quality iron ore or roll scale are added to decrease the blow time. When this is done it is also necessary to reduce the amount of scrap because of the extra heat required to reduce the iron oxide. The furnace is charged in the horizontal position, the air blast is turned on, and then the furnace is rotated to the vertical position, a procedure that prevents plugging of the tuyeres. The cold-air blast oxidizes manganese and silicon and some iron to form a slag in the first four minutes of the blow. During this time,

the flame emerging from the mouth of the vessel is short and transparent. Then the carbon is oxidized, producing a bright flame about 10 metres (30 feet) long. This flame results from carbon monoxide burning in the surrounding air. When most of the carbon is burned out of the steel, the flame drops. An operator learns from experience to judge the carbon content and temperature of the steel by the character of the flame. A blow lasts 12 to 15 minutes, and the final slag contains iron oxide, manganese oxide, and silica, the amount being approximately 7 percent of the weight of the steel made. Additional scrap, iron ore, cold pig iron, or steam in the blast can be used to cool a hot heat; the temperature of a cold heat can be raised by blowing with the vessel turned on its side. This oxidizes iron into the slag, which generates considerable heat, but reduces the amount of steel produced from the charge. Deoxidizing agents and alloying additions to bring the steel to the desired composition are added to the ladle as the steel is tapped from the furnace into the ladle. For a long time, high-carbon steels such as rails were made by recarburizing the steel with hot metal. The practice led, however, to a product of poor quality.

The basic Bessemer process. Developed to produce steel from the low-grade ores of Europe and similar ores in Britain, the basic Bessemer, or Thomas, process has had an existence in those areas similar to that of the acid Bessemer process in the United States. Production of Thomas steel rose rapidly until 1910 and, except for the period of high demand during World War II, declined as a result of competition from basic open-hearth steel. The process is suffering further decline as a result of the increased importation of high-grade ores that also are low in phosphorus.

In most external aspects, the basic Bessemer steel shop is similar to the acid Bessemer plant. There are essential differences in the way the process operates, however, because a basic refractory and a basic slag are employed.

The basic and acid Bessemer vessels are similar in shape, except that in the basic vessel the nose is usually concentric with the longitudinal axis. The capacity of basic furnaces and the relatively deep bath requires a blast pressure of 2.5 kilograms per square centimetre (35 pounds per square inch). The lining is a plastic mixture of tar and calcined dolomite or lime that is rammed into place. The wall and nose lining lasts up to 400 heats and the bottom refractory to about 40.

Thomas process in operation

The operation of the Thomas process varies so widely from plant to plant that only the general pattern can be described: the vessel is charged with hot metal, the composition of which is 1.5 to 2 percent phosphorus, 0.6 to 1 percent manganese, 0.2 to 0.4 percent silicon, about 3.8 percent carbon, and a maximum of 0.05 percent sulfur. It should be noted that the regular basic Bessemer process cannot operate successfully on pig iron containing less than 1.5 percent phosphorus, because insufficient heat is generated to meet the needs of the process. The weight of lime charged is approximately 15 percent of the weight of the finished steel. This lime is charged with the hot metal. Only about 10 percent of the metallic charge is steel scrap, and it is usually added late in the blow, because much of the heat is generated by oxidation of the phosphorus occurring late in the heat. The reactions proceed much as they do in the acid process, with silicon and manganese being oxidized first, followed by oxidation of carbon. The flame drops when the carbon reaches about 0.04 percent. About three minutes before the flame drops, scrap is added to the vessel, and the blow is continued for two to three minutes more. During this time, the principal reaction is oxidation of phosphorus into the slag. At the end of the blow, the carbon, manganese, and silicon levels in the metal are very low, and the phosphorus content is about 0.04 percent and nitrogen 0.01 percent. The carbon must always be brought to a low level to obtain a dephosphorizing basic slag with 15 to 20 percent ferrous oxide. Thus, it is necessary to add to the ladle large amounts of ferromanganese, other carbon-containing additions, and a deoxidizer as the heat is tapped to obtain steel of the desired final composition.

The final weight of the basic Bessemer slag is about 20 to 22 percent of the weight of the steel made. This slag contains 16 to 18 percent phosphoric acid and up to 50 percent lime, so that it has been sold profitably as a fertilizer. The recent production of less expensive fertilizer from phosphate rock has meant the loss of revenue to the steel plant from the sale of this slag.

Low-nitrogen steel

A number of special practices have been developed to permit the production of steels low in nitrogen by the basic Bessemer process. These consist of keeping the bath cool and shortening the afterblow by adding roll scale or ore (up to 2.5 percent of the metal weight) just before the flame drops, by first blowing half of the charge to completion in order to develop a slag containing considerable iron oxide and then adding the balance of the hot metal and completing the blow, and by equipping the vessel with tuyeres on the sidewall so that the air has a shorter path through the metal and is in contact with the bath for a shorter time. Use of a blast of oxygen and hydrocarbon gas permits refining of hot metal containing little or no phosphorus. Oxygen is used in some plants to enrich the air blast. This practice permits the use of hot metal with less phosphorus, because, with a lower level of nitrogen in the blast, less heat is carried from the system by the waste gases. Alternatively, the use of oxygen permits the percentage of scrap that dilutes the phosphorus in the hot metal and also shortens the afterblow.

Steels produced by the basic Bessemer process are used extensively except where high ductility and formability are required in low-carbon sheet steels. With special practices in the afterblow, particularly blowing with oxygen and steam, low-carbon steels, with low levels of phosphorus and nitrogen, can be made that are competitive in properties with open-hearth steels.

The basic oxygen process. Though the basic oxygen furnace was developed in form from the basic converter, the chemistry of the process is more closely related to that of the basic open-hearth process. The basic oxygen furnace is a large, cylindrical steel vessel with a conical nose and a solid bottom. The opening in the nose is usually concentric, and a taphole is placed about halfway up the conical portion. The refractory lining of the walls and nose cone is made of a mixture of tar and calcined dolomite. The refractory on the bottom may also be made of tar–dolomite brick or plastic mix rammed in place. A supersonic jet of oxygen is projected down onto the central portion of the charge by a large water-cooled lance inserted vertically through the mouth of the vessel. The vessel is mounted on trunnions so it can be tilted for charging, tapping, and repair of the refractories. Two or three vessels are usually installed together in a row, and one vessel is relined while the other one or two operate. The building housing the furnaces is tall to accommodate bins and weighing equipment for delivering lime, limestone, calcined dolomite, and spar to the vessel as required. The lance-support system and replacement lances are also located above each furnace. Waste gases and heavy red fumes from the furnace are collected in a hood, passed through cleaning equipment, and discharged through a tall stack. Often the hood is water-jacketed so the heat of the hot gases and that resulting from the combustion of the waste gases with air is used to generate low-pressure steam: a hood system has been developed in Japan that is sealed to the mouth of the vessel, and the waste gases are collected for use as a gaseous fuel in other areas of the steel plant.

Charging the vessel

The vessel is charged with hot metal from a large ladle carried by an overhead crane. Scrap may also be charged from a chute carried by the same crane or may be added from a large box mounted on a very large tilting device that travels on rails. The ladle for the steel and a pot to collect the slag are carried under the furnace by a car that travels on rails. This car moves the steel ladle and slag pot out from under the furnace on the side opposite to the charging side. The ladle of steel is then carried by another overhead crane to the place where the steel is poured into ingots or into a strand-casting machine. There are several positions from which the furnace is operated. The controls for the furnace, oxygen lance, waste-gas system, and bins for additions to the furnace are located in a small, well-protected building on the level of the furnace trunnions

and where the operator can see the mouth of the vessel during blowing. There are usually auxiliary control stations for tilting the vessel on the charging side and on the tapping side.

The charge to a furnace is usually 27 to 33 percent scrap, with the balance being hot metal. The amount of scrap will be a little higher if the metal is hotter and the level of silicon is high. Scrap in the charge has been increased up to 40 percent by preheating the scrap and by the use of an auxiliary fuel such as silicon carbide. The furnace will operate successfully on a broad range of compositions of hot metal—3.5 to 4.3 percent carbon, 0.2 to 2.0 percent silicon, 0.1 to 2.5 percent manganese, up to 0.45 percent phosphorus, and up to 0.06 percent sulfur. When making steels with sulfur below 0.025 percent, however, it is necessary that the sulfur in the hot metal be less than 0.04 percent. The operation of the process is not particularly sensitive to the kind of steel scrap in the charge. The oxygen required is approximately 0.07 tons per ton of steel made (55 cubic metres per metric ton [1,750 cubic feet per net ton]). Gaseous oxygen is delivered to the lance at a pressure of 10 to 13 atmospheres gauge (150 to 180 pounds per square inch). The amount of lime and calcined dolomite required varies depending upon the silicon content of the iron, but an average figure is 8 percent of the weight of the finished steel. Fluorspar (CaF_2) is used almost universally as a fluxing agent to help dissolve the lime in the slag and to produce a thin, fluid slag. The amount used also varies widely and increases as the level of manganese in the hot metal decreases. An average figure for U.S. practice is 0.7 to 0.8 percent of the weight of steel. The slag produced is 10 to 15 percent of the weight of the steel. The scrap is usually charged to the vessel first, and then the hot metal is poured in. The oxygen lance is lowered to within 1.5 to three metres (five to nine feet) of the metal surface, and the oxygen jet is turned on. The blowing rate depends primarily upon the size of the furnace, the rates being approximately 0.035 ton of oxygen a minute per ton of steel capacity; *i.e.*, a 200-ton furnace is blown at a rate of 0.7 ton of oxygen per minute. Usually within seconds after the oxygen is turned on, flame and fumes appear at the mouth of the vessel, and the refining reactions begin. Shortly thereafter the desired amounts of lime, calcined dolomite, fluorspar, and ore, if used, are added. Within four to six minutes an early slag is formed by oxidation of iron and of impurities in the charge. After eight minutes, oxidation of carbon begins and continues until the flame drops. If the reaction becomes too violent, slag and metal may be ejected from the furnace, and the operator adjusts the lance position and oxygen flow to correct the condition. Usually the amount of the charge and the blowing rate are designed to give a blow time of from 15 to 20 minutes. In most furnaces, the flame drops when the carbon is about 0.06 percent. If the heat of steel is to have a higher concentration of carbon, the blowing is stopped earlier. When the blow is ended, the furnace is turned down, a thermocouple is inserted into the bath to measure its temperature, and a sample of metal is taken for chemical analysis. Often a sample of the slag is also taken. The carbon content of the bath is determined within a minute or two, and the contents of metals such as manganese, phosphorus, sulfur, copper, tin, and nickel can, if needed, be determined by a spectroscope in a few minutes. The ferrous-oxide content of the slag and the lime–silica ratio may also be obtained. If there is too much carbon, sulfur, or phosphorus in the steel or if the temperature is too low, the heat is reblown for an interval of from one-half to three minutes. If the steel is too hot, limestone or scrap is added. As the steel is poured into the ladle, ferro-alloys and deoxidizers are added to bring it to the desired composition.

Heat times

The time from the beginning of charge of one heat to the beginning of charge of the next heat ranges from just under an hour for large furnaces to less than 30 minutes for small furnaces. There is vigorous competition among furnace crews, steel plants, and nations for the fastest furnace; a good performance is 30 heats a day for one furnace, but an average of 50 heats a day for a month has been achieved. The composition of a typical slag from a low-carbon heat is 55 percent lime, 25 percent ferrous oxide, 16 percent silica, 2 percent magnesia, and a few percent each of manganous oxide and phosphorus pentoxide. The sulfur content would range from 0.05 to 0.20 percent.

The reactions among silicon, manganese, and carbon proceed in much the same manner as in the acid and basic Bessemer converters. Silicon and manganese oxidize rapidly to silica and manganous oxide. Some ferrous oxide is formed, and these oxides react with the lime, calcined dolomite, and fluorspar to form a slag within four or five minutes after the fluxes have been added. As soon as a slag has formed, some phosphorus is oxidized to phosphorus pentoxide and is absorbed by the slag. After approximately eight minutes of blowing, only a trace of silicon is left in the metal, and most of the manganese that will oxidize is in the slag. At this point the reactions with carbon become much more vigorous, and the slag and metal are highly agitated and mixed by the release of very large volumes of gas. The high speed of the basic oxygen process results from this agitation. When the carbon in the bath has reached 1 percent, the phosphorus in the bath is usually less than 0.05 percent, so satisfactory steel with a carbon level as high as 1 percent can be made. As the carbon drops, the ferrous oxide in the slag rises, and the lime becomes fully dissolved to produce a fluid, basic slag that will dissolve sulfur and more phosphorus. Nitrogen is not absorbed by the metal in the process, because high-purity oxygen containing less than a 0.1 percent of nitrogen is used, and the vessel nose is designed to prevent any air from entering the converter when the metal is being agitated violently. Levels in the final steel between 0.003 and 0.005 percent nitrogen are obtained.

A great deal of attention has been devoted to finding means for automatic control of the process. Instruments have been developed for determining the temperature and carbon content of the bath and the amount of carbon in the waste gases during the blow. Computer programs have been developed that calculate the charge for a heat and determine such things as the lance position, oxygen delivery rate, and time to stop blowing. Although much has been learned, a wholly complete and automatic control system has not yet been devised.

The original furnaces installed at Linz and Donawitz, Austria, in 1952 and 1953 had capacities of 39 tons. Since then the size of furnaces has grown rapidly, and many vessels with capacities of as much as 200 and 300 tons are now in operation in Europe, North America, Australia, Great Britain, Russia, and Japan. The basic oxygen furnace is rapidly replacing the basic open-hearth furnace and the Thomas converter for producing carbon steels. Virtually all production of carbon steels in Japan is by oxygen converter. The first oxygen converter was installed in the United States in 1956, and, by 1981, about 60 percent of the steel made was by the process. Large plants have also been installed in the United Kingdom, Italy, and France.

Increasing use of the basic oxygen process

A special two-slag operation is carried out when the basic oxygen furnace is used to refine high-phosphorus hot metal, which previously required the use of the Thomas process. An early slag that is high in phosphorus is formed by carrying powdered lime into the vessel along with the oxygen stream in a special lance. A fluid slag is formed, the blow is interrupted, and this slag is poured off. The blow is continued, and more lime is added, to form a second slag typical of normal operations. The first slag can be sold as fertilizer because of its high phosphorus content. In some cases, the second slag is retained in the vessel after the steel is tapped and becomes the first slag in the next heat. This two-slag practice is used in Luxembourg, Belgium, France, and Great Britain.

The open-hearth process. The open-hearth furnace has a large, rectangular, shallow hearth that contains the molten pool of metal and the liquid slag floating on its surface (see Figure 14). Walls of brick and an arched brick roof confine the large flame that is the source of heat in the process and reflect heat from the flame down onto the hot steel and slag. This brickwork is held in place by a framework of heavy steel members. Brick ducts at each of the short sides of the rectangle—*i.e.*, the ends of

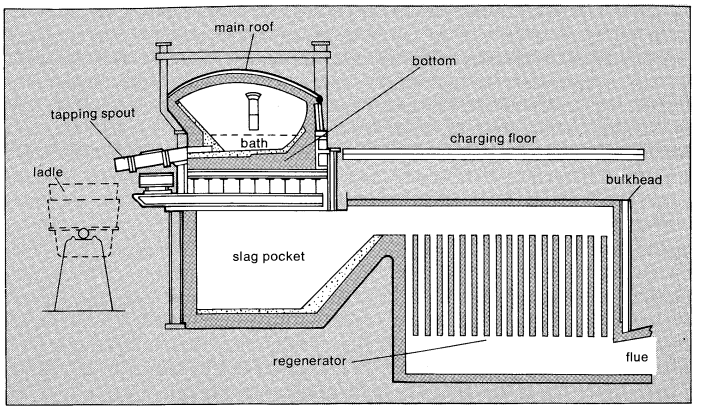


Figure 14: Open-hearth furnace (not to scale). Upper portion shows cross section through taphole of furnace; bottom portion shows cross section through slag pocket and regenerator.

From H.E. McGannon (ed.), *Making, Shaping and Treating of Steel*, 9th ed. (copyright 1971 by United States Steel Corporation)

the furnace—permit the passage of air and gases between the furnace chamber and the brick regenerator chambers located one floor below the hearth level. A large water-cooled burner is mounted at each end of the furnace and, when in operation, projects natural gas or atomized heavy fuel oil, pitch, or tar down the furnace chamber. This spray of fuel is burned by the stream of hot air from the regenerator. A jet of oxygen gas may also emerge from the burner under the fuel jet to increase the rate at which the fuel is burned. The furnace is charged through a number of openings on the front side of the furnace. These openings are closed by water-cooled, refractory-lined doors that slide up and down. It is probable that the name open hearth is derived from the fact that these openings extend the full length of the hearth to make the furnace chamber accessible for charging, working of the heat, and making repairs. Equipment for servicing the furnace operates on the charging floor at hearth level in front of these doors. The materials to be charged are brought to the furnace in large steel boxes carried on small railway cars. A charging machine that also runs on rails is equipped to pick up a charging box, thrust it through an open door, dump its contents on the hearth, and return it to the small car. Hot metal is poured into the furnace from a ladle carried by an overhead travelling crane. The metal is poured in through a movable spout inserted in one of the open doors.

A taphole for draining steel and slag from the furnace is built into the bottom of the furnace on the side opposite the doors. This hole is plugged with refractory, and when the heat is to be tapped the plug is partially dug out and then blasted open with a small explosive charge. The metal flows through a short trough called a launder and falls into a large ladle. The slag also flows into this ladle and floats on top of the liquid steel. When the ladle is full of steel, the slag cascades into a cast-steel slag pot set beside the ladle. The full ladle is transported by another overhead travelling crane to a stand where ingot molds are filled, or teemed. In some shops the ladle is moved in this fashion to a strand-casting machine.

Oxygen lances

Most modern furnaces are equipped with one or more oxygen lances that project down through the roof. A lance can be lowered so that its tip is close to the slag and the jets of oxygen from numerous small holes in the tip impinge on the surface of the metal bath. A lance also can be withdrawn completely from the furnace so as not to interfere with other operations, such as charging. These lances are smaller than those used in the basic oxygen furnace and deliver oxygen at a much lower rate to avoid heavy splashing of slag and metal. The waste gases leaving the furnace carry large quantities of fume, particularly when oxygen is being used. Gas-cleaning systems have been installed on many modern furnaces to collect this fume before the gases are discharged into the atmosphere.

The regenerators for each furnace are two large chambers, each of which may be as big as the furnace itself, located below the furnace and under the charging floor. Each chamber is partially filled with refractory brick stacked to form many small vertical flues. Air from a blower passes into the lower part of one of the chambers. It then turns and passes up through the flues. The checker brickwork is hot, and the air is heated from 1,200° to 1,550° C (2,200° to 2,820° F) in the passage. The air then passes up into the furnace, where it burns fuel. The gases from the flame and from the reactions in progress in the furnace pass out of the opposite end of the furnace and down into the other regenerator chamber. The ductwork has several abrupt changes in direction and cross section to cause particles of slag, soot, and iron oxide to settle out of the stream of gas before it enters the regenerator chamber. This very hot gas heats the brick flues in the chamber but still retains some heat when it leaves the regenerator chamber. To remove this heat, the gas is passed through a waste-heat boiler. It then goes to the gas-cleaning system and out the stack. The gas flow through the regenerator system and furnace is reversed at approximately 20-minute intervals to keep the air for combustion hot.

The capacities of furnaces range from 200 to 650 tons per heat. Eight to 12 furnaces are usually installed in a long row so that auxiliary equipment can service several furnaces. Hot metal is usually brought from the blast furnace in submarine-type (enclosed) ladles. It may be put in a mixer and transferred to the furnace; or the submarine ladle may serve as a reservoir, and the hot metal may be transferred directly to the open-hearth furnace as needed. Because of its complex shape and regenerator chambers, an open-hearth furnace requires large amounts of costly refractories, and considerable plant space and staff are devoted to the repair and reconstruction of the furnaces. Most furnaces are basic, and the hearth is formed by ramming a plastic magnesite (MgO) mixture into the desired form. In modern shops the walls and roof are made of chrome–magnesite ($Cr_2O_3 \cdot MgO$) bricks, each of which has a thin steel casing. This steel oxidizes in use and functions as mortar to seal the bricks in place. The hearth requires frequent repairs but will last 3,000 to 6,000 heats. The roof will last between 500 to 1,000 heats; its life depends on the care of the operator running the furnace. Each time a new roof is put on, the furnace is repaired, and the regenerators are cleaned out. The whole structure may be rebuilt when the hearth is replaced.

The furnace is able to melt virtually all kinds of scrap, ranging from discarded ingots and odd pieces that formed in the ladle from a cold heat to very lightweight clippings and trimmings left in the processing of steel and miscellaneous iron-bearing waste obtained from refuse.

Production of a heat

The production of a heat of steel starts with an inspection of the furnace and sealing of the taphole with a refractory plug. The charge usually consists of limestone, steel scrap, iron ore, and hot metal. The weight of the limestone is determined by the weight of silicon in the hot metal and normally ranges from 6 to 9 percent of the weight of steel made. The hearth is first covered with a light layer of plate scrap, and then limestone is spread over the hearth. Iron ore, if used, is spread over the limestone. The amount of ore required ranges from 0 to 25 percent of the weight of steel made and is determined principally by the percentage of hot metal and cold pig iron in the metallic charge. If more than 45 percent hot metal is used, some ore will be needed, and the amount is increased as the percentage of hot metal increases above this level. Charging of these materials may take up to 2.5 hours. The flame is kept on during this time to heat the charge, and much of the scrap is melted by the time charging is finished. During melting, some of the scrap is oxidized, and an early slag high in iron oxide forms. The hot metal is poured in as soon as possible after charging is complete. The desired composition of the hot metal is 0.8 percent silicon, 1 percent manganese, less than 0.03 percent sulfur, less than 0.15 percent phosphorus, and about 4 percent carbon. A wide range of compositions is possible. If there is ore in the charge, the hot metal reacts with the hot ore and early slag causing a vigorous boiling action of the metal and slag because of the large amount of gas (principally carbon monoxide) generated. It is desired that the ore boil be vigorous so that an appreciable portion of slag is flushed

from the furnace through a special port or out one of the doors. If there is no ore in the charge, no slag will be flushed from the furnace. The ore-boiling period will take up to 2.5 hours, during which virtually all of the silicon, much of the manganese, and approximately 60 percent of the phosphorus in the hot metal will have been oxidized into the slag. When the ore boil is almost completed, calcination of the limestone in the charge begins, and a heavy boiling action results from liberation of carbon dioxide from the limestone. The lime boil lasts one to 1.5 hours. At its termination, the slag is heavy and lumpy because of the presence of particles of undissolved lime. The metal is all molten, however, and relatively cold. The charge to the furnace is designed so the level of carbon in the bath at this time is 0.2 to 0.4 percent above the carbon level desired in the steel when the heat is tapped. Much heat is required to finish melting the scrap and to calcine the limestone so the flame is maintained in the furnace during the lime boil. The oxygen lances may be lowered and used to assist in melting the scrap; their action also speeds the ore and lime boils. With the lances in operation, the action in the furnace may become so vigorous and the rate of gas evolution so high that the fuel may be turned off to prevent overtaxing the waste-gas system.

Refining the steel

The refining period begins at the end of the lime boil. At this point the bath is cold, the slag is lumpy and raw, and the levels of carbon, sulfur, and phosphorus in the bath are higher in the steel than is desired at the end of the heat. A sample of the bath is taken for chemical analysis to determine the levels of carbon, copper, nickel, tin, phosphorus, and sulfur. Continued heating, use of the oxygen lances, and judicious additions of burnt lime and small amounts of fluorspar developed a fluid basic slag suitable for the removal of sulfur and phosphorus. Iron oxide in the slag and oxygen from the lances react with the carbon in the bath to form carbon monoxide. The heat from these reactions and from the flame raises the temperature of the bath. A skillful operator will bring the bath to the desired temperature and carbon levels simultaneously and will have both sulfur and phosphorus levels below the specified limits.

Heat times range from seven to nine hours and depend on the kind of steel being made, the use of oxygen, and the type of charge. The heat time does not vary appreciably with the size of the furnace. The amount of slag produced depends upon the amount of hot metal in the charge and on the silicon content of the hot metal. The weight will be from 10 to 15 percent of the final weight of the steel bath. The composition of the slags will be very similar to the range given for the basic oxygen furnace. The fuel composition will range from 6 to 12 percent by weight (fuel oil) of the steel made.

Many of the older basic open-hearth plants have been closed down, and most of the rest can be expected to cease operation in the face of competition from the basic oxygen furnace. It would appear, however, that existing modern, well-equipped shops will continue to operate for years to come. The reason is that the open hearth enjoys the advantages over the basic oxygen furnace of greater flexibility in the amount of scrap that can be used in the charge and the ease with which it can make all kinds of carbon steels. It also has the advantage over the electric furnace of being able to produce steels that are low in nitrogen.

The electric process. Steelmaking processes employing electricity as the chief source of heat are utilized in the production of most highly alloyed and specialty steels. The advantages of electric furnaces in steelmaking are absence of contamination of the steel with elements, such as sulfur, that are present in most fuels; flexibility of the system, which permits special slags to be made and removed easily; excellent control of temperature and very high temperatures obtainable; and less loss of expensive alloying elements.

Electric-arc furnace

The most commonly used electric furnace (Figure 15) for steelmaking employs an arc struck between graphite electrodes and the metal charge as the source of heat for the system. The furnace has a dish-shaped circular hearth, vertical sidewalls, and an arched roof. It is mounted on a large rocker assembly so it may be tilted forward for tapping and backward for charging, working the heat, and removing slag. Three cylindrical graphite electrodes (two are shown in the figure) are inserted through holes in the roof and are positioned at the apexes of an equilateral triangle. The electrodes are carried on arms projecting from a vertical mast located on one side of the furnace. The electrodes can be moved up and down individually by these arms. Each electrode is connected to the terminal of a large transformer by a heavy cable and a (copper) bus bar. Electrical energy is supplied to this transformer,

From C.E. Sims (ed.), *Electric Furnace Steelmaking;* the Metallurgical Society of AIME

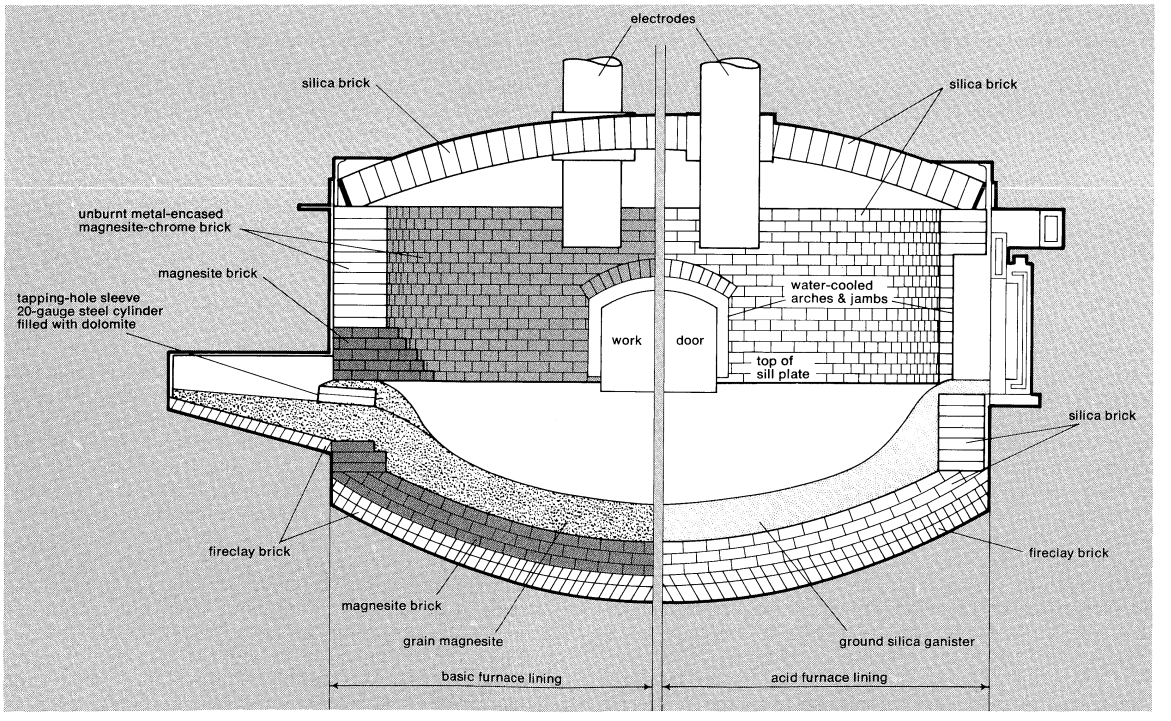


Figure 15: Cross section of a typical electric-arc furnace, showing construction details for (left) basic furnace lining and (right) acid furnace lining (see text).

usually at 22,000 volts, from another transformer and a high-tension line leading to the power generating system. An elaborate set of terminals and switching equipment on the furnace transformer permit energy to be supplied to the furnace in a range of from 50 to 700 volts. Modern furnaces are equipped with a mechanism that lifts the roof, electrode-support system, and electrodes and swings them clear of the furnace, to permit charging from the top by a large steel bucket with a drop bottom. Most furnaces are also equipped with a water-cooled oxygen lance that is inserted into the furnace chamber through a port in the sidewall or in the roof. An electric furnace also has two openings in the sides that are closed by water-cooled sliding doors. One port is opposite the taphole, and the other is 90° from it on the side opposite the electrode mast. These openings permit access to the furnace for adding materials, for removing the slag, and for "working the heat." Machinery for making additions to the furnace and for service and repairs operates on the charging floor at the hearth level. Large furnaces are usually located with the hearth level six to nine metres (20 to 30 feet) above ground level. The support structure is located below the furnace, and the slag pots and the steel ladles are carried on railcars at ground level. Charging buckets are carried by an overhead crane. The steel ladles and slag pots are carried by another overhead crane on the tapping side of the furnace. A plant usually has several furnaces in a line, with the centre line through the tap hole and front door of each furnace at right angles to the line of furnaces. The overhead cranes travel parallel to the furnace line.

Virtually all of the electric-arc furnaces employed for steelmaking have basic hearth refractories and utilize basic slags. The hearth is usually a layer of rammed grain magnesite (MgO) over magnesite brick, or chrome–magnesite materials ($Cr_2O_3 \cdot MgO$) are used when steels high in chromium, such as the stainless steels, are being made. The sidewalls are chrome–magnesite brick, and the roof is made of high-alumina or high-silica brick.

The electrodes are cylindrical lengths of graphite connected together by threaded graphite nipples. As an electrode is consumed, an additional section is added to the end outside the furnace, and the assembly is slipped a distance through the support clamp. Though electrode size is determined by furnace size, furnaces of more than 100 tons capacity usually employ 610-millimetre- (24-inch-) diameter elements; these are the largest practical size commercially available. Either carbon or graphite electrodes may be used. Carbon electrodes are cheaper, but the best quality graphite electrodes are required to meet the severe operating conditions of high-speed, large-capacity modern furnaces. Electrodes are consumed by oxidation, by the action of the arc, and by breakage. The weight of electrodes consumed is 0.4 to 0.75 percent of the steel produced. A modern 200-ton furnace is powered with a 50,000 kilovolt-ampere transformer, and each electrode carries up to 60,000 amperes at full power. A furnace of this size would have a shell 7.3 metres (24 feet) in diameter.

Considerable fume and smoke are generated when the arc furnace is in operation; therefore, to maintain air quality a furnace must be equipped with gas-collecting and cleaning equipment.

Single-slag oxidizing practice

An oxidizing practice is employed for producing plain-carbon and low-alloy steels in electric-arc furnaces of all sizes. Many fairly recently installed large furnaces use this single-slag oxidizing method to produce ordinary carbon steels in competition with the open hearth. Prior to charging, the furnace is inspected, the electrodes are raised, the roof is swung away from the furnace, and the scrap is discharged from a large bucket containing the charge. Some heavy scrap is placed on the bottom of the charge, and light scrap is positioned to fall against the furnace walls to protect them from exposure to the arcs during melting. Hot metal is rarely included in the charge. The burned lime, ore, and fluorspar necessary for the heat are added on top of the scrap. The roof and electrodes are repositioned on the furnace, and the electrodes are lowered, and the power is turned on at about one-third the rated capacity. The electrodes, on automatic control, strike arcs against the scrap and burn holes down through the pile. The power is then turned off, and the electrodes are raised slightly and then lowered again with a higher power setting. This is repeated until full power is reached. The arcs are then deep into the pile of scrap, so that all of their energy is absorbed by the pile. This method also produces a pool of steel in the hearth for the arcs to play on, which also expedites melting. If the charge is mostly light scrap, a second bucket of scrap is added to the furnace after the first is almost completely melted. The full charge is melted in from 60 to 120 minutes. At this time it is necessary to reduce the power to 20 to 30 percent of normal rating of the furnace to prevent severe damage of the roof and wall refractories because of their exposure to direct and reflected radiation from the arc. The bath is cold, and the level of carbon is 0.1 to 0.3 percent above the desired final value for the steel. A sample of the metal is withdrawn for composition analysis. The slag contains undissolved lime and is high in iron oxide from oxidation of the scrap that occurred during melting. The operator then adds ore, lime, and fluorspar as required to assist in obtaining a basic, oxidizing fluid slag, and he may also add the necessary alloying elements—nickel, molybdenum, and copper—that do not oxidize. These may be added either as the metal or as the oxide, because the oxides are reduced by reaction with carbon in the bath. The oxygen lance is also employed to speed the refining reactions and to assist in the formation of a fluid slag. About 30 to 60 minutes are required to reach the desired temperature and carbon level in the bath. If a heat of low-carbon steel is being made, the furnace is tilted, and the steel and slag are tapped into the waiting ladle. Additions and deoxidizers such as ferromanganese, ferrochromium, ferrovanadium, and ferrosilicon and aluminum are added to the ladle as the steel is tapped. The full ladle is then moved to the pouring stand, where a train of ingot molds is waiting or where the steel is strand cast (explained below).

Two-slag practice

Many alloy and plain carbon steels that must be manufactured to very high standards of cleanliness and quality are made exclusively in the electric furnace using what is called a two-slag practice. First the charge is melted and refined with the oxidizing practice described above. At the end of the refining period, the oxidizing slag is removed by tilting the furnace back and letting the slag drain into a pot beneath the furnace. Removal of the slag is assisted by rabbling (stirring the surface by hand) to drag as much as possible of the slag out of the furnaces. Some furnaces are equipped with induction stirrers that induce currents in the pool toward the door and thereby carry the slag to the door. Small amounts of graphite, ferro-silicon, and aluminum are then added to the bath to deoxidize and recarburize it. A new slag, formed by adding burned lime, silica sand, and fluorspar to the furnace, contains less than 1 percent iron oxide; it is very basic and very effective in removing sulfur from the metal. Once the slag is formed, the alloying elements, such as titanium, zirconium, and chromium, that normally oxidize readily can now be added to the bath without serious loss. The bath is stirred gently by rabbling or with the induction stirrer, and then the heat is tapped into the ladle. The steel produced in this way is very clean, because the bath is quiet so that the particles of slag and dirt float to the surface, and levels of sulfur below 0.005 percent can be reached.

The level of nitrogen in electric-furnace steels tends to be much higher than in open-hearth or basic-oxygen steels, because the bath is exposed to the atmosphere in the region of the arc. By contrast, there is virtually no nitrogen in the furnace gases of the basic oxygen furnace, and in the open hearth the bath is isolated from the atmosphere by a layer of slag. Electric-furnace steels may contain between 0.006 and 0.012 percent nitrogen. Ordinary carbon steels made with the single-slag practice generally stand at the lower end of this range.

The oxidizing slags employed are similar to basic open-hearth slags, except that the iron oxide content usually is lower at a given level of carbon in the bath. The reason for this is that the oxygen level in the furnace atmosphere is kept low because of the reducing action of the hot graphite electrodes in the furnace chamber. Accordingly, phosphorus removal presents a more serious problem in the electric

furnace than in the open-hearth or basic oxygen furnace. On the other hand, the phosphorus content in the original charge to the electric furnace is usually very low, so the limitation is not a serious one. The high temperature of the arc makes it possible to obtain a very hot slag; it is relatively easy to produce very basic slags that are fluid if needed. This is advantageous when working with the reducing slag, which is low in iron oxide and silica. The ratio of lime to silica in these slags ranges between 2.5 and 4.5. A typical reducing slag has the following composition: lime, 60 to 65 percent; silica, 18 to 22 percent; magnesia, 7 to 9 percent; calcium carbide (CaC_2), 1 to 2 percent; ferrous oxide, 0.5 percent; sulfur, 0.3 percent; and small amounts of alumina and manganous oxide.

Production of stainless steels

The electric furnace is used extensively in the production of steels high in chromium, especially the stainless steels that contain over 5 percent chromium. To economize on chromium, many plants use a large percentage of stainless scrap in the charge. At ordinary steelmaking temperatures around 1,600° C (2,900° F), oxidation of the bath to eliminate carbon to the low levels necessary in many types of stainless steels results in most of the chromium being oxidized into the slag and therefore wasted. By copious injection of oxygen into the bath, however, the temperature of the system can be raised as high as 1,850° C (3,360° F), which favours elimination of carbon and avoids much of the chromium loss. The slag is deoxidized and removed. The bath is cooled to 1,600° C subsequently, and a deoxidizing slag is formed. The heat is finished as described above for the normal two-slag practice. Two alternatives to the use of high temperatures are available to avoid loss of chromium, which also causes extensive wear of the furnace refractories. The first is to melt the charge and oxidize it modestly to reduce the carbon level of the bath without serious loss of chromium. After the bath has been refined, it is transferred to a ladle in a vacuum chamber in which the metal can also be heated by induction. The application of a vacuum and the use of some oxygen gas reduces the carbon level of the bath as desired, and the necessary additional chromium is then added to obtain the desired composition, and the steel is ready for pouring. A second alternative, coming into wider use in the United States and Japan, involves bubbling a stream of argon and oxygen through the metal, reducing the carbon in the bath to the desired level without the need for very high temperatures. This method avoids excessive loss of chromium into the slag.

Vacuum processing

If liquid steel comes into contact with moisture or water, hydrogen from the water may dissolve in the metal in the ensuing reaction. The usual levels encountered range from one to 10 parts per million by weight. On cooling, the solubility of hydrogen in the steel diminishes. If the steel is cooled too rapidly for the hydrogen to migrate to the surface and escape, pockets of hydrogen gas may form at weak points within the structure, and small internal ruptures will result. The problem is particularly acute with very large ingots of electric-furnace steel cast for big forgings used in turbines and electric generators, and is also generally present with alloy steels containing from 2 to 10 percent nickel used for smaller forgings. The presence of these ruptures make a casting unusable. Moisture in the furnace additions, moist runner and ladle refractories, and moisture on the ingot-mold walls contribute hydrogen to the steel. By carefully drying all additions and refractories and by cooling the ingots very slowly in special furnaces, the problem of hydrogen in forgings can be minimized but not eliminated.

A number of techniques have been devised in which the liquid steel is exposed to a vacuum in a chamber, and the hydrogen is pumped out to a level ranging from 1.5 to two parts per million. If the steel then does not come into contact with moisture until it is solidified, the problem of internal rupturing can be avoided. One method of preventing such contact is to place the ingot mold in a large steel chamber connected to vacuum pumps. The chamber is evacuated, and the ladle containing the steel is set on top of the chamber. The nozzle is sealed to a fitting closed with a thin metal disc. When the stopper rod is raised, the steel stream cuts through the disc, and the stream and the steel in the mold are exposed to the vacuum. Several alternative methods are also employed.

Vacuum processing is also used to eliminate oxygen from liquid steel. Carbon and oxygen react to form carbon dioxide gas at a low pressure maintained by the vacuum pumps. Under this condition carbon itself acts as a very strong deoxidizer, and no products of deoxidation are left to contaminate the steel. With a good vacuum and a clean system, the level of oxygen dissolved in the steel can be reduced to as low as 10 to 20 parts per million, a level that may not be reached with strong deoxidizers such as aluminum.

Solidification and casting. The quality of the surface and interior and the physical properties such as strength, resistance to impact, and corrosion resistance of a steel product are influenced by the way the liquid steel is solidified. In the usual practice, molten steel is poured into a cast-iron mold from a hole in the bottom of the ladle, the hole is closed by a stopper on a rod extending up through the metal, and the rod is protected from high temperatures and action of the liquid metal and slag by a refractory sleeve. A recent development is a sliding refractory valve with all parts outside the ladle and below the nozzle. Because the cast-iron mold is colder than the liquid steel, solid steel forms as a shell on the mold wall and bottom. Freezing proceeds inwardly from the wall as heat from the hot steel flows into the mold.

The refined steel in the furnace contains an appreciable amount of dissolved oxygen, as described above. As the liquid cools and freezes, carbon in the steel and dissolved oxygen will combine to form a large volume of gaseous carbon monoxide. If the evolution of this gas is uncontrolled, the metal will boil up and overflow from the mold, and the steel ingot will have a porous, spongy interior. This phenomenon caused serious difficulties in the early days of the Bessemer process until it was discovered that adding manganese to the steel to combine with the oxygen would prevent formation of the gas, so the ingots were sound and satisfactory. Subsequently, it was found that other elements, such as silicon, aluminum, titanium, and calcium, all of which form stable oxides, could also be used to deoxidize the steel. Furthermore, as liquid steel solidifies and cools, its volume shrinks 7 percent. Finally, carbon and many alloying elements and impurities such as sulfur and phosphorus tend to migrate, or segregate, into the unfrozen liquid during solidification, with the result that the final bit of liquid to freeze is much more impure than the portion frozen first.

Types of ingots and their manufacture. The internal structure of an ingot can be controlled by regulating the rate at which carbon monoxide gas is formed during solidification. If all the oxygen dissolved in the steel is removed by the addition of a deoxidizer, no gas will be generated, and a shrinkage cavity will form in the upper central portion of the ingot. This ingot is a killed ingot (Figure 16A), and the cavity is called "pipe." The steel immediately surrounding the lower portion of the pipe is relatively impure because of the segregation of impurities and carbon. As a result, the upper portion of this ingot is unfit for use, and as much as 30 percent must be discarded. This loss can be minimized by casting the steel in tapered molds with the large end up. Another method is to install a brick or clay reservoir on top of the ingot mold that maintains a small pool of liquid metal during freezing of the body of the ingot. The metal in this "hot top" flows down to compensate for shrinkage of the metal in the mold. The top of the ingot is cropped and discarded; loss is only 10 to 15 percent. Killed steels are employed when a dense, uniform product is desired, as is the case with alloy steels used in forging, tubing, and plates for pressure vessels. Most high-grade alloy steels are killed steels. A "semikilled" ingot (Figure 16B) results when the addition of the deoxidation agents is limited to allow the generation of a small amount of gas during the latter stages of freezing. If properly controlled, this gas forms small, well-distributed blowholes to compensate for the shrinkage of the metal, and no pipe appears. These blowholes collapse and weld shut when the ingot is rolled. Since it is not necessary to discard the top portion of the ingot, semikilled

Minimizing loss of steel

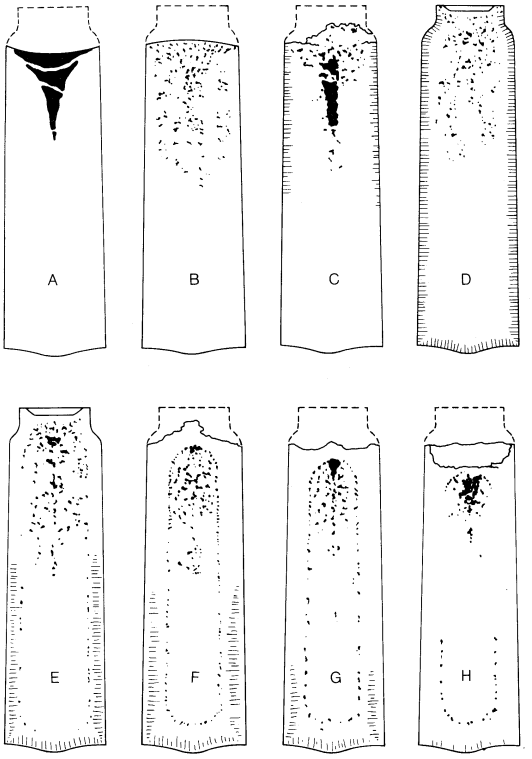


Figure 16: *Ingot structures.*
(A) Killed; (B,C) semikilled; (D,E) capped or bottle top; (F,G,H) rimmed (see text).
From G. Derge, *Basic Open Hearth Steelmaking*

steels are less costly to produce than killed steels. Semikilled steels are used for structural members, rails, some types of plates, welded pipe, and ordinary bars and common sections, such as angles and tees. Ingot C in Figure 16 is an example of a poorly controlled semikilled ingot in which pipe formed, gas pressure ruptured the top, and metal bled out. With a greater amount of oxygen in the steel when it is poured, it is possible to cause small bubbles of carbon monoxide to form at the interface between the liquid and solid metal. These bubbles grow, detach themselves, and rise toward the top of the ingot. In rising they sweep the liquid along and generate a flow of metal up the walls and down in the central portion of the ingot. If the circulation is gentle, blowholes will be present in the solidified rim, and if it is very vigorous a dense rim forms. After a period of rimming, the top of the ingot is sealed, and the resulting interior structure is similar to that of a semikilled ingot. Ingots D through H in Figure 16 show the types of structures that can be obtained. With ingot D and E, the rimming action is modest, and the level of the liquid rises because of the large volume of the blowholes in the rim. Within one to two minutes after the mold has been filled, the steel rises and freezes against a heavy cast-iron or steel cap placed on top of it that seals the ingot and stops evolution of the gas because of a rapid rise in internal pressure. This type is called a capped ingot, or a bottle-top ingot, because of its shape. Ingot E is a better structure, because the rimming action was stronger, and the blowholes in the rim are farther from the surface of the ingot. Plain carbon sheet steels containing between 0.04 to 0.20 percent carbon are often cast as capped ingots. Ingots F, G, and H are called rimmed ingots, because the oxygen level in the liquid steel is high enough to sustain a strong rimming action for from 12 to 16 minutes that produces a rim from 75 to 100 millimetres (three to four inches) thick. After a rim of the desired thickness is formed, the ingots are usually sealed by placing a heavy steel plate on top of the liquid surface, and the top of the mold is flooded with water. Of the three, ingot G has the best structure for most purposes. In practice, the rimming behaviour is controlled by adding gas-forming agents to the steel to stimulate the action or small amounts of aluminum to reduce the action. Plain carbon steels containing between 0.04 and 0.25 percent carbon can be rimmed successfully. This type of steel has a soft, pure rim and is used when good surface, high ductility, and good formability are desired. Large tonnages of the product are used in automobile bodies.

Control of nonmetallic particles

In addition to controlling the structure, control over the nature and composition of nonmetallic particles in steels is needed. In most steels it is desired to eliminate the particles formed by the reaction of the deoxidizer with the oxygen. Furthermore, small particles of slag may be emulsified in the steel, and bits of refractory may be entrained as a result of the flow and churning of the metal when the heat is tapped from the surface. Most of these particles float to the surface of the metal and combine with the slag in 10 to 30 minutes between the end of tap and the beginning of pouring of the ingots. Not all do, however, so many special deoxidizers and pouring methods are used to keep the remaining particles small and well distributed throughout the ingot. Free machining steels are exceptions where many small uniformly distributed particles of certain types are desired. These particles lubricate the cutting tool, prevent welding of the cut chip to the tool tip, and cause the chip to break into short pieces. Additions of lead, sulfur, manganese sulfide, selenium, and tellurium to the steel improve the machining properties of steels in this manner.

Variation of ingot size

The size of ingots varies enormously depending upon both the type of steel and its ultimate use. Ingots of highly alloyed tool steels may weigh only a few hundred pounds each. Steels for structural, bar, and rod products usually weigh from 1,800 to 7,300 kilograms (two to eight tons). Ingots from which slabs are rolled for large plates and large coils of sheet steels may weigh from 13,600 to 36,300 kilograms (15 to 40 tons). These larger ingots are blocks of steel that may be 2.3 metres (7.5 feet) tall by 2.1 metres (6.7 feet) wide and up to one metre (three feet) thick. Very large ingots for large, thick plates and forgings for turbine shafts, drive shafts, and pressure vessels may weigh as much as 272,000 kilograms (300 tons) each; these are cast in enormous cast-iron molds and take weeks to solidify. Forging ingots are from three to five times thicker than the thickness of the product forged or rolled from it to ensure adequate working of the steel. A greater reduction in thickness is desired for plates, structural products, rails, and most other products.

Ingot casting

In large steelmaking shops, two or three ingot molds are carried on a small ingot car; several cars are connected together so that one heat is carried on a train. These ingot molds are filled from the top. If particularly good surface quality is required, ingot molds may be filled from the bottom. Usually a number of molds are stood in a cluster on a large car. The steel is poured from the ladle into a tall, slim, trumpet-like tube. The metal flows from this trumpet through hollow clay brick to the bottom of each mold, filling all simultaneously and more slowly, with less splashing, than is the case with top-pouring single molds. After the ingots have cooled sufficiently, the train of cars is moved to a stripping device, which seizes each mold and frees the ingot from the mold by the action of a powerful plunger. The molds are set aside for cleaning and reuse, and the train of hot ingots is moved to the rolling mill. There they are inserted in a reheating furnace, where they are "soaked" for four or five hours to bring them to a uniform temperature in preparation for rolling. These furnaces are often called soaking pits. Six to a dozen ingots are stood on end in each pit; a gas or oil flame passing above the ingots supplies the heat. The air for combustion is heated in regenerators similar to those described for the open-hearth furnace. The furnace is covered, and a slightly oxidizing atmosphere is maintained so a crust of oxide forms on the exterior of each ingot. Care is taken so the ingots are not heated above the temperature at which any part of the internal structure becomes liquid, because the presence of liquid will result in severe cracking and rupturing of the steel during rolling.

Strand casting. Rapid progress is being made in the use of methods for casting steel in a long, continuous bar, or strand. Once called continuous casting, this method is now called strand casting. The process has long been

used in casting nonferrous metals such as copper and aluminum, but many problems peculiar to steel had to be solved. Many highly specialized machines, control methods, and deoxidation methods have been developed for the process. Only the essential aspects of the method need be dealt with here. It has been possible to cast square and rectangular sections of steel for a number of years, but the problems of surface quality, internal structure, and reliability of operation when casting slabs from 150 by 760 millimetres to 280 by 2,000 millimetres (six by 30 inches to 11 by 80 inches) have been resolved only since 1968. Facilities for strand casting the whole output of large basic-oxygen shops and electric-furnace shops are being installed in plants in many countries in the world.

Advantages of strand casting

There are two important advantages of strand casting over ingot casting. A greater yield of useful product from the liquid steel is obtained, because the large pipe in the cast product is avoided and less steel is lost by cropping the ends of the bar of steel during rolling. Also the structure and composition of the bar is uniform along its length, which is not so with a bar rolled from an ingot.

The essential characteristics of the strand-casting system are shown in Figure 17. The steel is poured from the ladle in the customary manner. It flows into a shallow refractory-lined basin (tundish) and then through a nozzle into the water-cooled mold. Flow into the mold is regulated by a stopper in the tundish. The tundish is used to permit precise regulation of the flow of metal into the mold and to allow additional time for nonmetallic particles in the steel to float out. Several molds are usually fed from one tundish. In recently built systems, the metal from the tundish is conducted through a ceramic tube and is discharged below the surface of the metal in the mold. This avoids splashing and further oxidation of the metal by contact with the air. The surface of the metal in the mold may also be covered with a powdery layer of material used for insulation. The layer in contact with the metal melts to form a thin slag, which dissolves particles of oxide that float on the metal and also coats the bar of steel as it is cast. Heat is extracted from the liquid steel by the water-cooled mold; a thin layer of solid steel forms on the mold wall. In most of the strand-casting machines, the mold oscillates rapidly up and down to aid in the separation of the metal shell from the mold wall. Below the mold the steel bar is sprayed vigorously with high-velocity streams of cold water to speed solidification.

From J. Szekely and N.J. Themelis, *Rate Phenomena in Process Metallurgy*, copyright 1971; used with permission of John Wiley & Sons, Inc.

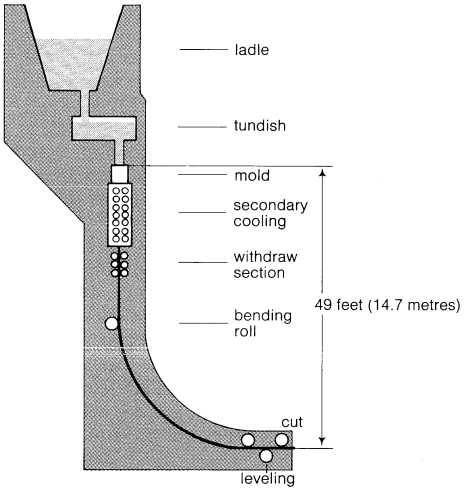


Figure 17: Continuous casting machine.

The casting operation is started by inserting a large steel bar of the cross section to be cast up into the mold. The liquid metal is poured into the mold and freezes around a large projection on the bar. This bar is withdrawn at the desired rate, and the flow of metal into the mold is regulated to maintain a constant level. The steel bar is supported by several sets of rolls after it emerges from the mold, and some of the rolls are driven at a controlled rate to regulate the speed of withdrawal of the bar. Considerable time is required to complete solidification, and the tapered liquid core may be 6.3 to 13 metres (21 to 43 feet) long. The bar is cut into lengths by a saw or an oxygen torch below the tip of the liquid region. The speed of casting is very fast for small bars: *e.g.*, 10 by 10 centimetres or 15 by 15 centimetres (4 by 4 inches or 6 by 6 inches). Casting speed ranges from one to two metres per minute (three to six feet per minute) for wide slabs that are 230 to 280 millimetres (nine to 11 inches) thick. To reduce the height of the casting machine, some machines bend the bar 90° in the cooling section (Figure 17) so that it comes out horizontally.

Other machines have a curved mold that casts a curved bar, in which case the bar is straightened as it emerges from the cooling sections and is horizontal as it leaves the machine. Strand casting avoids the need for large, expensive mills for rolling ingots.

Rolling and working of steels. *The rolling process.* The mechanical properties of a cast steel bar or ingot are improved materially by kneading and working the piece in a press, hammer, or roll stand when it is red-hot and plastic. The nonuniform, coarsely crystalline cast structure is converted to a uniform structure having a fibrous appearance when fractured; there are usually attendant increases in toughness and ductility. The steel is also altered to the desired form. During the working the steel is kept sufficiently hot for the small grains, or crystallites, produced to regrow to form a new structure. Massive gun tubes, drive shafts, and turbine rotors are forged slowly in enormous hydraulic presses. Axles, crankshafts, and heavy engine parts are forged by large steam hammers, while tools and small machine parts are worked by rapid-acting drop hammers.

The rolling mill

Most steel ingots and strand-cast bars are formed in rolling mills. As noted above, rolling mills vary widely in mechanical details depending upon their purpose. The working and shaping of the red-hot piece of steel is accomplished by passing it back and forth through one stand or by sending it through one stand of rolls after another. A bar has its cross section reduced and a sheet is made thinner with each pass through a mill. The rolls of a stand may be grooved to produce a bar with a particular cross section, such as a square bar, a rail, or a beam, or may be cylindrical for rolling flat pieces, such as strips or plates. Though a mill may consist of only one stand, more often several are employed, either in tandem (the steel passes straight from one stand to the next) or side-by-side (the steel bar is looped around to go from one pass to the next). In the latter case, the rolls of each stand may have a number of sets of grooves so that a bar may pass through a stand several times. In the past, a wide variety of arrangements of stand locations have been used in mills.

Modern mills usually have the stands in tandem; the steel passes directly and rapidly through them in a straight line from heating furnace to cooling beds or coiler. The products of rolling mills vary widely in weight and shape, but, in virtually all cases, each piece is uniform in cross section from end to end. Special mills are used to roll steel pieces that vary in cross section along their length. All modern mill stands are driven by powerful electric motors, and the mills are controlled by electronic equipment.

Primary mill

The first reduction of an ingot takes place in a "primary" mill, which may be called a blooming mill if it converts ingots of approximately square cross section into large square or rectangular bars, called blooms. The mill has two large, grooved rolls driven by electric motors of as much as 30,000 horsepower, and the ingot is passed back and forth as the mill is reversed. In each pass the spacing between the rolls is reduced so the bar is reduced in cross section and correspondingly lengthened. The grooves in the rolls also permit rolling blooms with special shapes to be converted into structural shapes, sheet piling, and other forms. This type of mill is called a reversing two-high blooming mill. Some blooming mills are three-high (Figure 11) and roll only in one direction. The ingot can be passed back and forth by alternately using the opening between the middle and upper rolls and then that between the middle and lower rolls. A two-high reversing mill is

often used to roll large ingots into slabs for plates and for the production of coils of sheet steel. In a modern steel plant these larger ingots may also be rolled into slabs on a universal mill, which has two stands. One of these stands is a two-high reversing mill for rolling the flat faces of the slab. Adjacent to this stand is another mill, which simultaneously rolls the edges of the slab. Such an installation is very fast and can easily roll 4,000,000 or 5,000,000 tons of steel a year. Each mill is equipped with soaking pits to heat the ingots before they are rolled and a shear to crop off the ends of the slab and to cut it into appropriate lengths. Many mills are equipped with a scarfing machine that burns off one to three millimetres (0.04 to 0.12 inch) of each surface of the slab or bloom to obtain a surface free of defects. The burning is accomplished by many jets of gaseous oxygen impinging on the surface of the hot steel.

Second stage of rolling

The blooms and slabs produced by the primary mills receive additional hot work in another mill to convert them into hot-rolled products or to process them further in preparation for cold rolling. The bars and slabs produced by strand casting are introduced into the rolling sequence at this point. Usually the steel is not hot enough to be rolled directly after it leaves the primary mill. Also, it is usually desired to let the steel cool down so it may be inspected and defects on the surface cut out by machining, chipping, or scarfing. The steel is then heated in reheating furnaces, fired by fuel oil or gas. The furnace hearth is inclined and is as wide as the blooms or slabs are long; the pieces of steel are pushed sideways into the furnace. The charge rests on water-cooled skids, and, as one cold piece enters, a hot piece ready for rolling is forced out the other side of the furnace.

Rails, bars, and structural steel

Rolled and worked materials. Rails, I beams, H beams, sheet piling, angles, large rounds, and similar products are rolled on a mill having stands with grooved rolls to shape the piece to the desired form. Since special rolls are required for each type of section, the stands are designed so the rolls may be quickly replaced. In modern mills, whole stands may also be changed, inserted, or removed when there is a change in the product being rolled. In a common rolling sequence, a reheated bloom receives an initial pass in a special stand called a scale breaker to remove the scale of iron oxide formed during reheating. The bloom then receives more passes in a two-high reversing stand with grooved rolls. The bar is then brought to its final shape by a number of passes in several subsequent stands. If 12 or more passes are required, it may be necessary to reheat the steel in a second furnace at an intermediate stage. Very close attention is given to the final passes to ensure that the final product is correctly shaped. The hot pieces of steel are sheared to length and cooled on tables. With alloy steels, it may be necessary to control cooling or even to cool slowly in a special furnace. The bars, rails, and beams are then straightened, trimmed, and given a final inspection. For H beams special, massive stands with four rolls arranged to roll the faces of the flanges and each side of the web of the beam are used simultaneously.

Mills used for rolling small, simple sections such as round and square bars, angles, and rods also vary widely in the design of the mill and the type of stand used. Blooms may first be reduced to obtain billets having a cross section of under roughly 10 centimetres (four inches) square. These billets are then reheated for further rolling. Economical operations require high speed; the most modern mills are continuous, with the stands set in a line. The first several stands are spaced apart sufficiently so the billet is in only one of them at a time, and subsequent stands are closely spaced so the steel passes continuously through all stands simultaneously. To avoid buckling of the steel bar between stands, the speed of the stands must be coordinated and precisely controlled. The steel emerges at a speed of a few thousand metres a minute and is either cut to length by a flying shear and cooled on beds or coiled in automatic coilers. These continuous mills may have several of the stands set with their rolls at right angles to the others, or square bars may be twisted 90° as they pass from stand to stand to work all corners of the bar and avoid corner cracks. Some continuous rod mills are equipped to roll four to six bars simultaneously. A very wide range of alloy compositions and types of bars and rods are rolled, and a mill has extensive facilities for inspecting, cutting, straightening, surface preparation, heat-treating, storing, packaging, and shipping.

Plates

The plate mill appears to be very simple in comparison to a structural or rail mill. In some mills, however, the stands and equipment must be massive to be able to roll plates of alloy steels up to five metres (16 feet) wide and smaller plates that may be up to 40 centimetres (16 inches) thick. The mills for producing large plates will be described only briefly here. Long plates less than approximately three centimetres (1.2 inches) thick and two metres (80 inches) wide are usually produced on a continuous-type mill, as described below. The production of large plates is a specialized operation because of the many different alloy compositions and sizes that must be made. The mill consists of a storage yard for slabs, several heating furnaces, a scale-breaking stand, the main mill stand and its drive system, run-out tables, a shear, a straightener, furnaces for heat-treating and annealing, and an area for inspection, cutting, trimming, and finishing of the plates.

The main mill stand is usually a four-high reversing mill in which two smaller cylindrical rolls are the work rolls. Slabs not more than approximately two metres wide can be produced by the primary mill. When making wide plates, it is necessary to give the slabs several passes in its long dimension and then turn the slab 90° and cross-roll it in a few final passes to increase its width. This sequence is required to obtain the desired width, because, in passing through the mill, a slab or bar increases its length but changes its width only slightly. It is found that the ductility of steel is greater in the direction of rolling than it is at right angles to this direction. Hence, cross-rolling is required to prevent the development of the highly directional properties that result when steel is rolled only in one direction.

Continuous strip mills

Much of the steel from the large integrated steel plants in the world today is produced in the form of long, wide strips, either flat or wound in coils. One large coil usually is obtained from a single slab that may be as much as 250 millimetres (10 inches) thick and 11 metres (36 feet) long and may weigh up to 40 tons. This slab may be strand cast or rolled from a single ingot. Sheet steel for skelp (pipe production), narrow, lightweight plates, and narrow and wide continuous-strip steels are produced on the hot strip mill. Such a mill consists of several heating furnaces, a scale breaker, four roughing stands, a shear, six finishing stands, a run-out table, and coilers. The roughing stands are equipped with edge rolls to control the width of the strip. Except for the scale breaker, the stands are four-high with smaller work rolls backed by two massive rolls to stiffen the work rolls. Each stand of a wide strip mill is driven by an electric motor rated at 3,000 to 6,000 horsepower. The scale breaker is a two-high stand followed by sprays of very high-pressure water to carry away the scale loosened by the light reduction of the slab as it passes through the stand. Removal of all scale is essential if a sheet with a surface of good quality is to be obtained. All of the mill stands are set in a line, and the roughing stands are set sufficiently far apart so the slab is in only one of them at a time. The mill may have a slab squeezer, which is a large press located after the first roughing stand to flatten the slab and true up its edges. On leaving the last roughing stand, the strip is about 25 millimetres (one inch) thick. After its ends are squared in a shear, it passes on into the finishing train. The long strip of steel is in all finishing stands simultaneously; the openings of the rolls in the stands are set to obtain the desired final thickness of the strip. These settings and the speed of the stands are controlled automatically. It is remarkable that the stands are only about five metres (16 feet) apart and the strip is travelling a few thousand metres a minute as it leaves the last stand. The temperature of the steel must also be carefully controlled to obtain a satisfactory product. The strip then is cooled by water sprays, after which it may be coiled. The thicker strips may be transferred for cooling to another roller table, where they are cut up into long, narrow plates. The final thickness of the hot-rolled strip produced on the continuous hot mill ranges from three

to 10 millimetres (0.1 to 0.4 inch). A large modern mill producing strips up to two metres wide can roll 4,000,000 to 6,000,000 tons of steel a year.

The product from the continuous hot strip mill has a tightly adhering oxide coating, or scale, covering its surface. This scale may be desired on light plate and sheets taken directly off the mill. For some products, however, the sheets may be pickled to remove the scale and then coated with oil. Pickled products may be sold flat or in coils; wide strips may be split into several narrow strips suitable for making such items as pipe. Much of the product from a wide continuous hot strip mill must be cold rolled to obtain the desired surface quality and physical properties. The coils are first pickled continuously in long tanks to remove the scale. They are then cold reduced 40 to 90 percent in thickness at high speed in powerful mills consisting of four to six four-high stands that are set closely spaced and in tandem. During rolling the strip is kept under tension and driven reels pay out and take up the strip. The product of the mill may be annealed (heat-treated to remove internal strains) in batch furnaces or may be uncoiled and processed continuously through a long annealing furnace. Often coils of either hot- or cold-reduced steel are rolled on single- or double-stand four-high mills to obtain precisely controlled physical and mechanical properties and also a special surface finish. The product is oiled and inspected and may be sold in coil form or may be cut up into sheets. Many steel plants process cold-rolled strip steel into coated products. Coils of tin-plated steel are produced continuously by passing the steel through a bath in which a coating of tin is electroplated onto the surfaces, and this coating is fused to obtain a bright surface. Galvanized sheet is produced continuously by passing the strip through a long furnace to reduce any oxides on the surface and to soften the strip and then passing it into and out of a tank of molten zinc. After cooling, the familiar bright, spangled galvanized steel is obtained. Many other types of coatings are also applied to steels.

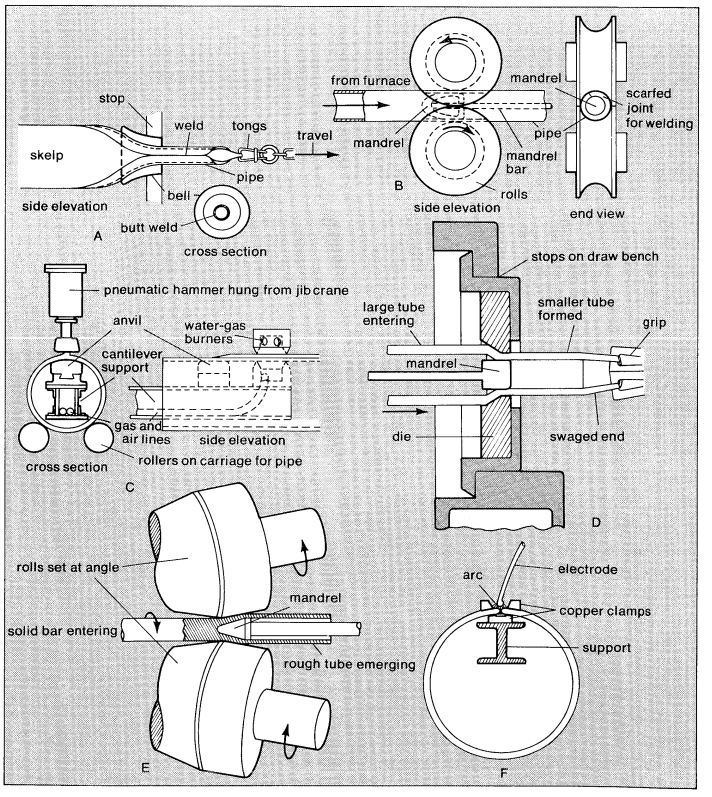


Figure 18: *Various methods of manufacturing pipe.* (A) Butt welding a five-centimetre (two-inch) pipe; (B) lap welding; (C) hammer finishing and gas welding of 76-centimetre (30-inch) pipe; (D) drawing seamless pipe; (E) Mannesmann method; (F) arc welding of 76-centimetre pipe.

Pipe

Pipe in all sizes from six millimetres to 1.3 metres (0.25 to 50 inches) in diameter is produced from flat-rolled strip and plate steels. The methods of production vary widely; some of the more common are illustrated in Figure 18A, B, C, and E. The smaller sizes are produced from narrow strips called skelp by method A. Lengths of skelp six metres (20 feet) long are pointed at one end and are heated flat in a furnace to a white heat. At this temperature the steel is coated with a shiny layer of molten iron oxide and welds easily under pressure. The strip is drawn rapidly through a welding bell, which simultaneously forms the strip into a tube and forms a butt weld. Smaller sizes of pipe, with inside diameter ranging from three to seven millimetres (0.1 to 0.3 inch), are made by lap welding (B). The edges of the strip of skelp are first bevelled in a set of rolls, and then the strip is formed into a tube by pulling it through a ring. The tube is then heated to the welding temperature and rolled over a mandrel (core) by a special set of rolls, as shown. The pressure on the overlapping edges from the roll and mandrel causes the steel to weld. The butt-welded or lap-welded pipe is straightened, descaled, cut to size, and threaded, if required, tested, and coated with oil, paint, or plastic coating. The steel used in making these pipes is plain carbon steel containing not more than 0.1 percent carbon.

Large-diameter pipe is produced by forming long plates into tubular form in special presses. First the edges are bent slightly, then the strip is formed into a U shape, and finally the U is bent into an O. The weld is made either by gas (C) or electric (F) means, as shown. After welding, the pipe may be rerolled for shape, heat-treated or annealed if required, cut to size, pressure tested, and given a protective coating. In some plants, pipe of many sizes is produced continuously from strip by electric resistance welding. The strip is formed cold into a tube, with the edges tightly pressed together. Welding is accomplished by passing a heavy current across the joint, which rapidly heats the metal to the welding temperature. The current is supplied by two copper rollers, one on each side of the joint. Because high strength, good corrosion resistance, and other special properties are desired in large pipes, they are made of intermediate carbon and alloy steels.

Seamless tubing

Seamless tubing is usually employed if expected internal pressures are high, service is severe, and reliability is of great importance. Tubing for nuclear reactors, oil wells, and high-pressure chemical equipment are examples. The diameter of the finished tube may be up to 450 millimetres (18 inches), with wall thicknesses as high as 150 millimetres (six inches). The tube is made from a round billet with a total weight of steel that is slightly greater than that of the tube to be made. The billet at a red heat is first pierced by being rotated rapidly between two rolls whose axes are set at a slight angle to each other (Figure 18E). The rotation sets up forces in the billet, causing the centre to rupture. The action of the rolls also forces the billet to advance slowly over a mandrel set between the two rolls. The result is a length of rough tubular blank. After piercing, the tube is reheated and the inner and outer surfaces are smoothed by rolling it over a mandrel as in lap welding (Figure 18B). Subsequently, the tube is straightened, trimmed, and inspected. Tubes may be reduced in size and their strength improved by cold drawing over a mandrel held in a die, as shown in Figure 18D. If a large reduction in diameter or wall thickness is desired, the tubes may be annealed several times during the processing. Tubes produced in this way are certain to be of good quality as only the best quality steel can be pierced successfully. Seamless tubing and other complex shapes are also produced by extruding a hot billet of steel through a die in a powerful hydraulic press. To produce a tube, a billet with a hole down its axis is placed in the cylinder of the press, and the steel is forced to flow out the annulus between a die and a mandrel attached to the piston of the press. The billet is contained in a woven glass sock when it is placed in the press. This sock melts, and the liquid glass acts as a lubricant between the billet and the cylinder of the press. This method was developed in France. Extruded tubes are then processed similarly to the method used for pierced tubes. Steels can also be extruded cold.

Wire and cold-drawn rods

Steel wire is produced by drawing a hot-rolled rod continuously through a number of dies. The reduction in each die ranges from 10 to 40 percent. The rod is first pickled to remove scale and is then given a light coating of lime, phosphate, or copper for lubrication. During the drawing, the dies are also lubricated with soap. Because the steel is drawn cold, the wire becomes stronger and harder with each draw. To prevent breakage, the wire is annealed frequently. Wire is produced in a large number of steel compositions, diameters, stiffness, or tempers, and coatings, depending on its end use, which may be in steel cables, coil springs, nail fencing, or household screening. Hot-rolled rods may be cold reduced slightly by drawing them through several dies. This improves the surface finish and increases the strength of the steel.

Control of properties. Satisfactory production of steel requires that the temperature, composition, and properties of the metal and the processing methods be kept under close surveillance at every step. The compositions of iron ores, sinter, pellets, hot metal, blast-furnace slag, liquid steel, and steel-furnace slag are determined from samples taken during processing. Analyses are obtained by conventional chemical methods and with the use of emission spectroscopy and X-ray fluorescence. The gas content of steel is determined by collecting and analyzing the gases obtained when a sample of steel is remelted under vacuum or in a stream of argon gas. Temperatures are measured by thermocouples inserted in the furnace or in the liquid metal or by optical devices sighted on the steel or into the furnace.

Measuring steel properties

The composition and properties of the solid steel are measured at many stages in processing, from the ingot and strand-cast bar or slab to the final hot- and cold-rolled products. The quality of cast and hot-rolled sections of steel is indicated by the appearance of a fractured surface and by the texture of a cut surface as revealed by etching with acid. Traditionally, the grain size and toughness of a steel bar are judged from the appearance of the fractured surface of the steel. A trained eye is needed for this method to be effective in controlling the production of steel; it has been superseded by many newer and more quantitative techniques.

Sulfur print

The general structure and texture of a roughly polished surface cut through a rolled or forged piece of steel is revealed by etching the surface with a hot solution of hydrochloric acid. Lines of flow of the steel from hot working, areas of segregation, and collapsed blowholes and pipe can readily be seen. The distribution of sulfur in the cast or rolled structure can be observed by making a sulfur print of a smoothly ground surface of the steel. A sheet of photographic paper moistened with sulfuric acid is held in contact with the steel for approximately one minute. On being developed, the print shows the texture of the steel and the locations where sulfur is segregated. The structure of hot- and cold-rolled steel products is observed under the optical microscope to determine if the processing has produced the texture and phases necessary for the properties desired of the steel. Many test methods have been developed and are used regularly to determine the hardness, tensile strength, ductility, toughness, and drawability of the steel during processing. These measurements and metallographic examination of the steel are essential in controlling the quality and properties of the steel being made.

Design of the furnace. The size of a furnace used in heating steel for processing depends very much upon the nature of the product. Furnaces for heat-treating tool steels are small; the batch furnaces for annealing coils of strip steel are refractory-lined boxes of sufficient size to contain six or eight coils in a charge; and furnaces for heating and annealing strip steels and steel parts on a conveyor are block-long horizontal tunnels or may have several vertical sections through which the steel is looped. These furnaces have zones for heating, cooling, and holding the temperature constant so that the steel receives the desired heat-treating cycle automatically. Much attention has been given recently to the design and construction of furnaces of this type for many purposes. Gas, oil, and electricity are used as sources of heat. If reactions between the steel and combustion gases causing carburization, decarburization, scale formation, or pickup of sulfur must be avoided, electrical heating is often used. Furthermore, the steel can be isolated from combustion gases by burning fuel oil or gas within tubes of heat-resistant steel or silicon carbide that pass through the furnace. Heat is transferred from these tubes to the furnace charge by radiation and convection of hot gases within the furnace. Mechanical details of a furnace also depend upon the nature of the charge. Furnaces used for heat-treating and annealing large forgings will have large manipulators on wheels or an overhead crane for moving the steel to the furnace and to and from the press, forge, or hammer. Furnaces for treating many steel parts may resemble a tunnel the bottom of which consists of refractory-covered trucks mounted on small wheels that travel on rails. The charge is placed on these trucks, which are pushed through the furnace at the desired rate. Bars and billets may be moved through long or circular furnaces on rollers or conveyors. To conserve heat, many high-temperature furnaces are equipped with recuperators or regenerators for preheating the combustion air by waste heat extracted from the exhaust gases.

Furnace atmosphere. The atmosphere within a furnace is often important to the purpose of the furnace. Soaking pits for heating ingots and the reheating furnaces on primary mills are usually operated with a strongly oxidizing atmosphere to obtain a dry, brittle oxide coating on the steel. This is to assure that the oxide flakes off easily early in the rolling process so as not to blemish the surface of the finished piece. When heating strip steel or annealing tool steels for long periods, the surface must be preserved, so a nonoxidizing atmosphere is obtained by careful control of the ratio of water vapour to hydrogen and the ratio of carbon dioxide to carbon monoxide in the furnace atmosphere. In bright annealing at comparatively low temperatures (650° C [1,200° F]), a small excess of carbon monoxide over carbon dioxide in the atmosphere is employed to obtain a bright metallic lustre on the steel surface. Moisture content of the gas must be kept low, however, to prevent oxidation of the surface during the cooling cycle. Atmospheres that protect the steel from oxidation during heating to higher temperatures must be much more strongly reducing; these are obtained by combustion of natural gas or other gaseous hydrocarbons with a deficiency of air to produce the desired gas composition. Regulation of the ratios of carbon dioxide to carbon monoxide and water vapour to hydrogen permit a gas to be used to add or subtract carbon from the steel. If the steel is to be case carburized (externally hardened by addition of carbon), methane (CH_4) is added to the gas. The addition of both methane and ammonia (NH_3) produces a hard surface on the steel that is high in both carbon and nitrogen. The composition of the gases can also be adjusted in many ways: by partial combustion of a fuel gas, as mentioned earlier, and by cooling a partially burned gas and absorbing carbon dioxide from it or by chilling it sufficiently to condense out most of the moisture.

Non-oxidizing atmosphere

Cooling rate. Rapid extraction of heat from steel to cool it quickly from one temperature to another is important in developing certain desired properties. The cooling rate of the steel is determined by the nature of the quenching medium and the thickness of the piece being treated. An enormous variety of liquids has been employed since antiquity, but relatively few are now utilized. Often used mediums, in order of decreasing severity of quench, are cold water from high-pressure jets onto hot steel, a bath of iced brine, a bath of circulating cold water, a bath of oil (the oil must have a high flash point to be safe), a bath of boiling water, circulating liquid lead, a blast of air, still air, burial in a pit of lime, and slow cooling in a furnace or refractory-lined box.

Quenching media

The cooling rate required to harden a steel is reduced with increased levels of carbon and the presence of alloying elements. Hence, a less severe quench is used with alloy steels. With a given quench medium, however, the cooling rate in the interior of a piece of steel decreases with increased thickness of the steel. Thus, the greater the thickness of the piece, the more severe the quench must be to harden the steel throughout. With carbon steels,

air cooling may be sufficient fully to harden thin strips, but water quenching may be inadequate for heavy gears and forgings. Accordingly, alloyed steels are used for the heavier sections both to obtain hardening throughout and to avoid the severe quench rates that cause distortion and cracking. The quenching medium usually has its temperature controlled by a cooling system and is circulated to assure uniformity of quench from piece to piece and over the surface of each piece. Mechanical devices and conveyors are employed to save labour and provide rapid and consistent introduction of metal parts into and removal from the quenching medium. Uniformity and consistency of the quenching operation are essential to obtaining good control of the parts being processed.

Metallography. The nature and structure of a steel can be observed by examining the polished and etched surface of the piece of steel in an optical microscope. The sample is given a high polish and then etched carefully to differentiate the various phases present. The choice of etchant is determined by the composition of the steel and the aspect of its structure under investigation. The magnification employed is also determined by the purpose of the investigation and ranges from one to 1,500 times. Since the latter figure is the approximate limit of ordinary light microscopy, electron microscopy is used if greater resolution is desired. The structures observed may be photographed for further study and to provide a record of the structure of the steel. Considerable attention is being devoted to the use of closed-circuit television and a computer to determine automatically the nature and amounts of various phases present in an image from the microscope. In the long run, this technique should reduce the tedious work of identifying and counting the nonmetallic particles present in steels, but more work is needed to make it practical.

There is a close connection between the heat-treatment given a steel and the structures obtained (see below *Steel metallurgy*). Thus, the microscope is important in determining whether the processing was satisfactory. Microscopic examination of a steel, coupled with measurements of tensile strength, ductility, fatigue strength, impact resistance, and other special physical measurements, is used to establish control of processing of all kinds of steels. Hardness is determined by measuring the size of indentation left by a steel ball, diamond cone, or diamond pyramid that is pressed into the surface of the steel with a specified force. Tensile strength is measured by subjecting a standard steel specimen to an increasing load until fracture. Toughness or impact strength is measured by the amount of energy absorbed by the instantaneous fracture of a standardized test piece when struck by a large swinging hammer.

PROPERTIES OF CARBON AND ALLOY STEELS

Carbon steels. Steels can be classified in a number of ways. One of the most common methods is based on the structure produced during solidification; types, already described above, are killed, semikilled, and rimmed steels. Only a few additional comments are needed here. Killed steels are used in applications requiring a dense, uniform structure. Virtually all alloy steels are produced as killed steels, because a high level of deoxidation is necessary to assure control of density and composition. Steels for strand casting must be killed for satisfactory handling during solidification. Thus, a number of kinds of steels that have been cast as semikilled or rimmed ingots must have their compositions and deoxidation practices adjusted for strand casting.

Chemical classification of steels

The primary basis for classifying steels is by chemical analysis. Most major countries have their own methods of classification, which may indicate levels of carbon, manganese, phosphorus, sulfur, or silicon in a steel, depending on type. If other elements, such as aluminum, lead, and copper, are required, their levels also must be specified when a steel is made.

Carbon is both the cheapest and the most effective alloying agent in steels. Low-carbon steels, containing between 0.08 and 0.15 percent carbon, are used largely where low cost, good surface, good ductility and toughness, weldability, and ease of forming and working are desired. Examples include the sheet steels used for automobile bodies, welded tubing, all kinds of steel containers, and wire and case-hardened machine parts. Mild steels, containing from 0.15 to 0.25 percent carbon, are easily machined and stronger but less ductile than the lower carbon grades. These are used in structural members in buildings and bridges, in truck bodies and railroad equipment, and in boilers, drop forgings, and bolts. Steels in the medium-carbon range, from 0.25 to 0.35 percent carbon, are harder and stronger than mild steel and are used for large tubes, ship structures, shafts, and a variety of machine parts. Steels containing 0.35 to 0.65 percent carbon are used for large forgings for heavy machinery. Steel rails and wrought-steel wheels that require good strength and ductility and good response to heat-treatment are made from steels containing 0.65 to 0.85 percent carbon. These steels respond well to heat-treatment. Steels with 0.85 to 1.2 percent carbon are used in the heat-treated condition when ranges of hardness and ductility are required. Uses include springs, cutting tools, and rock drills.

Alloy steels. A very wide variety of compositions of steels is possible with many alloying elements that may be used singly or in combination. It has been found generally that a combination of alloying elements gives better results in cost and properties than is obtained with the use of one element.

Nickel and chromium

Alloying metals. The most important alloying elements in steels are nickel and chromium. Nickel dissolves in ferrite (see below *Steel metallurgy*) and imparts strength to the steel but does not increase toughness. Because of this and because of its cost, nickel is no longer used extensively as the sole alloying element. Chromium at low levels increases the ability of a steel to be hardened by heat-treatment and joins with carbon to form hard, finely dispersed carbides; at levels above 5 percent it imparts corrosion resistance. A number of steels containing from 0.30 to 1.60 percent chromium are used for grinding rolls, roller bearings, and projectiles. Chromium and nickel are often used together or in combination with small quantities of other elements, such as vanadium, molybdenum, and boron. Chromium and nickel are used together because they increase the toughness and strength of a steel and improve its ability to be hardened. The other elements mentioned above aid in these improvements and reduce the amounts of nickel and chromium necessary. The chromium–nickel steels are used extensively for heat-treated machine parts, axles, gears, and armour plate. Rustless irons and stainless steels contain 12 to 18 percent chromium and up to 8 percent nickel. Heat-resisting alloys contain up to 25 percent each of nickel and chromium.

Manganese

Manganese is contained in all steels to ensure good hot-working properties; at levels above 1 percent it is considered an alloying element. Four steel compositions containing 1.75 percent manganese are used in farm implements and automobiles. Higher manganese killed steels are employed in ship construction where brittle fracture is to be avoided. Manganese is an important element in steels used in high-strength castings, rifle barrels, and heat-treated forgings. The tough Hadfield steel, containing 14 percent manganese, is described below in the section on *Steel metallurgy*.

Vanadium is a good deoxidizer. A few tenths of a percent yields steel with a sound structure, a fine grain, or texture, and good response to forging. High-speed steels and tools for hot work contain 0.5 to 2 percent vanadium.

Molybdenum is an effective alloying addition at low concentrations (0.2 to 0.3 percent). Molybdenum increases the depth of hardening, improves the resistance to creep at moderately elevated temperatures, and permits good control of grain size.

Advantages of boron

Boron at a level of 0.0005 percent reduces by half the amounts of nickel, chromium, and molybdenum normally necessary to achieve a desired strength and toughness. Boron was first employed extensively during World War II to economize on scarce alloy elements; its use has continued since that time. Because of its great potency and the very low concentration required, particular skill and care are required when producing, processing, and heat-treating the boron steels.

Tool steels. Ordinary high-carbon tool steels must be

used carefully to avoid overheating the tip of the tool, which destroys the cutting edge. Modern high-speed steel can be used at six or seven times the cutting speeds possible with carbon-tool steels, and the cutting edge can be allowed to reach a red heat. These steels contain tungsten, chromium, and vanadium and may also contain either molybdenum or cobalt. When forged and annealed, they are soft and can be machined easily into complex forms. The usual heat treatment first involves preheating to 850° C (1,560° F), followed by rapid heating to 1,300° C (2,400° F), cooling in oil or air to room temperature, and finally a reheat to 600° C (1,100° F), where the steel is held for 30 minutes before it is cooled to room temperature.

There are many compositions of alloy steels used for special-purpose tools. Steels containing 1 to 5 percent tungsten are used for woodworking tools. Tool steels that can be hardened with very little distortion contain 1 percent each of carbon, tungsten, chromium, and manganese. Die steels used for forging contain 0.3 percent carbon, 12 percent tungsten, 3 percent chromium, and 0.4 percent vanadium. The tool-steel industry is highly competitive; many proprietary compositions exist.

Stainless steels. The three principal types of stainless steels are austenitic, ferritic, and martensitic (see below *Steel metallurgy* for the identification of these structures). The most familiar stainless is an austenitic type that contains 18 percent chromium and 8 percent nickel; it is very ductile and has good resistance to corrosion and staining. It is used in cooking utensils, food-processing equipment, equipment for the chemical industry, hospital equipment, cutlery and flat tableware, and architectural trim. To reduce the cost, manganese may be substituted for some of the nickel. If superior corrosion resistance is required, small amounts of niobium (columbium) may be added. A number of special grades of austenitic steel have been developed for their unique properties.

Martensitic stainless steel, which contains chromium but little or no nickel, is readily formed when annealed; its strength, hardness, and other mechanical properties may be improved by heat treatment. Martensite steels are used for equipment in the chemical and oil industries. Ferritic stainless steels are similar in composition, except for a low carbon content, and have better corrosion resistance. They are employed for automotive trim and interior architectural trim. Stainless steels containing up to 1 percent carbon are used in cutlery.

Ferritic stainless steels

Alloying agents. Various pure metals and ferro-alloys are added to liquid steel to obtain a desired composition in the final product and to control the levels of oxygen and sulfur in the steel. Pure aluminum, cobalt, copper, chromium, lead, manganese, nickel, and titanium are used when extra-pure alloying elements are required. Generally, aluminum, cobalt, copper, and lead are added in the pure form; the other metals are more expensive as pure metals, however, than when obtained as ferro-alloys. Hence, these ferro-alloys, which are alloys of iron and one or more of the alloying elements, are the principal alloying agents used today. Most contain appreciable levels of carbon, although a number of grades are produced with little or no carbon for use in special situations. An example is low-carbon ferrochromium, which is used in producing extra low-carbon stainless steels; *i.e.,* the steel contains not more than 0.03 percent carbon.

Ferromanganese is usually produced in a blast furnace from high-grade manganese ores low in phosphorus, silica, and iron. The usual product contains between 74 and 82 percent manganese, 7.5 percent carbon, not more than 1.25 percent silicon or 0.35 percent phosphorus, and the balance iron. The operation of the blast furnace is somewhat more difficult when producing ferromanganese than when producing hot metal for steelmaking; in consequence, older and smaller furnaces are usually used. The capacity is reduced by about two-thirds, and the coke rate is doubled. Approximately 20 percent of the manganese is lost to the slag. Ferromanganese can also be produced in submerged-arc electric furnaces (described below). Cost of the process is prohibitive, however, except where power is inexpensive, as in Norway.

Ferrosilicon

Ferrosilicon is used extensively in steelmaking both as an alloying agent and as a deoxidizer. Several grades are produced in the submerged-arc electric furnace: 50, 75, 80, and 90 percent silicon, with the balance principally iron. Ferrosilicon contains less than 0.5 percent carbon and less than 0.1 percent each of sulfur and phosphorus. Phosphorus levels must be kept low to prevent the formation of poisonous phosphine (PH_3) when the alloy is exposed to moist air or water. The submerged-arc electric furnace used to produce ferro-alloys is a steel shell having a shallow hearth usually lined with carbon or carbon blocks. Three large carbon electrodes project vertically down into the hearth with their tips well immersed in the furnace charge. Furnaces may or may not have roofs. The electrodes may be preformed or may be made in place by filling thin steel cylindrical shells with a tar–coke mix. This mix is baked as it passes into the hot zone of the furnace. Heat for the process is generated by currents from the electrodes passing through the charge. The charge, consisting of pure quartzite, coke, and iron turnings, is fed in around the electrodes. Silicon is reduced to the metal at a location low in the furnace near the electrode tips, and the charge is heated in part by hot carbon-monoxide gas rising from the reaction zone. The molten iron–silicon alloy collects in the hearth and is periodically drained from the surface.

Ferrochromium

Many grades of ferrochromium are produced; the usual commercial alloys contain approximately 65 to 72 percent chromium and between 4.5 and 8.5 percent carbon. This high-carbon product is usually produced in a so-called three-phase Heroult-type electric-arc furnace having a carbon hearth. The charge is high-grade chromium ore, iron or iron ore, and a flux. The reductant is usually anthracite coal. Action of the arc heats the charge, and the oxides are reduced by the carbon. Molten ferrochromium collects in the hearth. When the reactions are complete, the metal is cast into slabs, and when cold these slabs are broken and the pieces crushed and sized for use. Ferrochromium low in carbon is produced by melting high-carbon chromium, chromium oxide, lime, and fluorspar in an electric furnace. The reaction between chromium oxide and carbon in the ferrochromium eliminates the carbon as carbon monoxide. The lime and fluorspar combine to form a slag. The molten metal contains from 0.2 to 2 percent carbon and is processed as described above. Ferrochromium with an intermediate carbon content can be produced in a submerged-arc furnace, but ferrosilicon and ferrochromium with a high level of silicon instead of carbonaceous materials are used as the reducing agents.

Ferrovanadium is difficult to produce, because the metal is highly reactive. It is manufactured from a charge of ore, carbonaceous material, and fluxes in a small graphite-lined arc furnace. The alloy contains between 35 and 55 percent vanadium.

Ferrotungsten is produced in a small, single-electrode arc furnace in which the hearth acts as one electrode. The charge consists of tungsten ore, carbonaceous material, iron ore, or iron and fluxes. The tungsten ore is easily reduced, but the alloy formed has a high melting point. During the heat, a chunk of solid alloy builds up on the hearth. When sufficient alloy has collected, the furnace is shut down and dismantled. The lump of alloy is crushed, cleaned, and prepared for shipment. The usual grade of ferrotungsten contains 70 to 80 percent tungsten and a maximum of 0.6 percent carbon.

Nickel is often charged to the steel furnace as nickel oxide, which is easily and rapidly reduced to metal by carbon in the charge. Because nickel does not oxidize into the slag, nickel-steel scrap is usually charged to a heat when a nickel-bearing steel is made. Recently, metallic iron pellets containing 1 percent nickel have been used as a source of nickel when making low-alloy steels.

Alloys of many types are marketed for use in steelmaking. These contain barium, boron, calcium, niobium (columbium), molybdenum, magnesium, phosphorus, sulfur, selenium, tantalum, titanium, zirconium, etc. These metals may be in simple combinations with iron or with another metal, such as calcium silicide, or they may be in complex combinations as proprietary mixtures sold for special purposes.

STEEL METALLURGY

Structural forms of iron. Pure iron has three structural forms. The structure that is stable from room temperature to 911° C (1,672° F) is called alpha. The iron atoms in this structure are arrayed in a regular fashion on a lattice made up of simple cubes in which iron atoms are at the cube corners and one is at the centre of each cube. Accordingly, the alpha form of iron is said to have a body-centred cubic-crystalline structure. Between 911° C and 1,392° C (2,538° F), the gamma form is stable. In gamma iron the crystalline lattice is composed of identical cubes in which iron atoms are located at the corners and in the middle of each cube face; *i.e.,* the structure is face-centred cubic. Between 1,392° C and 1,538° C (2,800° F) the stable structure is again the body-centred cubic form, which is designated delta iron. Iron melts at 1,538° C, and in the liquid there is no well-ordered arrangement of the atoms. The physical and mechanical properties of a metal are influenced strongly by the nature of the atomic arrangement, and it is found that metals having the face-centred cubic structure are particularly ductile. The properties of a metal are also influenced significantly by defects in the lattice structure; much research has been directed to understanding the nature of these defects. Among the important defects are (1) a vacancy—*i.e.,* one atom is missing from a lattice point; (2) an iron interstitial, in which an iron atom is wedged between the regularly spaced atoms; and (3) large-scale effects called dislocations, one form of which is an extra plane of atoms that extends only partway through the crystal. The crystal structures of iron and steel have been explored by projecting beams of X-rays, neutrons, or electrons through thin foils or powdered specimens and observing the way that these beams are scattered and bent (diffracted) because of the regularity and irregularity of the atomic structure. Many changes in the properties of iron and of metals in general, such as hardening by cold work, brittle fracture, and plastic flow under stress, have been explained in terms of defects.

Structural defects

Phases in steel. Three phases may be found in unhardened steels: ferrite, austenite, and cementite. Ferrite is the body-centred cubic form of iron (alpha) in which some carbon has dissolved; the maximum amount at equilibrium is 0.025 percent at 727° C (1,341° F). Austenite is the face-centred form of iron (gamma) in which some carbon can dissolve; at 727° C the maximum amount is 0.76 percent, rising to 2.1 percent at 1,147° C (2,097° F). Cementite is a compound of iron and carbon having the formula Fe_3C. Cementite can decompose into ferrite and graphite below 727° C and into austenite and graphite above that temperature. The reactions are so slow, however, that they are not important in the processing of most steels. The iron–cementite phase diagram (Figure 19) shows the regions of stability of the phases found in slowly cooled steels. Below 727° C these steels, containing between 0.025 and 2 percent carbon, consist of ferrite and cementite. As has been noted already, with increasing carbon content, the amount of ferrite decreases and the amount of cementite increases.

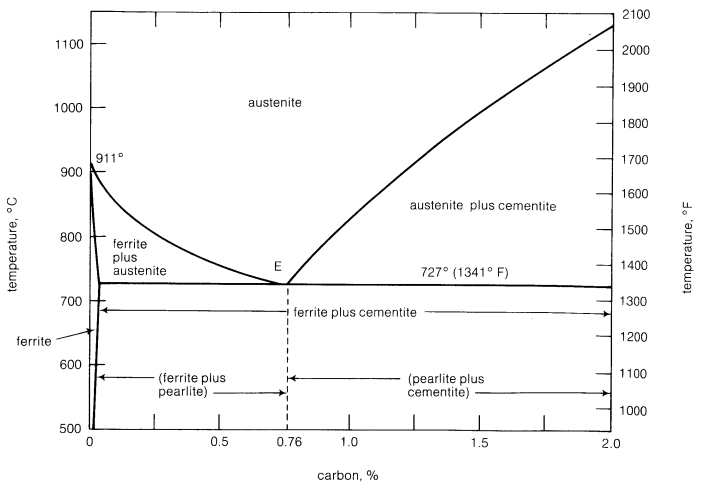


Figure 19: Steel portion of iron–cementite diagram (see text).

A 0.76-percent-carbon steel. The significance of other areas of the phase diagram can be illustrated with steels of three different compositions, beginning with a steel containing 0.76 percent carbon. Below 727° C this steel contains ferrite and cementite. When the steel is heated to just above 727° C (point E), the ferrite dissolves the cementite, and the whole structure is changed to austenite. The temperature at which austenite begins to form on heating is called the Ac_1 temperature (*c* for *chauffage,* or "heating"). Initially the austenite crystals are very small, but either holding the steel at this temperature for a long time or heating it to a higher temperature and holding for a shorter time causes some of the crystals to grow at the expense of others and the crystal size to increase. If this steel is then cooled to a little below 727° C, the austenite will decompose, forming alternate platelets of ferrite and cementite, in a structure called pearlite because of its appearance (see Figure 20A). This reaction of one solid phase decomposing to form two different solid phases is called a eutectoid change, and the temperature at which the austenite decomposition is complete is designated Ar_1 (*r* for *refroidissement,* or "cooling"). Because the reactions are somewhat sluggish, for a given composition, Ac_1 is slightly higher than Ar_1.

The distribution of the ferrite and cementite in pearlite is dependent on the rate of cooling and will be described later.

A 0.3-percent-carbon steel. The second steel, one containing 0.3 percent carbon, will also consist of ferrite and cementite below 727° C. When heated just above that temperature, austenite containing 0.76 percent carbon forms by a reaction between the cementite and some of the ferrite. The resulting structure contains only ferrite and austenite. With increased temperature the ferrite dissolves in the austenite until at approximately 810° C (1,490° F) only austenite containing 0.3 percent carbon remains. The temperature at which the transformation to austenite is complete is given the designation Ac_3 (there is no Ac_2).

All carbon steels containing between 0.025 and 0.76 percent carbon undergo this same change on heating; the temperature (Cr_3) at which the steel becomes all austenite rises, however, as the carbon content decreases. This temperature range (see Figure 19) within which the structure changes on heating from ferrite plus cementite to ferrite plus austenite and then to all austenite is known as the critical range for carbon steels. The crystals of austenite again grow in size as the temperature of this steel is increased above 810° C. A steel containing less than 0.76 percent carbon that is slowly cooled from a temperature above the critical range will begin to form ferrite at the Ar_3 temperature. Very small crystallites of ferrite will form first at the boundaries of the austenite crystals. This ferrite is low in carbon so that the remaining austenite gradually becomes enriched in carbon, reaching 0.76 percent at 727° C.

Critical range for carbon steels

A 1.1-percent-carbon steel. The third steel contains 1.1 percent carbon. If cooled slowly, it, too, will contain only ferrite and cementite below 727° C. If this structure is heated to just above 727° C, the ferrite will react with some of the cementite to form austenite containing 0.76 percent carbon. The structure then will contain only austenite and cementite. With further heating the cementite will decrease, and at 860° C (1,580° F) the steel contains only austenite. When the steel is cooled from a temperature at which it is all austenite, cementite begins to form just below 860° C at the austenite grain boundaries. On further cooling, the amount of cementite increases, and the austenite is depleted in carbon until it reaches 0.76 percent at 727° C. The austenite then decomposes into pearlite, and the final structure is pearlite and cementite or ferrite and cementite, pearlite consisting, as noted earlier, of ferrite and cementite. In steels containing more than 0.76 percent carbon, the temperature at which the solution of cementite in austenite is complete is designated Ac_{cm}, and the temperature at which cementite begins to form from austenite is designated Ar_{cm}.

Effects of carbon content on structure and properties of steel. *Slowly cooled steels.* At temperatures below 727°

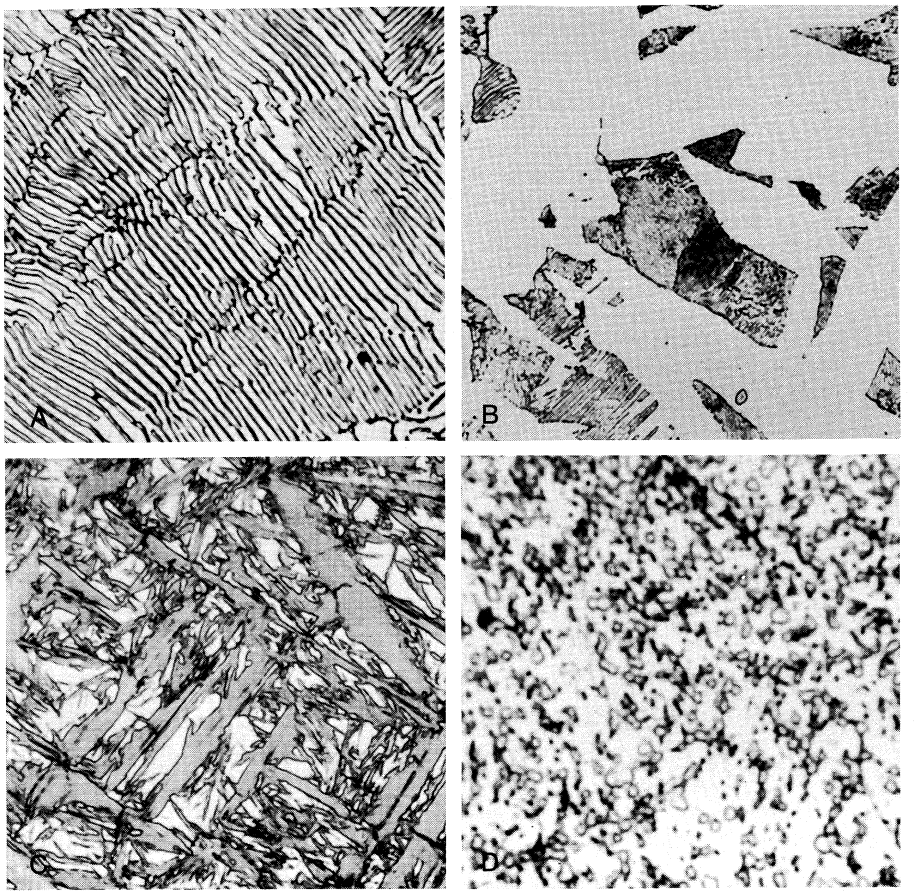


Figure 20: *Microphotographs of various steels.*
(A) Pearlite × 1,000; (B) ferrite (light) and pearlite (dark and lamellar) × 600; (C) martensite × 850; and (D) tempered martensite × 2,000.
D. Lloyd Thomas

C, slowly cooled steels containing more than 0.025 percent carbon contain ferrite and cementite. Ferrite is soft and ductile, and cementite hard and brittle. The amount of cementite seen under the microscope increases with the amount of carbon in the steel. A steel with 0.3 percent carbon contains approximately 94 percent ferrite and 6 percent cementite, whereas at 1 percent carbon these amounts are 85 percent ferrite and 15 percent cementite. In a slowly cooled steel the austenite remaining at 727° C will transform into hard and strong pearlite that consists of alternate layers of ferrite and cementite (Figure 20A). With increased carbon content up to 0.76 percent, slowly cooled steels contain an increased amount of pearlite and less ferrite (Figure 20B); strength and hardness increase, but ductility and toughness are reduced (see Figure 21A). Steels having maximum ductility and toughness are low-carbon steels that have been slowly cooled or annealed below 727° C. The final structure of a steel will be coarse if the austenite crystals were coarse at the beginning of cooling. Hardness and tensile strength are not affected materially by the temperature at which the steel is held before cooling, nor do they depend upon the size of the austenite crystals. The finer the final structure, however, the tougher will be the steel when subjected to shock at and below room temperature. Because recrystallization takes place both on heating (Ac_1 to Ac_3) and cooling (Ar_1 to Ar_3), a coarse structure can be converted to a fine structure by reheating the steel to just above its critical range of temperature and then slowly cooling it.

A steel containing more than 0.76 percent carbon when cooled from a temperature where it is all austenite will first precipitate cementite at the austenite grain boundaries. With further cooling the structure will consist of grains of austenite surrounded by cementite. Subsequently, the austenite will form pearlite, and the final structure will consist of islands of pearlite in a continuous network of cementite. Though steel formed this way is strong, it is also very brittle, so the cementite is dispersed by kneading the steel. It is important that these steels not be reheated above 760° C (1,400° F) in subsequent heat treatment; otherwise, the embrittling network will form on cooling.

Hardened steels. The temperature and the nature of the decomposition of austenite into pearlite in plain carbon steels depends strongly upon the rate of cooling. With in-

Carbon steels

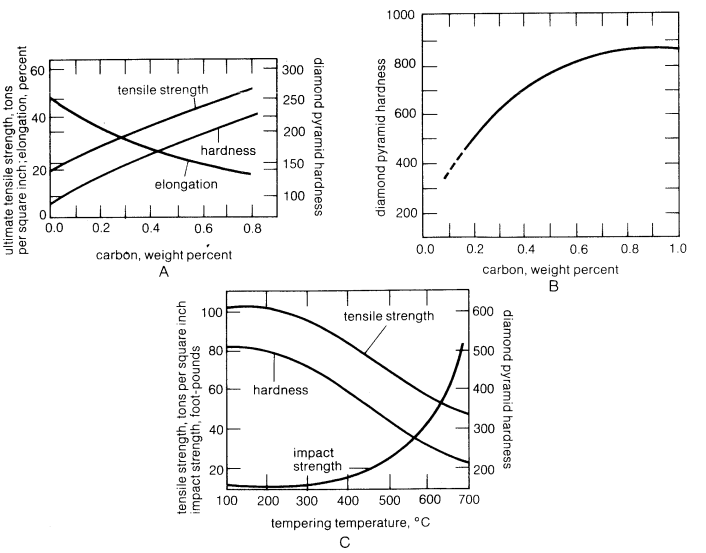


Figure 21: *Effects of composition and heat treatment on mechanical properties of steels.*
(A) Mechanical properties of annealed steel containing various percentages of carbon. (B) Hardness of martensite as a function of carbon content. (C) Effects of tempering temperature on mechanical properties of tempered martensite.

creased cooling rates the Ac_1 temperature can be lowered progressively to about 550° C (1,022° F), and the pearlite will become so fine that it cannot be resolved under the light microscope. Accompanying this change is increased hardness and strength of the pearlite. With a further increase in the cooling rate, the transformation of austenite to pearlite may be suppressed altogether. A broad range of heat treatments are used to take advantage of these effects. A steel is annealed by cooling it slowly in a furnace from just above its critical temperature range. The result is coarse pearlite, and the steel is soft. A steel cooled in air from just above its critical temperature range is said to be normalized. The result is a finer pearlite, and the steel is harder, stronger, and less ductile than it would be if annealed. By quenching the steel very rapidly in cold water or brine, the formation of pearlite may be partially or completely suppressed, and austenite can be retained down to about 200° C (400° F). At that temperature, the austenite begins to form a structure called martensite (Figure 20C and D). The transformation from austenite to martensite is complete at about room temperature. The minimum cooling rate necessary to suppress the formation of any pearlite is called the critical cooling velocity for the steel. Quenching to form martensite is often used because of the special properties obtained when the steel is annealed.

Forming martensite

Martensite results when face-centred cubic austenite transforms to the body-centred structure without cementite being precipitated. Carbon is retained in the structure, which is stretched slightly so that it is no longer cubic; thus, martensite can be thought to be ferrite supersaturated with carbon. This supersaturation results in severe distortion of the crystalline lattice, which increases with increasing carbon content of the steel. The hardness of martensite increases to a maximum as the carbon content of the steel increases. To obtain a completely martensitic structure in a steel containing less than 0.8 percent carbon, the steel must be quenched from a temperature above its critical temperature range. This is necessary to suppress the formation of both ferrite and cementite. Steels containing more than 0.8 percent carbon are quenched from 760° C, usually immediately after forging, thus producing a structure of cementite dispersed in a matrix of martensite. The cooling rate necessary to obtain martensite in plain carbon steels very often causes distortion of the part being heat-treated. In severe cases, cracking may also result, especially in a part with a complex shape. Thus, even with the most severe quench in brine or cold water, a maximum thickness of a plain carbon steel that can harden fully to martensite is only about 20 millimetres (0.8 inch), because the cooling rate decreases from the surface to the interior of a part.

Low-alloy steels

Nickel, chromium, manganese, and molybdenum in steels reduce the critical cooling rate necessary to obtain martensite. These elements are said to increase the steel's ability to harden. On the other hand, they do not increase the hardness of the martensite that forms. The use of these alloying elements singly or in combination reduces the tendency for distortion and cracking because a less drastic quench is required to obtain martensite. A quench in oil or air is usually sufficient. Alloy steels containing nickel, chromium, and molybdenum in sections 150 millimetres (about six inches) thick can be made martensitic throughout by an oil quench. Because alloy steels are much more expensive than carbon steels, they are used only when their special properties are required.

Hardened and tempered steels. Martensite is usually much too brittle and too hard for most uses, and residual internal stresses are present in the steel as a result of the rapid cooling necessary to form the martensite. Reheating the steel to a selected temperature in the range of 200° to 600° C (400° to 1,100° F) and then cooling it in air will relieve the internal stresses and decrease the hardness and strength of the steel. A marked increase in toughness is obtained, however. This operation is called tempering. The temperature and time for the tempering operation must be carefully selected on the basis of the type of steel and the final properties desired. A tempering temperature near 680° C (1,260° F) causes the greatest loss in hardness and strength but imparts the greatest toughness as measured by impact strength (see Figures 20D, 21C). Tool steels are tempered in the range of 200° to 250° C (400° to 500° F) to relieve internal stresses without the loss of hardness. Spring steels are tempered at 350° to 500° C (650° to 900° F) to obtain good strength without a serious loss of impact properties. Machine parts subject to impact loadings may be tempered in the range of 500° to 600° C (900° to 1,100° F).

In tempering, martensite decomposes into ferrite and cementite. The amount of decomposition increases as the tempering temperature increases up to 400° C (750° F). In the decomposition, carbon forms small particles of cementite, and the highly stressed martensite relaxes to ferrite. The product is a structure consisting of very fine particles of cementite imbedded in a matrix of ferrite. The structure obtained in tempering at a low temperature is so fine that it can be resolved with the electron microscope but not with the optical microscope. The structure is coarser with higher tempering temperatures, and those obtained at 600° C can be resolved in the optical microscope at 1,000 magnifications (see Figure 20D).

Mechanical properties of tempered alloy steels

The mechanical properties obtained after tempering alloy steels containing nickel and chromium are much superior to those of the carbon steels. This is because these alloying elements strengthen the ferrite matrix. At the same strength, alloy steels are much tougher than carbon steels. The highly alloyed steels are softened much less by tempering than are carbon and low-alloy steels. High-speed tool steels, for example, are quenched from 1,300° C (2,400° F) and tempered at 500° to 600° C with little if any loss in hardness. This resistance to softening at high temperatures makes these steels useful as cutting tools where very high tool temperatures are encountered at high cutting speeds.

The phenomenon known as temper brittleness occurs when steels containing relatively high levels of phosphorus, manganese, or chromium are tempered in the range of 400° to 550° C (750° to 1,000° F). It also occurs if these steels are tempered at a higher temperature and then are cooled slowly through this range. The effect can be avoided by tempering at a temperature above 550° C, followed by a quench from the tempering temperature. Molybdenum in the steel reduces the tendency toward temper brittleness.

Surface hardening. In many instances steel parts such as gear teeth, the teeth on power shovels, and the tires of railroad wheels are subject to severe wear or abrasion at the surface as well as shock and dynamic loading. The needs for high hardness at the surface and a tough interior are conflicting requirements, but they can be met by treating the surface to obtain a hard "case" while the interior of the steel part remains tough. Common methods for accomplishing this are described briefly below.

Carburizing. Carburizing is the modern form of an old cementation process. The part, made from a carbon or alloy steel low in carbon, is heated to about 900° C (1,650° F) in a bed of carbonaceous material or a carburizing atmosphere. Carbon diffuses into the surface, the amount of carbon added and the depth of penetration depending upon the carburizing medium, the temperature, and the time of exposure. The resulting case will contain about 0.8 percent carbon and be fully martensitic when the part is quenched to room temperature. Because of its lower carbon content, the core will not be fully hard and will be tough. A variety of heat treatments after carburization are used, depending upon the combination of properties desired for the case and core.

Nitriding. A steel containing small amounts of chromium, molybdenum, and aluminum will form a very hard, thin case on the surface when exposed to an atmosphere of ammonia. This case consists of nitrides of aluminum and chromium formed when nitrogen obtained from the decomposition of the ammonia diffuses into the steel. The depth of the case increases with the duration of the nitriding process. The procedure is first to quench and temper the steel part at 550° to 650° C (1,000° to 1,200° F) to obtain a tough case. The hard case is then obtained by exposing the part to an atmosphere of ammonia for

Nitriding procedure

about 80 hours at 500° C. Nitriding has the advantage over carburizing in that the case remains hard when exposed to temperatures up to 500° C, whereas carburized cases begin to soften at about 200° C.

Flame and induction hardening. In this method a part is made from a medium-carbon steel, with or without alloying elements. The part is quenched and tempered to produce the desired case properties. Then the portions of the surface that are to be hardened are quickly heated to above the Ac_1 temperature of the steel and quenched in oil or water. The heated region becomes martensitic, and the core retains its toughness. Heating is accomplished by numerous small gas flames or by high-frequency induction.

Other methods. As indicated earlier, Hadfield manganese steels, which contain approximately 1.4 percent carbon and 14 percent manganese, are extremely tough and wear-resistant. These properties are obtained by heating the steel to between 1,000° and 1,060° C (1,800° to 1,940° F), followed by a water quench. This treatment preserves the austenitic structure down to room temperature. Abrasion of the surface and cold work causes the surface to transform to martensite. Thus, the hard surface is continually renewed as the surface wears away.

A number of other methods are also used to obtain a hard surface. One way is to apply a hard facing to a steel part by welding, brazing, or spraying a coating of hard material. Materials used for this are tungsten carbide and Stellite, tungsten, chromium, and carbon.

WORLD STEEL PRODUCTION

During the early 18th century, Sweden was the chief producer of the relatively small volume of steel made in the world. With the development of mass-produced iron and steel, Great Britain became dominant. It has been estimated that the output of pig iron in Great Britain rose from 28,000 tons per year in 1720 to 75,000 tons in 1788 and to almost 3,000,000 tons by 1852 (unless otherwise noted, figures are in short tons). It is notable that in the early part of the 19th century much of this iron was exported: 20 percent in 1806 and almost 50 percent in 1856. Pig-iron production continued its rapid increase to 7,100,000 tons in 1875, when the United Kingdom produced almost half the world's output. In that year, production in the United States was 2,200,000 tons, or 15 percent of world production. In 1880 British production reached 8,700,000 tons and then rose gradually to about 10,000,000 tons during the remainder of the century. During this time the British share of the world's production of pig iron declined from 42 percent in 1880 to approximately 22 percent as production of pig iron in the United States rose rapidly to over 15,000,000 tons in 1900, making that country's share of world production about 35 percent. Great Britain's annual output of pig iron was first surpassed by that of the United States in 1890, by Germany and France in 1921, when the coal supply was interrupted by a strike, by the Soviet Union in 1931, and by Japan in 1954.

British pig-iron production in the 19th century

Great Britain also was preeminent in the production of steel in the 19th century. The older methods for making iron and steel products with a great expenditure of labour rapidly gave way to the Bessemer and open-hearth processes after 1870. In 1875 Great Britain produced 790,000 tons of steel, or 40 percent of world production. In 1896 the output rose to about 4,400,000 tons, or approximately 22 percent of the world's total. U.S. production of steel and wrought iron had passed that of the British by 1880. British production continued to rise, however, to about 7,000,000 tons in 1900, 18,000,000 tons in 1950, 27,000,000 tons in 1960, reaching a high of 31,000,000 tons in 1970 but declining by the late 20th century to less than 20,000,000 tons.

Table 4 shows the production by processes in Great Britain from 1896 to 1906. During the interval, the increase in production was chiefly due to the increased output of open-hearth steel. In 1905 the output of Bessemer steels reached its peak for the 20th century, and, subsequently, the basic open hearth became the primary process for steel production in the United Kingdom.

From 1896 to 1970 the United States was the largest producer of steel in the world. In 1971, however, the Soviet Union surpassed the United States. The rise and decline of processes in use in the United States for the past 100 years illustrates in a general way the changes that have occurred throughout the world. The first production of acid Bessemer steel was in 1867, when 30,000 tons were made. In 1870, production of acid open-hearth steel was begun, and the output was 1,500 tons. In the same year, 42,000 tons of acid Bessemer steel were made, and the output by this process rose rapidly, reaching its all-time peak of 13,800,000 tons in 1906. In the interval from 1870 to 1906 the acid Bessemer was the primary steel-producing process in the country. Subsequently, it declined in output, becoming of less relative importance each year until production by this method ceased in 1967. Initially, the use of the open-hearth process grew more slowly, because it was operated with acid hearth refractories and an acid slag and, as a result, was not competitive with the acid converter. Its growth quickened after 1884, when it was found that the process would operate satisfactorily with basic refractories. The basic furnace rapidly replaced the acid furnace for most purposes in the United States, because it could treat readily the moderate levels of phosphorus in pig iron made from the ores from the Great Lakes ranges. These ores were used extensively because they were plentiful, cheap, and of good quality. The basic open-hearth process surpassed the acid Bessemer in output in 1908 and was the primary means for producing steel in the United States until 1970, when it was surpassed in output by the basic oxygen process. The reasons for the long ascendancy of the basic open hearth has been its flexibility in producing virtually any type of carbon steel and a number of the low-alloy steels, its flexibility in utilizing scrap, hot metal, and cold iron in virtually any proportion, and its ability to make products of good quality. Between 1928 and 1961 the open-hearth process produced over 85 percent of the steel made in the United States each year.

Table 4: Production of Steel and Wrought Iron in Great Britain
(in 000 metric tons)

year	open hearth			Bessemer			total steel	wrought iron
	acid	basic	total	acid	basic	total		
1896	2,180	174	2,354	1,380	465	1,845	4,209	1,234
1900	2,912	298	3,210	1,275	499	1,774	4,984	1,183
1905	3,092	808	3,900	1,418	588	2,006	5,906	954
1906	3,435	1,195	4,630	1,328	610	1,938	6,568	1,026

Source: J.C. Carr and W. Taplin, *History of the British Steel Industry* (1962).

By courtesy of the American Iron and Steel Institute

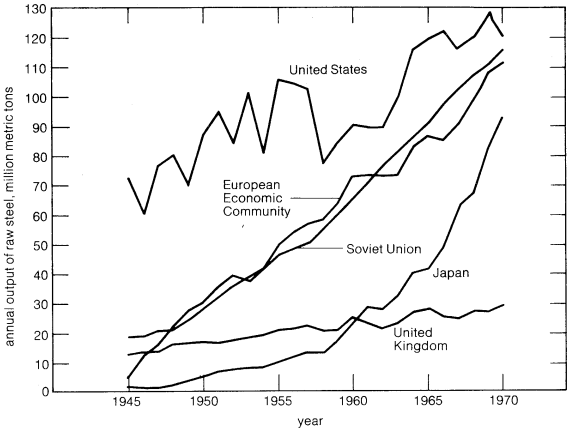


Figure 22: Growth of raw-steel production of major producing countries since World War II.

The use of the basic oxygen processes by the U.S. steel industry did not begin until 1955, six years after the basic oxygen furnace (or L–D furnace) had been developed in Austria in 1949. The subsequent growth in output of oxygen steel by the basic oxygen furnace has been dramatic. Production rose from 725,000 tons in 1955 to more than 70,000,000 tons in the late 20th century, accounting for

Growth of oxygen steel

more than one-half of the total U.S. steel production. The primary advantages of the basic oxygen furnace over the open hearth are its rapid operation, lower cost, and ease of control. It is best suited to the production of low-carbon steels, which are used in large quantities today. To meet the disadvantage of its inability to operate normally with more than 35 percent scrap in the charge, a number of steel plants have also installed basic electric furnaces for the production of carbon steels. With the two processes in effect, the inability of the oxygen furnace to melt scrap can be avoided.

High-quality and specialty steels were produced in the United States by the crucible process between 1870 and 1915. Growth was slow, and the peak in output of 146,000 tons was reached in 1907. Production held steady at approximately that level until 1920, but thereafter it declined rapidly to a few thousand tons a year. The commercial use of the electric-arc furnace for producing alloy and specialty steels began in 1906. Output by this process exceeded that by the crucible process in 1916 and increased rapidly to an average of approximately 650,000 tons per year, which was sustained from 1918 to 1938. The demands of World War II pushed production to about 4,500,000 tons in 1943, and the output declined only slightly in the immediate postwar years. The rapid growth in production between 1950 and 1970 resulted from the use of the process for melting carbon steels with scrap as the principal charge. The electric furnace came into wider use during the 1970s by virtue of increased production of metallized ore and increased recovery of iron and steel scrap from waste. By the late 20th century, electric-furnace steel production had surpassed open-hearth production in the United States, accounting for more than one-quarter of the domestic output.

In the year 1900 the basic Bessemer process produced 65 percent and the basic open hearth produced 31 percent of the total of 7,500,000 tons of steel made in Germany. In 1905 for France the corresponding figures are 60 percent and 34 percent, respectively, and the total production of steel that year was 2,500,000 tons. The balance of the output was by acid methods. With the increased importation of ores low in phosphorus from abroad, European steel plants began to use the basic oxygen furnace and the Kaldo process more extensively. The basic oxygen furnace with a two-slag operation also has been used to refine the high-phosphorus hot metal obtained from the local ores. The use of oxygen and steam and of oxygen and hydrocarbons in the basic Bessemer has received renewed interest. In the Soviet Union, the predominant steelmaking process continues to be the basic open hearth. In Japan, where steelmaking facilities were heavily damaged or destroyed in World War II, virtually the whole of the production of carbon steels is by the basic oxygen furnace. Electric furnaces are used to produce alloy and stainless steels, but a limited tonnage of stainless steel is also produced in the basic oxygen process. Throughout the world the acid processes have given way almost completely to the basic processes, the basic open hearth, the basic Bessemer, the basic oxygen processes, and the basic electric-arc furnace.

World production of iron and steel has maintained an upward trend except for brief periods following World Wars I and II and in 1929 to 1932 during the Great Depression; production also slackened during the late 1970s and the early 1980s as a result of diminished demand. The production of pig iron has been closely related to the production of steel, because it is used primarily as a raw material in steelmaking. In recent years the large industrial producers of cast-iron products have obtained the iron by melting steel scrap and then carburizing and alloying it. Discarded cast iron from the foundry is sent back to the melting furnace, but there is relatively little pig iron in the charge.

The place of steel in international trade has changed since World War II. Just after the war, the United States was a major exporter of steel. With the rapid growth of steelmaking capacities in other highly industrialized countries, this has been reversed. The flow of steel products between countries depends strongly on the economic climate and varies greatly from year to year. In the past, major steel-producing countries have always exported steel products to the less industrialized countries. To aid in reducing the drain on their currencies because of the importation of steel products, however, many small countries construct small steel plants using electric furnaces to melt scrap and metallized local ores. (J.F.E./Ed.)

Aluminum production

Aluminum, or aluminium, is the most modern of the common metals used industrially in large quantities. Although crude salts of aluminum were used as early as the 5th century BC, the metal was not isolated until 1825 and was not produced commercially on a large scale until 1888. Today, aluminum has surpassed all other metals except iron and steel. The industry's growth has been rapid in Europe and spectacular in the U.S., where its average annual rate of growth since 1893 has frequently approached 15 percent. Other developed countries also manifest a large demand for aluminum—and that of the underdeveloped countries is growing rapidly.

HISTORY

Before 5000 BC, people in Iraq were making fine pottery from a clay that consisted largely of a compound containing aluminum, and almost 4,000 years ago, Egyptians and Babylonians used aluminum compounds in various chemicals and medicines. Pliny refers to *alumen,* known now as alum, a compound of aluminum widely used in the ancient and medieval world to fix dyes in textiles. By the 18th century, the earthy base "alumina" was recognized as the potential source of a metal.

The English chemist Sir Humphry Davy in 1807 attempted to extract the metal. Though unsuccessful, he satisfied himself that alumina had a metallic base which he named "alumium" and later changed to "aluminum." The name has been retained in the U.S. but modified to "aluminium" in many other countries.

First laboratory production of aluminum

A Danish physicist and chemist, Hans Christian Ørsted in 1825 finally produced aluminum. "It forms," Ørsted reported, "a lump of metal which in color and luster somewhat resembles tin."

A few years later (1845) Friedrich Wöhler, a German physicist working at the University of Göttingen, made metallic aluminum in particles as large as pinheads and became the first man to determine the following of aluminum's properties: specific gravity, ductility, colour, stability in air, etc.

Deville process. Nine more years passed before a French scientist, Henri Sainte-Claire Deville, announced a momentous improvement in Wöhler's method, which permitted Wöhler's "pinheads" to coalesce into lumps the size of marbles. Deville's process became the foundation of the aluminum industry. Bars of aluminum, made at Javel Chemical Works and exhibited in 1855 at the Paris Exposition, introduced the new metal to the public.

Although enough was by then known about the properties of aluminum to indicate a promising future, the cost of the chemical process for producing the metal was too high to permit widespread use. But important improvements presently brought breakthroughs on two fronts: first, the Deville process was improved; and second, the development of the dynamo made available a large power source for electrolysis, which proved highly successful in separating the metal from the compound sodium aluminum chloride.

The work of Hall and Héroult. The modern electrolytic method of producing aluminum was discovered almost simultaneously, and completely independently, by Charles M. Hall of the United States and Paul-Louis-Toussaint Héroult of France in 1886. (By an odd coincidence, both men were born in 1863 and both died in 1914). The essentials of the Hall-Héroult processes were identical and remain the basis for today's aluminum industry. Purified alumina is dissolved in molten cryolite and electrolyzed (broken down) with direct current. Under the influence of the current, the oxygen of the alumina is deposited on the anode and is released as carbon dioxide, while free molten aluminum—which is heavier than the electrolyte—is deposited on the carbon lining at the bottom of the cell.

Hall immediately recognized the value of his discovery. He applied July 9, 1886, for a U.S. patent and worked energetically at developing the process. Héroult, on the other hand, although he applied for French patents, apparently failed to grasp the significance of the process. He continued work on a second successful process that used copper as a cathode and produced an aluminum–bronze alloy. Conveniently, in 1888, a German chemist, Karl Joseph Bayer, discovered an improved method for making pure alumina from low-silica bauxite ores.

Hall and a group of businessmen established the Pittsburgh Reduction Company in 1888 in Pittsburgh. The first ingot produced by electrolytic smelting was poured in the fall of that year. Demand for aluminum grew and a larger reduction plant was built at New Kensington, Pennsylvania, using steam-generated electricity to produce one ton of aluminum per day by 1894. The need for cheap, plentiful hydroelectric power led the young company to Niagara Falls, where, in 1895, it became the first customer for the new Niagara Falls power development.

In a short time, the demand for aluminum exceeded Hall's most optimistic expectations. In 1907 the company changed its name to Aluminum Company of America (Alcoa). Until World War II it remained the sole American producer of primary aluminum.

European industry. Neuhausen, Switzerland, is the "nursery" of the European aluminum industry. There, to take advantage of waterpower available from the falls of the Rhine, Héroult built his first aluminum–bronze production facility, which later became the Aluminium-Industrie-Aktien-Gesellschaft. The British Aluminium Company Limited, organized in 1894, soon recognized the wealth of cheap electric power available in Norway and became instrumental in building aluminum works at Stongfjorden in 1907 and later at Vigeland. In France the Société Électrométallurgique Française, also based on Héroult's patent, was started near Grenoble around 1888. Little aluminum was produced in Germany before 1914, but the war brought an urgent demand, and several smelters went into production employing electricity generated by steam power. Later, the Soviet Union began producing substantial amounts of aluminum in the Ural industrial complex. By 1940, Japan, Sweden, and Yugoslavia had become producers.

ORES AND MINING

Aluminum is the third most abundant element on the Earth's surface. Only oxygen and silicon are more common. The Earth's crust to a depth of 16 kilometres (10 miles) contains an estimated 8.05 percent of aluminum. Aluminum has a strong tendency to combine with other common elements and so never occurs in nature in the metallic form. Its compounds, however, are an important constituent of virtually all common rocks. It is found in clay, shale, slate, schist, granite, syenite, and anorthosite.

Occurrence in nature

Aluminum is found as hydrated aluminum oxide—a combination of alumina and water—in combination with such impurities as iron oxide, titania, and silica.

Bauxite. The most important aluminum ore, an iron-containing rock consisting of about 52 percent aluminum oxide, was discovered in 1821 near Les Baux in southern France. The material was later named bauxite. Though now applied loosely to any aluminum-bearing substance, bauxite is best defined as an aluminum ore of varying degrees of purity in which aluminum in the form of hydrated oxide is the largest single constituent.

Bauxite varies greatly in physical appearance, depending on its composition and impurities. It ranges in colour from yellowish white to gray, or pink, to dark red or brown, if high in iron oxides. It may be earthy, or range in form from clay to rock. Bauxite has been found in all the world's continents except Antarctica. The richest deposits generally lie in areas that, during formation, were in tropical and subtropical climates, providing optimum conditions of heavy rainfall, constant warm temperatures, and good drainage.

Principal deposits of bauxite

Large deposits are found in the United States, the Caribbean islands, northern South America, Hawaii, Australia, India, Indonesia, Malaysia, China, the Soviet Union, Ghana, Greece, Yugoslavia, Hungary, Italy, and France.

Not all bauxites are economical for aluminum production. Only earth with an aluminum oxide content of 45 percent or more is considered practical for making metal. Two minerals, gibbsite and boehmite, which contain 40–60 percent alumina, are presently considered economical enough to be processed. Gibbsite is found largely in tropical areas on either side of the Equator, while boehmite is found largely north of the subtropical belt in southern Europe, the Soviet Union, Turkey, and Mainland China.

Known deposits of bauxite can supply the world with aluminum for hundreds of years at present production levels. When high-grade bauxite deposits are depleted, substantial reserves of secondary ores will remain to be exploited: laterite deposits in the northwestern United States, Hawaii, and Australia; anorthosite in the western United States; apatite and diaspore in Europe; kaolinite in the southeastern United States. Other nonbauxite sources of alumina are also available: alumina clays, dawsonite, alunite, aluminous shales, igneous rocks, and saprolite and sillimanite minerals. Eastern Europe, led by the Soviet Union and Poland, already is refining alumina from nonbauxitic ores, namely nepheline syenite and high-alumina clays. Vast new bauxite developments in Australia, Guinea, and Indonesia have tended to postpone interest in secondary ore technology elsewhere.

Mining. By far the greatest quantity of commercially exploited bauxite lies at or near the Earth's surface. Consequently, it is mined in open pits requiring only a minimum removal of layers of soil and rock covering the ore. Bauxite beds are blasted loose, dug up with power shovel or dragline, and the ore transported by truck or rail to a processing plant, where it is crushed for easier handling, washed of clay and dirt, kiln-dried, pulverized, and shipped to a refining plant. Refining plants are located near mine sites, if possible, since transportation is the major item in bauxite costs.

Approximately 90 percent of all bauxite produced is refined into alumina, which is ultimately smelted into aluminum. The remaining 10 percent is used in other applications, such as abrasives and refractories, and in the manufacture of petroleum catalysts and paper. Bauxite yields one ton of aluminum from four to six tons of ore.

ORE PROCESSING

The production of aluminum from bauxite is a two-step process: refining bauxite to obtain alumina and smelting alumina to produce aluminum. Bauxite contains a number of impurities, including iron oxide, silica, titania, and water. If these impurities are not removed during refining, they will alloy with and contaminate the metal during the smelting process. The ore of aluminum, therefore, must be treated to eliminate these impurities. Purified alumina suitable for commercial production usually contains less than 0.1 percent of other oxides. Bayer's process, with some modification, is still the most widely used and commercially successful method for the production of alumina, and virtually all aluminum is smelted from alumina using the Hall-Héroult electrolytic process.

Refining the ore. There are a number of alkaline, acid, and thermal methods of refining bauxite, clay, or other ores to obtain alumina. Acid and electrothermal processes generally are either too expensive or do not produce alumina of sufficient purity for commercial use. A process that involves treatment of ore with lime and soda is used in the Soviet Union. Smelting is accomplished with the aid of limestone in Norway; this process may also have been used in the Soviet Union during World War II.

The Bayer process

The Bayer process (Figure 23) involves four steps: digestion, clarification, precipitation, and calcination.

In the first step, bauxite is ground, mixed with a solution of caustic soda (sodium hydroxide), and pumped into large pressure tanks called digesters, where the ore is dissolved under steam heat and pressure. The sodium hydroxide reacts with the hydrated aluminum oxide of bauxite to form a saturated solution of sodium aluminate; insoluble impurities, called red mud because of their iron oxide content, settle to the bottom. During clarification, the mixture is passed through a series of pressure-reducing

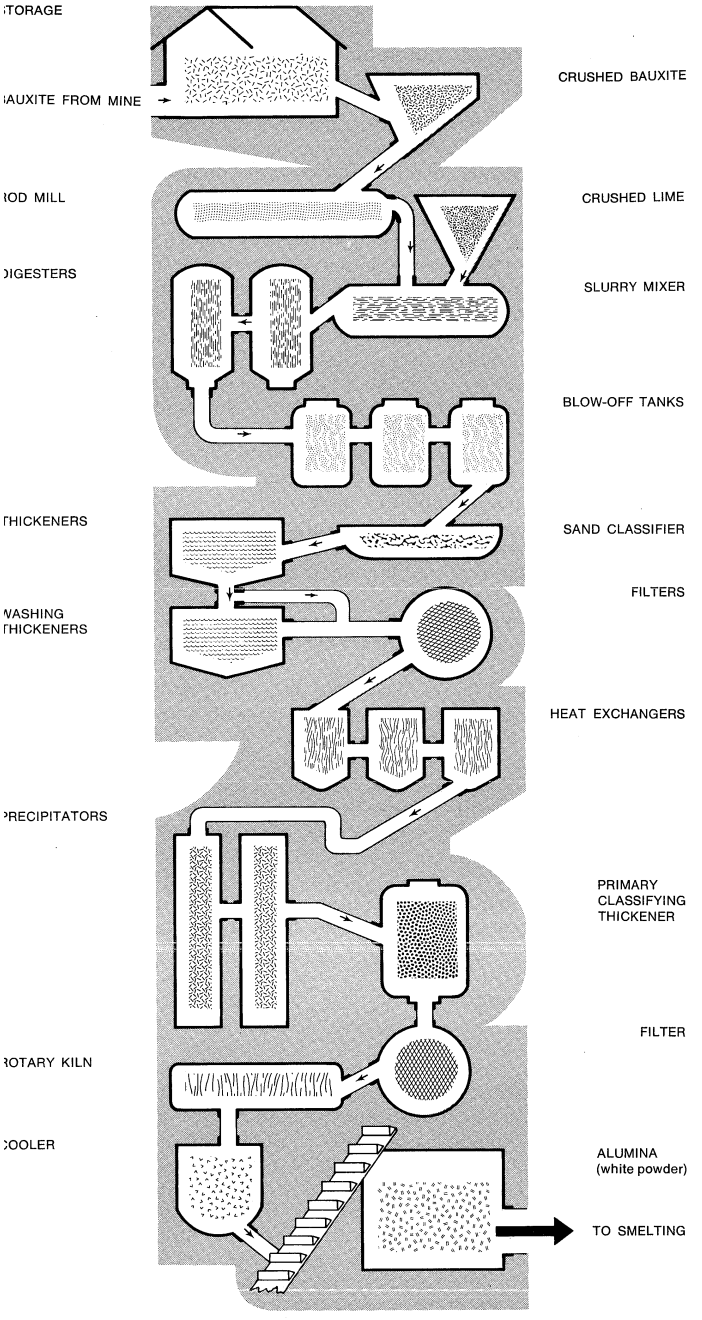


Figure 23: Principal operations in the Bayer aluminum refining process.

tanks (called blow-off tanks) and filter presses (shown in Figure 23 as the sand classifiers). Cloth filters catch the red mud, which, except at plants where lower grade ores are refined, is discarded. A resultant green liquor passes through filters into cooling towers (heat exchangers) and then into tall, silo-like precipitators. Sizable amounts of hydrated alumina crystals are added to the solution in the precipitators as "seeding" to hasten crystal separation.

The seed crystals attract other crystals and form groups physically heavy enough to settle out of solution. This precipitate of aluminum hydrate crystals is filtered out, washed to remove impurities, and heated in long, rotary kilns at temperatures in excess of 980° C (1,800° F). Free water and water that is chemically combined with the aluminum hydrate are driven off, leaving commercially pure alumina—or aluminum oxide—a superdry, fine white powder similar to sugar in appearance and consistency. It is still half aluminum and half oxygen by weight, bonded so firmly that neither chemicals nor heat alone can separate them.

Alcoa combination process

During World War II, the Alcoa combination process was developed for processing lower grade ores containing relatively high percentages of silica. Very briefly, this process reclaims the alumina that has combined with silica during the digestion process and has been filtered out with the red mud. The red mud is not discarded but is heated with limestone (calcium carbonate) and soda ash (sodium carbonate) to produce a sodium aluminate solution. This solution is returned to the digesters and proceeds through the remaining steps of the Bayer process. The waste product is a silicate material called brown mud.

Alumina produced by the Bayer process is quite pure, containing only a few hundredths of 1 percent of iron and silicon and traces of residual soda. In addition to being the primary raw material for producing metallic aluminum, alumina is an important chemical in its own right. It is used widely in the chemical, refractories, ceramic, and petroleum industries (see below *Aluminum chemicals*).

Refining four tons of bauxite yields about two tons of alumina. A typical alumina plant, using the Bayer process, can produce almost 3,000 tons of alumina per day. The cost of alumina can vary widely, depending on plant size and efficiency, labour costs, overhead, and cost of bauxite.

Smelting. Although there are several methods of producing aluminum, only three are worth attention. The Deville process, which involves direct action of metallic sodium upon aluminum chloride, was for years the basis of European aluminum production. It has been commercially abandoned in favour of the more economical electrolytic process, which involves dissolving purified alumina in melted cryolite and passing a current through the mixture to obtain metallic aluminum and carbon dioxide. A carbothermic approach, the oldest and the classical method for reducing (removing oxygen from) metallic oxides, has been for years the subject of intense research by the metal industries. This involves heating the oxide together with carbon. Carbon and oxygen combine to produce carbon monoxide or dioxide; the pure metal is left behind. The great attraction of carbothermic smelting is the possibility of bypassing the alumina refining cycle and of starting with lower-grade ores than bauxite. Despite many years of intensive research, however, no economic competitor has been found for the Bayer-Hall-Héroult approach.

The Hall-Héroult process

Although unchanged in principle, the Hall-Héroult smelting process of today differs greatly in detail from the original process. Modern technology has produced substantial improvements in equipment and materials, and has lowered final costs.

In a modern smelter (Figure 24), alumina is dissolved in smelting cells—deep, rectangular steel shells lined with carbon—that are filled with a molten electrolyte or con-

By courtesy of Aluminum Company of America

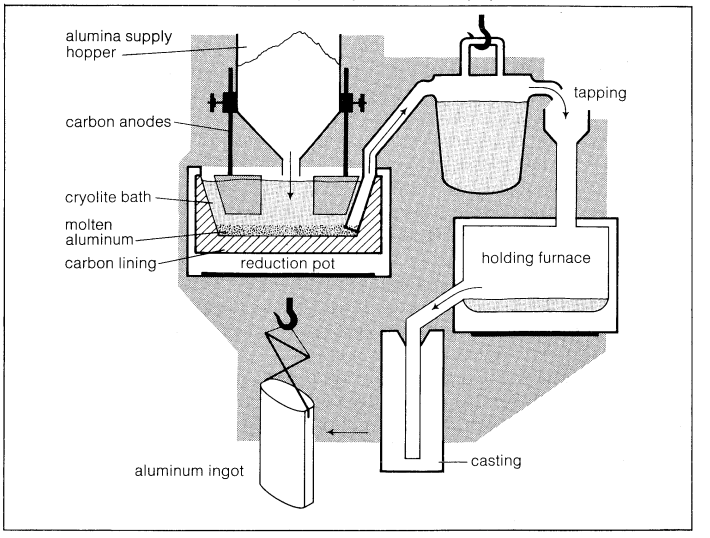


Figure 24: Simplified cross section of a modern aluminum smelter, showing the basic steps in the smelting process (see text).

ductor, consisting mostly of an aluminum compound called cryolite.

By means of carbon electrodes, direct current is passed through the electrolyte to the bottom of the cell, causing a crust to form over the surface of the molten bath. Alumina is added on top of this crust, where it is preheated by the heat from the cell (about 950° C [1,710° F]) and its adsorbed moisture driven off. Periodically the crust is broken, and the alumina is stirred into the bath to maintain proper concentration.

The actual process by which alumina is dissolved in the molten bath and electrolytically decomposed is a matter of dispute. The end results, however, are deposit of molten metallic aluminum on the bottom of the cell and the combination of the oxygen with the carbon, which then escapes as carbon dioxide.

The smelting process is continuous. Additional alumina is added to the bath to replace that consumed by reduction. Heat generated by the electric current maintains the bath in a molten condition so the fresh alumina is dissolved. Periodically, molten aluminum is siphoned off.

Because some aluminum fluoride from the cryolite electrolyte is lost in the course of the reaction, aluminum fluoride is added, as needed, to restore the chemical composition of the bath. An acid bath condition with an excess of aluminum fluoride provides maximum efficiency.

About 200 grams (0.5 pound) of carbon is consumed for every 450 grams (one pound) of aluminum produced.

Smelting cells

In actual practice, long rows of smelting cells, called potlines, are electrically connected in series. Normal voltages for cells range from four to six volts, and current loads range from 50,000 to 150,000 amperes. From 50 to 150 cells may form a single potline with a total line voltage of up to about 750 volts, thus making clear why power is one of the most costly ingredients of aluminum. Since 1900, aluminum producers have searched for sources of cheap hydroelectric power but have also had to construct many facilities that use energy from fossil fuels. Over the last four decades, technological advances have reduced the amount of electrical energy necessary to produce about 0.5 kilogram (one pound) of aluminum. In 1930, the figure was 12 kilowatt-hours. By the early 1980s the amount of electrical energy consumed for each 0.5 kilogram of aluminum produced had declined to about 4.5 kilowatt-hours.

Finished molten aluminum is siphoned from smelting cells into large crucibles. From there, the metal may be poured directly into molds to produce foundry ingot, it may be transferred to holding furnaces for further refining or for alloying with other metals, or both, to form fabricating ingot. As it comes from the cell, primary aluminum averages 99.80 percent purity.

Automation and computer control have had a marked effect on smelter operations. The most modern reduction facilities use fully mechanized carbon plants and sophisticated potline computer control for monitoring and automating potline operations.

ALUMINUM METAL AND ITS ALLOYS

Properties of the metal. A ductile, silvery white metal, usually with a dull lustre due to a surface film of aluminum oxide, aluminum is very light, weighing approximately one-third as much as an equal volume of copper or steel. It is corrosion resistant, an excellent conductor of heat and electricity, reflects both light and radiant heat, is nonmagnetic, does not readily absorb neutrons, can be safely used with foods and medicines, and can be formed by all known metalworking processes.

Aluminum can be joined by welding, brazing, soldering, adhesive bonding, rivetting, stitching, or stapling, by means of a number of mechanical assemblies, such as nuts and bolts, screws, and nails. It can be given a wide variety of mechanical finishes by grinding, polishing, buffing, abrasive blasting, and burnishing. A variety of chemical finishes can be used, such as alkaline or acid etches, bright dips (these give an extremely shiny finish to metal), chemical milling, and immersion plating. It is suited to an electrochemical process called anodizing. Or it can be electroplated with other metals, or given organic coatings such as paint, lacquer, and plastic films. Aluminum also is frequently finished by porcelain enamelling or metallizing.

Strength of aluminum

High-purity aluminum (99.90 percent) is relatively soft and has a fairly low tensile strength of about 500 kilograms per square centimetre (7,000 pounds per square inch) in the annealed condition. (Annealing involves heating and then cooling slowly to make the metal less brittle.) By alloying and proper thermal and mechanical treatment, however, it can be made much harder and stronger, with tensile strengths as high as 7,000 kilograms per square centimetre (100,000 pounds per square inch). Unlike some other metals, the strength and ductility of aluminum increases at supercold temperatures. Upon melting, the solid metal expands about 7 percent in volume, the solidification shrinkage being 6.6 percent of the liquid volume. Hydrogen is the only gas known to be appreciably soluble in molten aluminum; its solubility increases with temperature but becomes nearly zero when the metal freezes.

Aluminum may act as a base to form salts with acids or as a weak acid to form salts with strong alkalies. It is stable in air because of a thin, transparent oxide film that forms on exposure to air, protecting the aluminum from further oxidation and reaction. Growth of this natural oxide film is self-limiting; that is, when a thin layer is formed, further growth is halted. Molten aluminum is protected in air by a thicker oxide coating, which also deters further oxidation. Finely divided atomized or flake aluminum mixed with air and ignited will explode violently. Aluminum reacts rapidly with boiling water to liberate hydrogen and form aluminum hydroxide.

Alloys. In its superpure condition (99.99 percent), aluminum lacks strength and hardness but is formable, weldable, corrosion-resistant, and an excellent conductor of electricity. Super-purity aluminum has many applications; in chemical equipment, reflectors, as a catalyst in making gasoline, in fine jewelry, and in electronic components. Most aluminum used today, however, is alloyed with other metals to improve its properties for special uses.

Although almost all the metallic elements readily alloy with aluminum, only a few are important major ingredients in commercial, aluminum-base alloys. These are copper, manganese, silicon, magnesium, and zinc. Smaller amounts of nickel, tin, iron, cadmium, lead, bismuth, cobalt, chromium, titanium, vanadium, boron, sodium, and zirconium may be added for particular purposes.

Aluminum–copper alloys, following heat treatment, possess mechanical properties superior to those of mild steel. These alloys have good ductility and machinability, but they do not approach aluminum–magnesium for corrosion resistance and weldability. They are well-suited for screw-machine products and structural applications requiring high strength-to-weight ratios, such as truck-trailer panels and aircraft components. Addition of about 1 percent manganese yields non-heat-treatable aluminum alloys that are stronger than commercially pure aluminum, have good corrosion resistance, and are highly workable.

Aluminum–manganese alloys are popular for cooking utensils, heat exchangers, chemical equipment, storage tanks, highway signs, roofing, siding, and many architectural applications.

Adding major amounts (about 10 percent) of silicon to commercially pure aluminum yields an alloy with a relatively low melting point. Thus, aluminum–silicon alloys are suitable for welding wire and brazing material. Because silicon imparts great fluidity to molten metal, this alloy is used also in castings. The addition of up to 5 percent magnesium yields an alloy with good tensile strength, weldability, hardness, and corrosion resistance in marine atmospheres. Current uses for aluminum–magnesium alloys include packaging (can stock), ships and boats, light standards, appliances, and cryogenic (low temperature) tanks. Adding both silicon and magnesium to aluminum produces alloys that are easily formed, machined, welded, and finished, have good resistance to corrosion, and are of medium strength. End uses include transportation equipment, architectural installations, bridge and highway structures, and machine parts.

Aluminum–zinc alloys are hard, heat-treatable, and possess very great strength. Additions of lesser amounts of

magnesium, copper, and chromium—along with zinc—yield excellent alloys used widely in aircraft structures and mobile equipment.

Other metallic elements may be added to these basic aluminum alloys for special purposes. For example, bismuth, lead, and tin yield an alloy that is easier to machine; boron helps increase electrical conductivity; nickel gives improved strength at higher temperatures; and addition of titanium refines the grain of the metal, resulting in greater strength and ductility.

Casting alloys. Just as different casting methods, such as sand, die, or permanent mold, are best suited to specific applications, so are certain alloys best suited to particular casting methods. The elements found most often in aluminum casting alloys are silicon, magnesium, copper, and nickel. Silicon (up to 20 percent) improves the fluidity of molten aluminum, allowing it to flow more easily through a mold cavity. It also reduces shrinkage, decreases leaks, and makes welding easier. Copper may be added as a hardening constituent and to increase the strength of both heat-treated and non-heat-treated parts. Magnesium improves strength and corrosion resistance, and makes the alloy easier to machine. Nickel improves dimensional stability and strength at high temperatures.

In the early days of the industry, the casting alloys most used in the U.S. were of the aluminum–copper type. Britain and Europe preferred aluminum–zinc alloys. After 1920, aluminum–silicon was introduced. A 5 percent silicon alloy became popular in the U.S. for architectural and other applications in which good casting properties and high resistance to corrosion were more important than strength. After World War I, alloys containing about 10 percent copper were used throughout the world for pistons and other parts of internal-combustion engines. Most of the aluminum alloy castings made in the U.S. up to 1940 were of 7–10 percent copper, although heat-treated alloys containing 4–5 percent copper and 10 percent magnesium were employed extensively where higher strength and ductility were required.

Heat treatment required for casting

The heat treatment of castings requires a relatively long heating (6–24 hours) at temperatures of 370–535° C (700–1,000° F), the exact temperature and time depending on the alloy composition and type of casting. This is followed by quenching the casting in hot or cold water or oil, with subsequent aging at either room temperature or a temperature somewhat above the boiling point of water. The purpose of this "solution heat treatment" is to make certain that as much as possible of the alloying elements, particularly copper, magnesium, and zinc, are dissolved in the aluminum in what is termed a solid solution.

The reason for the long duration of heat treatment is inherent in the structure of the castings. When a casting is made from an aluminum-base alloy containing copper, for example, the crystals that first form during the cooling process are nearly pure aluminum. Most of the copper remains in solution in the liquid aluminum, and as the aluminum crystals grow and use up the liquid aluminum, the concentration of copper in the liquid increases. Since increasing the proportion of copper in aluminum lowers its freezing point, this process continues as the casting cools, until that temperature is reached at which the lowest melting point solution (eutectic) of aluminum and copper freezes. In the meantime, however, the crystals have grown until they almost touch one another, so that when the eutectic freezes it fills the small intergranular spaces and acts as a cement between the crystals. Aluminum can hold approximately 4.0 percent of copper in solid solution at the heat-treating temperature (about 510° C [950° F]). Only a small fraction of this, however, is actually in solid solution. The rest is in the intergranular eutectic. In order to obtain a nearly saturated solid solution in the aluminum, it is necessary to hold the casting at high temperatures long enough for the copper to diffuse from the grain boundaries and distribute itself throughout the grains or crystals. The larger the grains, the longer this takes. Consequently, fine-grained castings (such as those cast in permanent molds) require a shorter heat-treating time than coarse-grained sand castings. Wrought material (in which the eutectic has been broken up and mechanically dispersed) requires a much shorter heat-treat time than permanent mold or sand castings.

Supersaturated solutions

If the casting is suddenly cooled or quenched to room temperature, a supersaturated solid solution results, since the solid solubility of copper in aluminum at room temperature is about 0.25 percent. Unless some accelerating metal, such as magnesium, is present, this solid supersaturated solution is unstable and the copper precipitates slowly during a long period of storage at room temperature. If the temperature is raised to the boiling point of water or higher, however, the mobility of the atoms of copper in the alloy is sufficiently increased so that they are able to get together in a few hours. They apparently precipitate in the form of submicroscopic particles of a copper–aluminum compound. With longer heating and at higher temperatures, these particles grow large enough to become visible under the microscope upon a polished and properly etched surface of the alloy. They are found to be arranged in quite definite patterns along certain planes in the individual crystals. Apparently these minute particles of hard constituent remain between the crystals, just as particles of fine sand, placed between sheets of paper, cause them to resist the tendency to slip. The precipitation process thus increases the tensile strength and yield strength of the alloy but decreases its ductility. In the presence of 0.5 percent magnesium, this precipitation occurs spontaneously in a few days at room temperature. Such precipitation at room temperature is spoken of as natural aging, while precipitation induced by heating at elevated temperatures is called artificial aging.

In certain aluminum–magnesium alloys, the solid solubility of magnesium at the heat-treating temperature and at lower temperatures is considerably greater than that of copper, but the processes and results of heat-treating, quenching, and aging are similar. The solubility of silicon varies little with temperature, so that heat treatment of an aluminum–silicon alloy has little effect on its properties. If magnesium is present, however, a magnesium–silicon compound is formed, which has an appreciable solid solubility at heat-treating temperature and precipitates by artificial aging after quenching. A heat-treatable casting alloy of this type may contain about 5 percent silicon, 0.5 percent magnesium, and 1.5 percent copper. Its tensile strength is lower than that of the heat-treated copper or magnesium alloys, but it has much better casting characteristics.

The mechanical properties of castings vary greatly, depending not only on composition but also on casting conditions. Thin sections generally show a somewhat higher tensile strength than do thick sections. Heat-treated alloys as a rule show higher strength and ductility than those not heat-treated. Typical strengths for ordinary sand-cast alloys are in the range of 1,300–2,100 kilograms per square centimetre (19,000–30,000 pounds per square inch), while heat-treated alloys may show 2,100–3,200 kilograms per square centimetre (30,000–45,000 pounds per square inch). As a general rule, an alloy cast in permanent (metal) molds will show tensile strengths 350–700 kilograms per square centimetre (5,000–10,000 pounds per square inch) higher than when sand-cast.

Typical strengths of sand-cast alloys

Wrought alloys. Wrought alloys are available in worked form, such as sheet, foil, plate, extrusions, forgings, rod, bar, and wire. Working operations and heat treatments change ingot structure into a recrystallized or severely fragmented structure, characteristic for each wrought alloy, which determines strength, corrosion resistance, and other properties. Wrought alloys fall into two basic classes: non-heat-treatable alloys and heat-treatable alloys.

The first wrought-aluminum alloy was probably a relatively pure aluminum, since the pure metal's softness and high ductility permitted it to be rolled, forged, or drawn with a minimum of effort. Alloying elements usually cause the pure metal to become stronger, harder, and less ductile and, consequently, more difficult to fabricate. A high grade of pure aluminum was used early in this century for electrical conductors. Known as EC (electrical conductor) metal, this wrought material is usually 99.45 percent aluminum. Commercial "pure aluminum" sheet contains 99.5–99.79 percent aluminum. Foil is generally

made of high-purity metal, ranging up to 99.99 percent. For the flexible packaging market, products such as collapsible tubes are usually made of 99.8 percent aluminum. Wrought products of pure aluminum are used in many fields, with major consumption in the packaging, chemical, electrical, and architectural fields.

In the manufacture of wrought aluminum articles, maximum ductility is produced by annealing the article, usually by heating for some time at about 340° C (650° F). The hardening and strengthening effect of even small amounts of alloying metals is evident in comparing typical properties of commercial-purity annealed aluminum sheet with sheet of 99.99 percent purity. The typical tensile strength of about 500 kilograms per square centimetre (7,000 pounds per square inch) for the superpurity metal increases to 900 kilograms per square centimetre (13,000 pounds per square inch) for commercial-purity sheet, while the yield strength increases from 105 to 350 kilograms per square centimetre (1,500 to 5,000 pounds per square inch). (The yield strength of aluminum is defined as the load in pounds per square inch that produces a permanent stretch or elongation of 0.2 percent.) At the same time, the elongation in two inches is reduced from 50 percent to 35 percent in commercial-purity sheet.

Cold-working (rolling, forging, drawing, and so on, at ordinary temperatures) hardens the metal, increasing both tensile and yield strength. This effect increases with the amount of cold work. The magnitude of change is seen from the fact that a typical tensile strength for hard-rolled commercial-purity aluminum sheet is 1,700 kilograms per square centimetre (24,000 pounds per square inch) with a corresponding yield strength of 1,500 kilograms per square centimetre (22,000 pounds per square inch) and an elongation of only 5 percent in five centimetres (two inches). Intermediate amounts of cold working produce intermediate tempers.

Aluminum–magnesium alloys, containing 5 to 6 percent magnesium, form a useful series of non-heat-treatable, solid solution, wrought alloys. Strengths after the metal has been annealed range from 1,100 kilograms per square centimetre (16,000 pounds per square inch) tensile strength and 400 kilograms per square centimetre (6,000 pounds per square inch) yield strength to 3,600 kilograms per square centimetre (52,000 pounds per square inch) tensile strength and 1,700 kilograms per square centimetre (24,000 pounds per square inch) yield strength. Elongations may exceed 25 percent. Aluminum–magnesium wrought alloys have good strength, forming, and welding characteristics and high resistance to corrosion. Finishing and decorative capabilities are also characteristic of the aluminum–magnesium alloy-system.

Aluminum–magnesium–silicon forms a family of alloys used widely in wrought products for architectural and transportation applications. Aluminum–zinc–magnesium alloys, when heat-treated and aged, have very great strength and are employed principally in aircraft components and mobile equipment.

Aluminum–copper wrought alloys combine wide strength range with good weldability, superior resistance to stress corrosion, and good elevated-temperature properties. Additions of lead and bismuth provide superior machinability. Aluminum–copper–magnesium alloys, introduced over 50 years ago, are outstanding structural alloys in the aircraft industry.

Miscellaneous alloys in heat-treatable and non-heat-treatable classes span a broad range of composition and applications. Increasingly popular and useful are composite wrought products, consisting of an aluminum-alloy core coated on one or both surfaces with a thin layer of corrosion-resistant pure aluminum or aluminum alloy. The combination of core and cladding alloys is selected so that the cladding provides electrochemical protection for the core. When a corrosive solution (electrolyte) is in contact with the product, current from the cladding flows through the electrolyte to the core and the cladding is dissolved, thus protecting the core. The most widely used clad aluminum products are sheet and plate.

ALUMINUM POWDER AND PARTICLES

The phrase "powder and particles" encompasses a family of forms in which 99.5 percent unalloyed aluminum ingot is formed into powders or granules. Ingot is melted in a furnace and passed through streams of air or inert gases. The molten aluminum forms a spray that solidifies into granules and is collected. These atomized granules range in size from 0.015 micrometre (one micrometre equals 1/1,000 of a millimetre) up to 17,000 micrometres. Powders normally are defined as being finer than 1,000 micrometres, and particles range up to 17,000 micrometres. In shape, powders and particles may be granular, globular, irregular, spherical, elongated, ribbon, flake, or other, with varying surface area, unit weight, and density.

Applications include paints and pigments, missile fuels, chemicals, explosives, and a variety of putties, cements, and solders. Aluminum pigments are used widely in metallic paints on storage tanks, bridges, and towers, where durability and corrosion resistance are required. Particles in flake form in aluminum paint arrange themselves in parallel layers five to 10 flakes deep, forming, in effect, a thin layer of metal. Atomized powders are major components of rocket propellants, flares, and fireworks. Mixed with uranium compounds, they are used in nuclear fuel elements. Aluminum powder is utilized to produce aluminum-clad steel wire for the electrical industry. In aluminothermic reactions, aluminum powders are combined with the oxides of other metals with the production of heat. Aluminum's great affinity for oxygen makes it useful in the aluminothermic smelting of steel, chromium, and vanadium.

Use in metallic paints

Aluminum powders are used in powder metallurgy processes in which the powders serve the same function as molten metal in casting operations. Powders are put in a die and subjected to pressure. The compacted piece is heated to produce aluminum parts having specific desirable properties.

ALUMINUM CHEMICALS

Aluminum oxide. Aluminum oxide exists in several different crystallographic forms, of which corundum is most common. Corundum is characterized by a high specific gravity (4.0), a high melting point (about 2,015° C or 3,660° F), great insolubility and hardness.

Aluminum oxide is the major ingredient in the commercial chemicals known as aluminas. Of the pure, inorganic chemicals, aluminas are among the largest volume produced in the world today. Rubies and sapphires are crystalline, nearly pure varieties of alumina, coloured by small amounts of impurities. Synthetic rubies and sapphires are made commercially by fusing a mixture of high-purity aluminum oxide with colouring agents in an oxyhydrogen blowpipe flame. Most are cut and drilled to form tiny "jewel" bearings in watches and various precision measuring instruments.

Activated alumina is a porous form of aluminum oxide from which much of the chemically combined water has been driven off. It is chemically inert to most gases, nontoxic, and will not soften, swell, or disintegrate in water. It has the ability to adsorb and hold moisture without change in form or properties, and has high resistance to shock and abrasion. Activated alumina is used in oil, chemical, and petrochemical industries as a dehydration agent and purifier in the manufacture of gasoline, petrochemicals, natural gas, and hydrogen peroxide.

Calcined alumina is aluminum oxide that has been heated at temperatures in excess of 1,050° C (2,000° F) to drive off nearly all chemically combined water. In this form, alumina has great chemical purity, extreme hardness (nine on the Mohs hardness scale, on which diamond is 10), high density, and a high melting point (above 2,000° C [3,700° F]). It possesses good thermal conductivity, heat and shock resistance, and great electrical resistivity at elevated temperatures. This combination of properties makes calcined alumina useful in abrasives, glass, porcelains, spark plugs, and electrical insulators, but the greatest quantity of calcined alumina is smelted to obtain aluminum.

Tabular alumina

Tabular alumina is aluminum oxide that has been heated

to temperatures above 1,650° F (3,000° F). Composed of tablet-like crystals, it has very high capacity and thermal conductivity, and exceptional strength and volume stability at high temperatures. For these reasons, a major usage for tabular alumina is in the production of high-quality refractories, the materials used for lining industrial furnaces. High-alumina refractories are used in the metal and glass industries in boiler installations, large furnaces and kilns for smelting metals and firing glass, pottery and porcelain, and in the manufacture of building bricks.

Most refractories are produced in the form of brick, bonded and fired in furnaces. Some castable refractories are made in the form of mortars, usually tabular alumina with calcium aluminate cement as a binder. These mortars, called grog, are sprayed under pressure to form the linings of the steel industry's basic oxygen furnaces, electric ore furnaces, steel ladles, and coke ovens, and for steam boilers, rotary kilns, and many other high-temperature applications.

In addition, fused aluminas are used in special refractories for the glass industry. Fused alumina is calcined alumina that is melted in electric arc furnaces, cooled, crushed, and recast into desired shapes. In another application, industrial processes requiring hot gases use a unique heat-transfer device called a pebble heater. Gases to be heated are passed through a bed of tabular alumina balls that have been heated to extreme temperatures. In still another application, an aluminous insulating material is formed by melting alumina and silica in an electric furnace and subjecting the molten mixture to high-velocity gases to produce fine white fibres.

Aluminum hydrates and other compounds. The hydrous forms of alumina, called aluminum hydrates, may contain either one (monohydrate) or three (trihydrate) molecules of water. Each may exist in two different crystalline forms, known as "alpha" and "beta." In both hydrates, the alpha variety is more common. The alpha trihydrate (gibbsite) and alpha monohydrate (boehmite) occur in bauxite.

Hydrated alumina is used extensively in the production of aluminum chemicals, such as alum, sodium aluminate, aluminum fluoride, and hydrous aluminum chloride. It is a constituent of petroleum catalysts, and an ingredient in the manufacture of plastic and rubber goods, paper, glass and vitreous enamel, adhesives, varnishes, and toothpastes.

Aluminum sulfate is used in water purification. Aluminum chloride in various forms is used as a catalyst in organic chemistry, and in the cosmetic industry as a deodorant. Aluminum fluoride is a popular compound used widely in the purification of aluminum.

ALUMINUM FABRICATION

Aluminum has been called the obliging metal because it can be fabricated by all the techniques known to metalworking. Similarly, it can be joined to itself or to other materials by the total range of joining procedures and finished easily and economically by many methods (see also above *Metallurgy*).

Fabricating methods. *Casting.* In casting, molten aluminum is poured—or forced under pressure—into a mold, the shape of which it then takes. There are many variations. In sand-casting, expendable molds of special sands and clays are formed around a pattern into a shape. Molten aluminum at about 700° C (1,300° F) is then poured into the sand cast. The cast is discarded after use. The process is employed for large shapes and small quantity runs. Plaster mold casting, a refinement of sand-casting, provides parts with smooth surface finishes and fine dimensional control. Permanent mold castings employ metal molds and cores into which molten aluminum at about 730° C (1,350° F) is poured. Such a mold may be used thousands of times to produce parts with smooth surfaces and superior strength and soundness. In die-casting, molten metal at about 650° C (1,200° F) is injected into water-cooled, heat-treated steel dies at high velocities under pressures up to 1,400 kilograms per square centimetre (20,000 pounds per square inch). This is the preferred method for long production runs and for products requiring good finishes, dimensional accuracy, and internal soundness. Shell mold casting uses a thin shell of sand, which has been preformed and baked in a metal pattern. Investment casting is a two-step process in which wax is first cast in permanent molds, a new plaster mold prepared around the wax, the wax melted out, and molten aluminum poured into the cavity. Centrifugal casting uses centrifugal force to drive molten metal into a complex die.

Die-casting

Newly developed is premium-engineered casting that utilizes combinations of different casting processes to develop final products with excellent mechanical properties, close control of dimensions, thin sections, and smooth surfaces.

Rolling. Rolling is the process by which aluminum ingot is reduced by passing it between pairs of rolls. Rolling ingot, which may be up to 9 metres (30 feet) long, 90 centimetres (36 inches) wide, and 50 centimetres (20 inches) thick, generally passes through a scalper that trims and smooths the rolling surfaces and edges of the ingot. The ingot then is heated to about 425° C (800° F) and passed through a hot reversing mill that in successive passes rolls the ingot down to plate size which varies from 150 millimetres (six inches) to six millimetres (0.25 inch). For further reduction to sheet, six millimetres to 0.152 millimetre (0.006 inch), plate is passed through a series of cold-rolling mills. Before the final cold-rolling, it is usually annealed to attain optimum metallurgical qualities. Heat-treatable sheet is passed through a heat-treating furnace after it has been cold-rolled. Non-heat-treatable sheet goes directly from cold-rolling to finishing operations. Plate, which is not generally cold-rolled, goes from hot-rolling to heat-treating or finishing operations. Further rolling of sheet down to less than 0.152 millimetre produces foil. Aluminum foil can be rolled so thin that it takes 10 layers to equal the thickness of a newspaper page. Rolling also may be used to produce rod and bar stock. Specially formed blocks or bars of the metal, called blooms, are passed through successively smaller and variously shaped rolls to achieve the desired shapes and sizes.

Aluminum foil

Extruding. In extruding, a hydraulic ram forces an ingot or billet, which generally has been heated to about 425° C, through an orifice in a die. The metal takes the shape of the orifice, in cross section, and is sawed to length. Theoretically, there is no limit to the variety of shapes that can be extruded. Size is limited only by the power of the press and the size of the die block. Hydraulic presses range from about 500 tons pressure for soft alloys to 14,000 tons for hard alloys. Extruded shapes generally fall into three categories: solid, a shape other than rod and bar that does not enclose a void, as in the letter I; semihollow, a shape with part or all of its cross section partially enclosing a void, as in the letters Y and E; and hollow, a shape with part or all of its cross section completely enclosing a void, as in the letters P and O.

Forging. In aluminum forging, preheated ingots or billets are formed into desired shapes by repeated blows of a hammer, or by the application of pressure in a press. Forging hammers range from 227 to 22,700 kilograms (500 to 50,000 pounds) capacity. Drop forging employs a hammer blow to force the metal into a die cavity. Dies are usually produced from forged alloy steel blocks. Press forging uses a fast push, rather than a blow to force metal into the die. Die forging, hand forging, and ring forging are variations of the basic process.

Drawing. Aluminum can be drawn; that is, pressed into a hole or cavity whose shape it retains. This basic process, called deep drawing, is used to produce aluminum cups, pots, and similar containers. In cold-drawing, a pointed length of rod is pulled through a die orifice that is smaller than the overall diameter of the rod. Aluminum wire is formed by this process, being drawn through successively smaller die orifices until the desired gauge is reached. Generally, rod becomes wire when it is less than 9.52 millimetres (0.375 inch) in diameter. Cold-drawn tubular products are made in much the same manner, except that a bar of steel (mandrel) is employed to form the core of the tube and is withdrawn after the tube is complete.

Other processes. Aluminum and its alloys are distinguished by the ease with which they can be machined. Castings, forgings, and other products are brought to final shape, dimension, and finish on lathes, joiners, routers, drill presses, shears, grinders, and many other high-speed

metal-working machines. Other fabricating techniques used successfully with aluminum are spinning, stamping, embossing, coining, bending, rotary swaging, and cold-heading.

Heat treatment. Besides processes that change the metal's shape or mold the metal into desired forms, there are certain processes that insure the metal's final characteristics and metallurgical qualities. They include heat-treating and stress relieving.

There are four general types of heat treatments applied to aluminum and its alloys: preheating, annealing, solution heat treatment, and precipitation heat treatment.

Cast ingots are usually preheated before cold- or hot-working. Preheating improves workability, helps control recrystallization and grain growth, and causes desirable diffusion of alloying elements. Most aluminum alloys are heat-treated in gas-fired or radiant-heated-air or controlled-atmosphere furnaces. Depending on size and alloy, typical preheats for aluminum forging or extrusion ingots range from four to 48 hours at temperatures of 425 to 650° C (800 to 1,200° F).

Annealing results in a metal of softened, workable condition, relieves internal stresses, and stabilizes properties and dimensions. Certain non-heat-treatable, wrought-aluminum alloys are heated rapidly to annealing temperature to insure fine grain size. Typical annealing conditions for heat-treatable wrought alloys are two to three hours at 410° C (775° F). Careful control of cooling rate is essential to obtain maximum softness and stability of properties.

Typical annealing conditions

Solution heat treatment improves mechanical properties by developing maximum concentration of hardening constituents in solid solution. This requires heating the metal to a temperature close to the eutectic temperature, holding it there for the required time, and cooling it fast enough to retain the desired solid solution. Wrought alloys frequently are treated in molten salt baths, usually sodium nitrate. Proper quenching is vital. As the metal cools, alloying elements start to precipitate, affecting the mechanical properties of the metal. Immersion quenching in water at room temperature is used primarily for wrought alloy products. Most casting alloys are quenched in hot (66–100° C [150–212° F]) water, which reduces internal stresses and minimizes distortion in the product. Hot oil and cold water quenching also are used for certain products.

Precipitation heat-treating, often called "artificial aging," is heating at elevated temperatures to harden the alloy. This process also is used where reduction of internal stresses is required to stabilize the dimensions of a part or to prepare the metal for extensive machining.

Stress relieving reduces internal pressures in a metal body. These pressures, or stresses, are unequal plastic deformations in the metal, caused by fabricating or temperature gradients or other external forces. Such residual stresses may be relieved by annealing or by mechanical operations, such as stretching or roller levelling, or by combinations of mechanical and thermal methods.

Joining. Aluminum can be joined by most methods used for other metals. Of 40 welding processes, aluminum has been commercially welded by 32. Aluminum can also be joined by brazing and soldering—that is, by welding with the addition of an alloy with a high melting point (brazing) or a low melting point (soldering). The light metal can be joined by a wide range of mechanical fasteners, including rivets, nuts and bolts, screws, nails, clips, and special structures. It can be adhesive bonded and, in lighter gauges, stitched or stapled.

Finishing. Aluminum lends itself to a wide variety of finishing techniques for appearance, or for corrosion, erosion, or abrasion resistance. These may be classified into mechanical, chemical, electrochemical, and electroplated finishes and applied coatings.

Mechanical finishing. Mechanical finishing operations are carried out to remove surface irregularities, smooth the surface, and apply a texture to the surface. Grinding, the rapid removal of surface metal, is accomplished with an abrasive belt or wheel. Polishing, a form of grinding, uses finer abrasives with special lubricants. Buffing brings out high lustre by eliminating surface marks produced during polishing. In buffing, a fine abrasive powder is embedded in a grease binder and rubbed on, instead of being glued to the face of the wheel. Satin finishing imparts a soft, smooth finish with low reflectance by means of fine, irregular lines scratched on the surface of aluminum with a rotating wire brush, buffing wheel, or steel wool.

Satin finishing

In barrel burnishing, abrasive chips (in water, lubricant, detergent, or other medium) rub against the metal in a rotating or vibrating container, imparting a bright, smooth surface. This is an economical way to finish small parts. In abrasive blasting, handsome matte or satin finishes are produced by subjecting the metal to an air blast or air–water blast into which has been introduced an abrasive, such as fine silica, glass beads, or steel shot. Shot peening is similar to abrasive blasting except that large-diameter steel shot is used. Peening cleans and textures aluminum surfaces, while at the same time it work-hardens the metal and improves its fatigue resistance. In embossing, patterns are applied to aluminum sheet by passing it through engraved rolls.

Chemical finishing. Chemical finishing treatments on aluminum function as cleaners, etchers, and depositors. Chemical cleaners prepare the surface for subsequent operations. Chemical etches roughen the natural surface layers of aluminum. Chemical depositors add metallic film or coatings to the surface.

Alkaline etches impart an inexpensive matte finish to aluminum. The most frequently used alkaline etchant is an aqueous solution of sodium hydroxide (caustic soda). An acid bath is necessary to remove smut produced by the chemical reaction. The surface is microscopically pitted, which gives the metal a characteristic nonreflective appearance. Acid etches achieve a variety of results. Mixtures of sulfuric and chromic acids are used as mild etches to remove stains made during heat treatment, or to prepare aluminum alloys for printing, anodizing, or other finishing operations. Mixtures of nitric and hydrofluoric acids produce semibright matte finishes on aluminum alloys. Concentrated solutions of hydrochloric acid are used to deep-etch aluminum. The addition of cobalt or nickel salts hastens the action, yet leaves a smooth surface.

Both alkaline and acid etchants are used widely in chemical milling, by which the metal thickness is reduced in a controlled manner. Dissolution rates of about 0.025 millimetre (0.001 inch) of metal per minute are typical. Chemical milling (also called deep etching) is used to obtain surface patterns through selective surface metal removal, or to remove metal from highly complex shapes, or to achieve overall reduction in dimension and weight of large aluminum parts in cases in which machining would be infeasible or too costly.

Immersion plating is a type of chemical finish in which thin films of other metals are deposited galvanically when aluminum is immersed in aqueous solutions of certain lead, tin, copper, silver, brass (zinc and copper), cadmium, iron, or zinc compounds. These have different purposes: zinc or tin may serve as a surface preparation for electroplating; tin, lead, or brass surfaces reduce friction; silver promotes surface electrical conductivity.

Electrochemical finishing. Anodizing is an electrochemical finishing process unique to aluminum; an electric current creates an aluminum oxide coating on the surface that is integral with the aluminum. Billions of pores in the film permit the electrolyte to reach the metal surface and the current to flow. Different alloying constituents in the base aluminum can produce a variety of attractive colours in the product. Also, high porosity permits the oxide coating to adsorb applied dyes, pigments, and lubricants. The oxide film can then be effectively sealed by immersion in pure boiling water. Thus, the colour or dye or additive becomes integral with the metal. Anodic coatings on aluminum are very popular for architectural applications.

Electroplating. Electroplated coatings are applied to aluminum alloys to obtain a specific appearance or to impart a special property, such as increased wear resistance, electrical conductivity or improved solderability.

Applied coatings. Aluminum's inherent resistance to corrosion makes it an ideal receiver for applied coatings such as paints and lacquers, plastic films, and porcelain

enamels. Organic coatings, paints, lacquers, and enamels can be applied by any of several conventional techniques, such as spraying, roller coating, dipping, and flow coating. Plastic films can be laminated to aluminum sheet, using proper adhesives. Plastic powders can also be applied by electrostatic spraying; particles, negatively charged as they leave the spray gun, are attracted to the object to be painted, which is grounded. The term porcelain enamel refers to a family of finishes with widely varying composition and properties. An opaque porcelain coating can be applied to aluminum by fusing a finely milled and pigmented glass frit (a partly fused mixture) to a prepared surface by high-temperature firing. Such coatings give aluminum the characteristics of glass: hardness, resistance to chemical attack, and colour fastness. Finally, certain metallic coatings can be applied to aluminum by vacuum metallizing, hot dipping, and spraying to add the properties of the applied metal in specific applications.

'orcelain namel

(K.R.V.H./Ed.)

Copper production

Copper is one of the most useful and widely used of the metallic elements. Although commercial deposits of copper ores occur in almost every continent, 70 percent of the world's known reserves are found in seven countries: Chile, the United States, the Soviet Union, Zaire, Peru, Zambia, and Mexico. The greatest known reserve of copper ore in one body is the deposit at El Teniente mine in Chile. Many lesser deposits are being exploited, such as those in Alaska, China, Australia, and Europe.

HISTORY

Copper was discovered and first used by humans during the Late Stone Age. Though the exact time of this discovery will probably never be known, it is believed to have been about 8000 BC. Copper is found in the free metallic state in nature; this native copper is the material that humans employed as a substitute for stone. From it they fashioned crude hammers and knives and, later, other utensils. The malleability of the material made it relatively simple to shape implements by beating the metal. Pounding hardened the copper so that more durable edges resulted; the bright reddish colour of the metal and its durability made it highly prized.

)iscovery)f casting ınd educing

The search for copper during this early period led to the discovery and the working of deposits of native copper. Sometime after 6000 BC the discovery was made that the metal could be melted in the campfire and cast into the desired shape. Then followed the discovery of the relation of metallic copper to copper-bearing rock and the possibility of reducing ores to the metal by the use of fire and charcoal. This was the dawn of the metallic age and the birth of metallurgy.

The early development of copper probably was most advanced in Egypt. As early as 5000 BC copper weapons and implements were left in graves for the use of the dead. Definite records have been found of the working of copper mines on the Sinai Peninsula about 3800 BC, and the discovery of crucibles at these mines indicates that the art of extracting the metal included some refining. Copper was hammered into thin sheets and the sheets formed into pipes and other objects. During this period bronze first appeared. The oldest known piece of this material is a bronze rod found in the pyramid at Maydūm (Medum), near Memphis, the date of origin being generally accepted as about 3700 BC.

As noted earlier, bronze, an alloy of copper and tin, is both harder and tougher than either; it was widely employed to fashion weapons and objects of art. The period of its extensive and characteristic use has been designated the Bronze Age. From Egypt the use of bronze rapidly spread over the Mediterranean area: to Crete in 3000 BC, to Sicily in 2500 BC, to France and other parts of Europe in 2000 BC, and to Britain and the Scandinavian area in 1800 BC.

About 3000 BC copper was produced extensively on the island of Cyprus. The copper deposits there were highly prized by the successive masters of the island—Egyptians, Assyrians, Phoenicians, Greeks, Persians, and Romans. Cyprus was almost the sole source of copper to the Romans, who called it *aes cyprium* ("ore of Cyprus"), which was shortened to *cyprium* and later corrupted to cuprum. From this name comes the English name copper. The first two letters of the Latin name constitute the chemical symbol Cu.

When copper and bronze were first used in Asia is not known. The epics of *Shu Ching* mention the use of copper in China as early as 2500 BC, but nothing is known of the state of the art at that time or of the use of the metal prior to that time. Bronze vessels of great beauty made during the Shang dynasty, 1766–1122 BC, have been found, indicating an advanced art. The source of the metals, however, is unknown.

The Copper Age in the Americas probably dawned between AD 100 and 200. Native copper was mined and used extensively and, though some bronze appeared in South America, its use developed slowly until after the arrival of Columbus and other European explorers. Both North and South America passed more or less directly from the Copper Age into the Iron Age.

Later uses for copper

As man learned to fashion his weapons from iron and steel, copper began to assume another role. Being a durable metal and possessed of great beauty, it was used extensively for household utensils, water pipes, and for marine uses and other purposes that required resistance to corrosion. The unusual ability of this metal to conduct electric current accounts for its greatest use during the 20th century.

ORES AND MINING

Mining techniques. Though it sometimes occurs in native form, copper is most often found mixed with other minerals in the form of various ores (see below). In these ores, the amount of copper may vary from less than 1 percent to more than 10 percent. When the ore contains small quantities of the desired metal, very large amounts of rock must be mined and processed to recover a profitable quantity of metallic copper. In deposits of this type, the open-pit method, involving the removal of ore from extensive deposits exposed by the large open pits, is used. In cases in which the percentage of copper in the ore is considerably higher and the deposit far below the surface, the underground method, which requires sinking a shaft to the deposit, is employed.

Important ores. Principal forms in which copper ores are found include native copper, porphyry copper, massive deposits, and mixed ores. Native copper is simply the metal found unadulterated in nature. Porphyry copper deposits, in which the copper materials are more or less uniformly scattered throughout the rock, account for the greatest tonnage of metal in the producing areas of the world. The copper minerals in the upper portions of such deposits are in general oxides (copper chemically combined with oxygen), those in the lower levels sulfides (copper with sulfur). The host rock is porphyry, schist, or other rock. Massive deposits are of higher metal content but of more limited extent and may be oxidized in the upper portion with sulfides lower down. In mixed ores, nickel, zinc, or lead usually accompany the copper; when such ore is mined, these other metals also are refined and sold as by-products. Table 5 shows the ore minerals of copper and their compositions.

Main source

ORE PROCESSING

Copper metal of very high purity is recovered from mined rock by a three-step process. Since much copper ore is found mixed with large amounts of material that does not contain copper, the first step (ore dressing) involves separating the ore from the gangue. The second step frees copper metal from the concentrate; *i.e.,* separates it from other elements, such as sulfur and oxygen, with which it is chemically combined. The third step, refining, involves removing impurities from the copper metal in order to produce a metal of very high purity.

Ore dressing. Copper ores usually are treated either by smelting (pyrometallurgical methods) or by leaching (hydrometallurgical methods). Generally speaking, sulfide

Table 5: General Classification of the Ore Minerals of Copper

	formula	copper (percent)*
Native copper ore		
Native copper	Cu	99.9
Sulfide ores		
Chalcocite	Cu_2S	79.9
Covellite	CuS	66.5
Chalcopyrite	$CuFeS_2$	34.6
Bornite	Cu_5FeS_4	63.3
Enargite	Cu_3AsS_4	48.4
Tetrahedrite	Cu_3SbS_2	46.7
Oxide ores		
Cuprite	Cu_2O	88.8
Tenorite	CuO	79.9
Malachite	$CuCO_3 \cdot Cu(OH)_2$	57.5
Azurite	$2CuCO_3 \cdot Cu(OH)_2$	55.3
Chalcanthite	$CuSO_4 \cdot 5H_2O$	25.5
Brochantite	$CuSO_4 \cdot 3Cu(OH)_2$	56.2

*Approximate.

ores are first concentrated by the selective flotation method described below, while oxide ores are leached.

Selective flotation. In selective flotation the finely ground ore, mixed with water and selected reagents, is violently agitated with air to produce a heavy froth. The reagents are so selected that a great attraction exists between the surface of the valuable mineral particles and the air bubbles, with only a small attraction between the gangue particles, which are readily wetted by the water, and the air. As a result, the mineral particles cling to the bubbles of the froth and are carried to the surface while the gangue particles sink to the bottom of the container. Removing the froth then separates the minerals from the gangue material. Perfect separation is impossible, but with constant supervision, repeated treatment, and exacting control it is possible to concentrate 95 percent or more of the copper ore into 10 to 20 percent of the original weight and to eliminate most of the worthless gangue.

Roasting. In roasting, the copper concentrate is heated in air to expel volatile matter, such as arsenic and antimony, and to expel part of the sulfur from sulfide ores by oxidation to produce sulfur dioxide gas and copper oxide. The material that results from roasting, known as calcine, is a fine mixture of the sulfides and oxides of iron and copper, gangue material, and nonvolatile impurities. The total sulfur content has been reduced to that desired for the subsequent smelting operation.

Metal recovery. To separate metallic copper from the substances with which it is associated in copper ores, smelting of matte or electrolytic techniques may be employed. Sometimes copper is recovered from scrap.

Matte smelting and converting. The object of the first step in smelting is to produce a molten, artificial sulfide of copper and iron. This material, known as matte, should contain all the copper and the desired amount of iron and must be sufficiently heavy so that, when melted, it will separate from the gangue material, other undesirable compounds, and the balance of the iron. When metallic oxides fuse in the presence of fire, they react as bases and will combine with acid anhydrides to form stable compounds of relatively low weight. These compounds, known as slags, are usually formed by the use of silica, SiO_2, as the acid anhydride. The formation temperature and fluidity of the slag is a function of its composition. The action of the slag on the furnace lining is a function of its acidity. It is necessary then to balance the composition of the slag to achieve the desired results. This is done by adding to the furnace selected materials to give a slag of the desired composition and properties. These materials are known as fluxes; the most common are limestone and silica.

Matte is generally smelted in large oblong furnaces from six to nine metres (20 to 30 feet) in width and from 27 to 40 metres (90 to 130 feet) in length, constructed with heavy, silica brick side walls and a low roof. These are called reverberatory furnaces because heat is radiated from the roof onto the material treated. Matte may also be smelted in blast furnaces or electric furnaces.

The conversion of matte involves the transformation of molten matte into copper metal (blister copper). Matte converting is based on the fact that copper has a lower affinity for oxygen than has iron or sulfur and the fact that the oxidation of iron and sulfur liberates large quantities of heat. Molten matte can be oxidized by introducing into it a stream of air. At the point at which the air enters, iron, sulfur, and copper are oxidized. The copper oxide immediately reacts with any iron sulfide still present to reform the sulfide of copper and form the oxide of iron. As air continues to flow through the molten mass, a time arrives when all the iron is present as the oxide and all the copper as the sulfide, and the sulfur that was originally with the iron has left the charge as sulfur dioxide. Since iron oxide forms a fusible slag with silica, it is necessary only to add to the molten mass sufficient silica to form the desired slag, and all the iron may be removed as an iron-silicate slag, while the heavy copper sulfide remains behind. The oxidation of the iron and its sulfur liberates sufficient heat to keep the material molten and to melt the necessary silica and any other solid charge that it may be necessary to add. If air continues to enter the charge after the removal of the slag, the oxidation of the sulfur and the copper continues, and the copper oxide so formed at once reacts with the copper sulfide remaining in the charge to form sulfur dioxide and metallic copper. These reactions liberate just sufficient heat to keep the charge molten. Ultimately, all sulfur is oxidized, and metallic copper and a small amount of copper oxide are present in the charge. Small amounts of minor impurities are also present.

Producing blister copper

Hydrometallurgy. Hydrometallurgy, or leaching, involves the treatment of the ore with a suitable solvent that takes the copper compound into solution and leaves all or a major part of the undesirable material unaltered. The copper is then recovered in relatively pure form from the solution.

The oxide ores of copper lend themselves to leaching. Sulfuric acid readily dissolves these oxides and has little effect on the common gangue materials, except for some of the salts of iron. Many oxide ores are not amenable to concentration, so it is necessary to treat or leach the ore as mined. The material is agitated or soaked with the leach solution. Elaborate systems of washing and filtering or settling are employed to recover the optimum amount of copper. After use, the leaching solutions are usually subjected to some type of purification for removal of the soluble iron and other objectionable impurities. This often involves neutralization and oxidation followed by removal of the precipitated ferric salts.

Electrolytic recovery. When copper ores are leached with sulfuric acid, the metal may be recovered from the leach solution, in which it exists in the form of copper sulfate, by passing an electric current through the solution (electrolysis); this breaks down the sulfate and yields copper. An insoluble anode (positive terminal of an electrolytic cell) and a copper cathode (negative terminal) are used in the process; they are immersed in the leaching solution (copper sulfate), and an electric current is passed through the cell, and copper from the solution is deposited on the cathode.

Electrolytic decomposition of copper sulfate

Recovery from scrap. Because of its durability, large quantities of scrap or reclaimed copper become available each year from obsolete or discarded machinery and equipment. This material is readily converted into usable metal by its addition to smelter feed, or by remelting and refining.

Since the early 1970s, pollution problems and economic considerations have brought some shifts in the use of various methods of ore dressing and metal recovery. Smelting has been under criticism in many places because of the release of sulfur dioxide to the atmosphere. Continuous smelting processes that minimize the problem were developed in Australia and Canada, and a combined roasting and smelting process, called flash smelting, in Finland and Canada. Higher prices for copper at the same time were encouraging the use of leaching, effective in exploiting low-grade copper-bearing material. In this procedure, dilute sulfuric acid is allowed to trickle through the ore, extracting the copper in the form of a solution of copper

sulfate. The copper is recovered from the solution by precipitation on scrap iron.

Refining. Refining involves the removal of impurities from copper metal to yield a product containing less than 1/2 of 1 percent of impurities.

Fire refining. Blister copper contains copper oxide and small quantities of other impurities. Some of these impurities can best be removed by fire refining, which is done either in small reverberatory furnaces or in revolving furnaces, equipped with fuel burners that melt the charge, if necessary, and maintain it in a molten condition throughout the refining operation. Air is forced through the molten material to ensure complete oxidation of all impurities, and then the oxides are allowed to rise to the surface of the quiet pool from which they are skimmed. The oxidizing treatment is followed by a reducing treatment known as poling, in which the ends of green logs are forced into the pool of molten metal. The highly reducing gases resulting from the destructive distillation of the green logs reduce (remove the oxygen from) most of the copper oxide present in the metal. This reduction also can be effected by bubbling reformed natural gas through the bath. During this operation, frequent samples are examined to determine the degree of deoxidation. When the proper stage has been reached, the sample has a distinctive metallic rose colour. This metal is called tough-pitch copper, and may be readily cast into dense slabs.

Commercial tough-pitch copper is usable in this form. It still contains, however, any gold and silver recoverable from the original ore and traces of other impurities. The gold and silver are often worth recovering, and the removal of other impurities to produce the pure metal can be brought about by electrolytic refining.

Electrolytic refining. When an impure copper anode and a copper cathode are immersed in a solution of copper sulfate and sulfuric acid and an electric current is passed through the cell, copper ions from the solution migrate to the cathode and are deposited as metallic copper, and copper from the anode enters the solution. For each ion deposited on the cathode, an ion enters the solution from the anode. Impurities in the anode are liberated and settle to the bottom of the cell as a slime or accumulate in the solution. The net result is the consumption of the impure anode and the deposition of pure copper on the cathode along with the formation of a cell slime that contains the gold, silver, and other insoluble components.

(A.W.S./Ed.)

THE METAL AND ITS ALLOYS

The major portion of the world's production of copper is utilized by the electrical industries; most of the remainder is combined with other metals to form alloys.

Essentially pure metal. Typical samples of electrolytic copper contain from 99.92 percent to 99.96 percent copper. About 0.03 percent oxygen is purposely left in the copper, since this amount slightly improves density and conductivity. Copper in this condition has a conductivity of 100 percent to 102 percent of the International Annealed Copper Standard. On this standard, 100 percent denotes a resistance of 0.15328 ohm for a length of one metre (39.37 inches) weighing one gram (1/28 ounce) at 20° C (68° F); this standard has been universally adopted for industrial purposes.

Making copper wire

Electrical conductors. For making copper wire, electrolytic copper is first cast into wirebars, which are made in several standard sizes varying in weight from 60 to 227 kilograms (135 to 500 pounds). The wirebars are then reheated to 700° to 850° C (1,290° to 1,560° F) and are rolled without further reheating to rods approximately 10 millimetres (0.375 inch) in diameter. The rod is drawn cold into wire, through dies of successively smaller diameters until the desired size is reached. The dies are usually of tungsten carbide; for finer wires, diamond dies are used.

Much copper wire is marketed in the form of bare coils; a considerable tonnage is subsequently covered with paper, fabric, rubber, plastic, or other insulating material for use in the form of covered conductors. Much of the wire is also supplied stranded; all of these operations are carried out on special machines that are largely automatic in operation.

Copper cables are often covered to render them resistant to moisture. Lead covering, extruded onto the outside of the cable by means of a special apparatus, often is used; hemp or metal armouring sometimes may be added for additional protection.

The electrical industries also use large quantities of bare copper strip for incorporation in electrical machinery. This ribbonlike form of the metal is produced mainly from wirebars that are rolled in a mill similar to that used for the production of wire. Copper strip of greater width and much thinner gauge also is produced in long lengths and supplied in the form of coils.

Sheet and strip. The term copper strip as distinct from copper sheet is usually applied to material less than 61 centimetres (24 inches) wide that is supplied in long lengths. The majority of the strip used is less than 30 centimetres (12 inches) wide. In the preliminary stages of manufacture, the copper castings are rolled hot, but in the later stages of manufacture all the rolling is carried out cold, the material being coiled on coiling drums on each side of the rolling mills. Material produced by this method is of very even gauge and possesses an exceptionally good surface finish. The coils can be handled easily and are in general use for the manufacture of stampings in the electrical and other industries. Copper strip is supplied in various degrees of hardness according to the rolling it has received subsequent to the last annealing. These various tempers are selected according to the amount of subsequent mechanical deformation to which they will be subjected.

Hardness of copper strip

Copper sheets are produced by somewhat similar methods of manufacture. In the United States the majority of copper sheets are made from electrolytic copper, but in Europe fire-refined arsenical copper is used frequently in the manufacture of sheets and plates.

Alloys. In variety of uses, the alloys of copper surpass all other nonferrous alloys and comprise mixtures of copper with zinc, tin, nickel, aluminum, lead, manganese, and other elements.

Brass. Brass is an alloy consisting mainly if not exclusively of copper and zinc. The brasses may be conveniently divided into two groups in terms of their malleability, the dividing line being approximately the composition of 55 percent copper and 45 percent zinc. All the higher copper alloys are workable either hot or cold and in some cases both hot and cold, while the remainder are not malleable at all. The unworkable brasses, known as the white brasses, are not industrially important.

The general mechanical properties of brass vary widely; indeed, it is this wide range in tensile strength, elongation, and hardness that makes brass such an important alloy. Brass is readily drawn into fine wire, rolled into very thin strips, or drawn into tubes and extruded as rods or sections (see also below *Zinc production*).

Bronze. Bronze, an alloy formed by adding tin to copper, fuses at a lower temperature than copper and is thus better suited for casting; it also is harder and less malleable. A soft bronze or gunmetal is formed from 16 parts of copper to one of tin; a harder gunmetal, such as that used in the past for bronze cannon, contained about eight parts copper for each part of tin.

Phosphor bronze

Bronze is made harder and stronger when it is alloyed with phosphorus. Alloys prepared in this way, known as phosphor bronzes, may contain only about 1 percent of phosphorus in the ingot and a mere trace after casting, but their value is nevertheless enhanced for purposes in which a hard, strong metal is required, as for pumps, plungers, valves, and the bushings of bearings. Bronze is also improved by the presence of manganese in small quantities. Various grades of manganese bronze, in which there is little or no tin but a considerable percentage of zinc, are used in mechanical engineering.

Nickel. Because copper and nickel are completely miscible (mix thoroughly) in the solid state, forming a complete series of solid solutions, the useful range of alloys is not confined within any definite limits of composition, although certain compositions have come into general use.

Additions of 2 to 45 percent nickel to copper provide a series of alloys that are considerably stronger and more resistant to oxidation at high temperatures than is copper. Of these cupronickels, the one containing 30 percent nickel is the most important; it is widely used for steam-condenser tubes.

The alloy formed of 20 percent nickel with the remainder copper is one of the most ductile of commercial alloys and may be subjected to the most severe cold-working without the need of any intermediate annealing. It is also readily forged and rolled at a temperature above 800° C (1,470° F). These properties make it a very suitable alloy for drop forgings and cold stamping and pressing. It has also found a variety of uses in automobile construction for exposed fittings, as it takes a high polish and is resistant to atmospheric tarnishing. Other uses include bullet sheathing, a widespread application. Another alloy in this series, containing either 45 percent or 40 percent nickel, became widely known under the name constantan. It has a high electrical resistance, which remains almost constant over a wide temperature range.

Monel metal

Monel metal is a so-called natural alloy prepared by the reduction of a copper-nickel ore; it contains 65 to 70 percent nickel, iron and manganese in small amounts, and certain impurities that influence its properties to some extent. It has been widely used for various engineering and ornamental purposes, and possesses exceptionally high strength at both normal and elevated temperatures. Alloys of similar nickel content are also manufactured by melting nickel and copper together.

Beryllium–copper. Unlike many kinds of steel, most copper alloys are not susceptible to improvements of hardness and strength by processes of heat treatment. One useful exception is the heat-treatable alloy beryllium–copper. This consists of copper with addition of about 2 percent of beryllium, with or without a smaller addition of nickel or cobalt.

When beryllium–copper is heated to about 800° C, quenched in cold water, and then reheated to 275° C (525° F), it develops a tensile strength comparable to some of the stronger varieties of steel.

Other copper alloys. Copper also forms an important series of alloys with aluminum, classed under the general term aluminum bronzes. They may be classified into two main groups: those containing up to 7.5 percent aluminum are extremely ductile, whereas those containing 8 to 11 percent possess high tensile strength in the cast state. The ductile alloys containing less than 7.5 percent are especially useful for deep stamping, spinning, and severe cold-working of all kinds. They are useful as a substitute for brass and possess greater strength and resistance to atmospheric corrosion.

Silicon bronze usually contains about 96 percent copper. The remainder may be silicon alone, but more often a little manganese, tin, iron, or zinc is also added. These alloys were developed originally for the chemical industry because of their exceptional resistance to corrosion in many liquids. Their application later extended far beyond this field, chiefly because of their good casting qualities, strength, hardness, and ease of welding.

Manganese bronze

Manganese bronze is made in several varieties, exhibiting a range of compositions and properties. One type is in reality a brass to which a very little manganese has been added as a deoxidizer, less than 0.5 percent manganese remaining in the alloy. Another kind contains 2 to 5 percent manganese together with 2 to 4 percent iron and 3 to 7.5 percent aluminum. It has exceptionally high strength and is called high-tensile manganese bronze, or manganese-aluminum bronze.

(A.Bu.)

CHEMICAL COMPOUNDS

Copper forms two series of compounds, one in which it exhibits a valence (degree of combining power) of 1 (cuprous) and the other a valence of 2 (cupric); several unstable compounds in which a valence of 3 is exhibited are also known. Since the cuprous ion is unstable in aqueous solution, its salts readily decompose to form the metal and cupric salts.

Oxides. Copper forms two oxides in accordance with its two valences: cuprous oxide, Cu_2O, and cupric oxide, CuO. Cuprous oxide, a red crystalline material, can be produced by electrolytic or furnace methods. It is reduced readily by hydrogen, carbon monoxide, charcoal, or iron to metallic copper. It imparts a red colour to glass and is used for antifouling paints. It is soluble in mineral acids to form colourless cuprous salts, most of which rapidly oxidize to the cupric state. Cupric oxide, a black powder, can be prepared by the ignition of suitable salts such as the carbonate, the hydroxide, or the nitrate of copper, or by heating of cuprous oxide. This compound oxidizes carbon compounds and finds a wide laboratory and commercial use for this purpose. Since it imparts a green colour to glass, it is used extensively for that purpose. It is soluble in mineral acids and forms with them blue or green solutions.

Halides. Cuprous chloride, CuCl, can be prepared by treating metallic copper and cuprous oxide with hydrochloric acid or by treating metallic copper and cupric chloride with hydrochloric acid. The hydrochloric acid solution of cuprous chloride readily absorbs carbon monoxide and acetylene and is used for this purpose in gas analysis. Cupric chloride, $CuCl_2$, can be prepared by dissolving cupric oxide in hydrochloric acid. This material finds some use as the base salt for the manufacture of pigments. Cuprous iodide, CuI, is prepared by the direct combination of copper and iodine. Cupric iodide, CuI_2, exists only in combination with ammonium salts or in complex organic compounds.

Sulfates. Cupric sulfate, $CuSO_4$, commonly known as blue vitriol, is the most important salt of copper. It usually crystallizes as $CuSO_4 \cdot 5H_2O$ and has a bright blue colour. It is prepared by the treatment of copper oxides with sulfuric acid. While readily soluble in water it is insoluble in alcohol. The anhydrous salt is hygroscopic and is useful as a desiccating agent. Copper is readily displaced from aqueous solutions of the salt by metallic iron. Copper sulfate is the basic salt in the electrolytic refining of copper, and it also finds a wide use in the preparation of pigments.

Blue vitriol

Carbonates. Basic copper carbonates are formed when an alkaline carbonate is added to the solution of a copper salt. These compounds, which have a bright blue or green colour and are used in the preparation of pigments, occur in nature as the minerals azurite and malachite.

Other compounds. Copper forms a series of salts with arsenic, all of which are bright green in colour and poisonous. Wide use has been made of these compounds in the production of insecticides; some copper salts are used as superficial antiseptics. Cupric nitrate, $Cu(NO_3)_2$, can be prepared by dissolving metallic copper in nitric acid. The hydrated crystals are deep blue in colour. Copper silicates occur in nature and impart blue colour to many minerals.

(A.W.S.)

Zinc production

Zinc, one of the most important industrial metals, is a metallic element of moderate hardness that can be made ductile and easily worked with moderate heating. Its most important use, as a protective coating for iron, derives from two of its outstanding characteristics: it is highly resistant to corrosion, and, when in contact with iron, it provides sacrificial protection, corroding in place of the iron. Its second most important worldwide use is die-casting alloys. Zinc alloys melt easily, quickly assume the shape of even the most complicated die, and do not attack the steel dies into which they are injected.

Alloyed with copper, zinc forms the important group of alloys, the brasses. In sheet form, zinc is widely used to form the cans of dry batteries. The chemical compounds of zinc, particularly zinc oxide, have important industrial and pharmaceutical uses. Some 6,000,000 tons of zinc are mined per year.

HISTORY

Because the separation of pure zinc from its ores is difficult, the ancients used the metal in alloys for millennia before it was recognized as a metal in its own right.

Biblical and other ancient references to brass, an alloy of copper and zinc, are not reliable; many of them probably were erroneous identifications of bronze, an alloy of copper and tin.

Antiquity of brass

Brass was, however, known at least 3,000 years ago. In an early method of preparing the alloy, granulated copper was mixed with charcoal and zinc oxide derived from a natural zinc ore; the mixture was heated in a sealed crucible of clay or other refractory material. The zinc vapour from the reduction of the oxide combined with the copper to form brass.

Evidence suggests that zinc production originated in India and China. Zinc was used for coins in China at the beginning of the 15th century. Smelting was carried out in clay crucibles filled with a mixture of zinc ore and charcoal and fitted with necks in which the zinc vapour cooled and condensed.

An advance in zinc smelting took place near Bristol, England, in 1740 when William Champion operated a process to make zinc from zinc carbonate ore by distillation. This was a crucible method generally similar to the presumed Chinese method, but with a 100-millimetre-(four-inch-) diameter pipe extending downward from the crucible bottom into another crucible. At distillation temperature, zinc vapour passed down the pipe and condensed to liquid zinc, which was collected. Toward the end of the 18th century, zinc distillation was practiced in Silesia. A Belgian process, the basis of the present-day process using horizontal retorts, gradually developed during the 19th century.

ORES AND MINING

Principal zinc ores. Zinc ores are widely distributed throughout the world. The major producing countries include Canada, the Soviet Union, Australia, Peru, the United States, Japan, and Mexico. Zinc blende or sphalerite (zinc sulfide) is the major mineral, but oxidized ores such as the zinc carbonate smithsonite and a hydrated zinc silicate, hemimorphite, are also of considerable importance. Sphalerite is commonly found with lead ore (galena, or lead sulfide) and with iron sulfide, or pyrite. Calcium carbonate, in the mineral form calcite, is the most common waste material or gangue in zinc deposits.

Most of the world's zinc is extracted from geological replacement deposits—*i.e.,* deposits in which a mineral in solution gradually replaced an earlier formed mineral—in limestones and dolomites. Examples of ore bodies of this type include the large Sullivan Mine in British Columbia, Canada, and Broken Hill in New South Wales, Australia. The latter deposits contain—besides zinc—lead, silver, and minor but valuable amounts of copper and cadmium, and worthless components consisting principally of manganese silicates, calcite, fluorite, and silica.

Mining techniques. Zinc ores are recovered by many mining techniques, ranging from open-cut mining—mainly in the case of oxidized ore bodies, which are located closer to the Earth's surface—to the normal underground methods, used for the more deeply located sulfide ores.

Among the individual deposits at Broken Hill, variations—in concentration of ore, in size and shape of the ore body, and in strength of the surrounding stone—have necessitated the development of a flexible system that incorporates both open-cut and deep-mining methods.

But the most common method of ore extraction is cut-and-fill stoping, in which mine tunnels are dug to moderate depths, branching away from the mine portals. As in most mining, the trend has been toward increased mechanization.

REFINING AND RECOVERY

Beneficiation. The small fraction of zinc sulfide mineral present in the ore makes beneficiation (upgrading) necessary in order to produce a concentrate usable for treatment. The most common method for accomplishing this concentration is by froth flotation, a specialized form of liquid separation to isolate the sulfide mineral from the impure constituents (gangue) in the ore. In this process, the ore initially is crushed to about 1.9 centimetres (0.75 inch), combined with water, and ground to less than 0.1 millimetre (0.004 inch) in a ball mill, consisting of a horizontal cylinder containing various sizes of iron or steel balls which tumble and thus pulverize the ore as the mill is rotated at high speed. The finely ground particles and water form a slurry that flows from the mill to flotation cells or tanks, where, in the presence of selected chemical frothing agents that create a suspension of air bubbles, the slurry is agitated by beaters. The mineral particles cling to the bubbles and float to the surface, forming an oily froth that is constantly skimmed. The gangue material is wetted by the action of the chemicals and sinks in the cell. The proper choice of frothing agents makes it possible further to treat complex lead and zinc sulfides and economically to separate out each constituent mineral in a concentrated form.

Roasting and sintering. Zinc concentrates are roasted prior to reduction to convert zinc sulfide to zinc oxide and to produce sulfur dioxide gas, which is in turn converted to sulfuric acid. The roasting reaction gives off heat, and no fuel is required once the reaction is initiated. Roasting zinc sulfide at 1,200 to 1,300° C (about 2,190° to 2,370° F) to produce zinc oxide calcine has generally superseded other processes for the production of fine, powdery calcines. Two main roasting processes are operated. In one, associated primarily with the electrolytic extraction method, the finely divided particles of the concentrate are suspended, or fluidized, in a rising stream of air; the resulting powdered oxide is suited for leaching in acid. The other is a sintering process, in which the material becomes agglomerated into lumps. In one sintering process, new zinc concentrates are mixed with an excess of returns (*i.e.,* fine sinter separated or crushed from the lump sinter output), fed onto a moving grate, and fired. The resulting dense and lumpy product, by virtue of its strength and hardness, is an ideal feed for the blast furnace. Sometimes calcines are sintered with carbon to make improved feed material (briquettes) for vertical retort furnaces.

Principal processes

Extraction metallurgy. Each of the five main processes described below for obtaining zinc from the concentrated sulfide ores begins with conversion to the oxide. In all the processes except the electrolytic, carbon is used to bring about the reduction of zinc oxide.

In the two retort processes the reduction of zinc oxide by carbon yields zinc gas and carbon monoxide gas. Heat for the reaction is supplied from an external source. The concentration of zinc in the gas phase is high. In the blast-furnace process, heat is provided by the direct combustion of carbon in air; a special, rapid condensation technique is necessary to prevent the reoxidation of the zinc vapour by the large amount of carbon dioxide present in the hot gas mixture. The electrothermic process is described later.

The reduction of zinc oxide by carbon always produces the metal in the vapour form because the reaction does not take place at an appreciable velocity at temperatures below the boiling point of zinc, which is 906° C (1,663° F).

Horizontal retorts. Present-day practice is still based on the earlier Belgian retort furnace. A horizontal vessel of round or oval cross section, about 23 centimetres (nine inches) in diameter and 150 centimetres (60 inches) long, manufactured from fireclay and silica, is loaded with a mixture of roasted ore and coal; the zinc distills and is collected over a 48-hour cycle in a removable condenser unit fitted to the end of the retort. Such zinc recovered from condensers is called spelter and forms about 75 percent of the total zinc thus recovered. The rest is in the form of blue powder, or zinc dust. The daily output of such a retort is from 27–45 kilograms (60 to 100 pounds); several hundred retorts may be banked together and fired by gas. Improved methods of operation at some plants include mechanical loading machines.

Vertical retorts. An improved process features a continuous retort reaction, in which the vertical retort, constructed of silicon carbide brick, has a rectangular cross section of two metres by one-third metre (seven feet by one foot) and a height of about 11 metres (35 feet). The load is fed to the top of the retort, and residues are extracted at the base through a water-sealed quenching (cooling) section. A gas-fired combustion chamber heats the walls of the retort.

Continuous retort reaction

Load preparation is an important preliminary to smelting. Anthracite coal or coke breeze (the residue from coke making) and roasted concentrates are carefully sized and mixed prior to making them into briquettes, which are preheated in a coking furnace and fed into the top of the vertical retort.

Reaction gases consisting almost wholly of zinc vapour and carbon monoxide are removed near the top of the retort and the zinc vapour is condensed in a well-stirred, enclosed zinc bath. Output from each retort averages seven to nine tons per day; a typical plant may operate with about 20 retorts.

The electrothermic process. The electrothermic process generates the necessary heat for the reduction of zinc oxide by carbon through the direct electrical-resistance heating capabilities of the coke constituent of the charge. A typical furnace shaft is 2.4 metres (eight feet) in diameter and more than 12 metres (40 feet) in height, with graphite electrodes fixed near the top and bottom of the shaft spaced about nine metres (30 feet) apart. Preheated coke and sinter is loaded by a rotary distributor to achieve the optimum distribution and temperature pattern (around 1,300° C [2,400° F]). The spent load is removed below the lower electrodes by use of a rotary discharge table. Zinc vapour and carbon monoxide are led from the furnace to a condenser through special vapour ports. The average daily production of an electrothermic unit is more than 55 tons.

Zinc blast furnace. In early attempts to operate a zinc blast furnace, the main difficulty was in the condensation of zinc vapour from a gas containing oxidizing reaction products. This problem has now been overcome. Roasted zinc and lead concentrates (sinter) and metallurgical coke are loaded into a shaft furnace, in which rising air burns the fuel (Figure 25). Preheated air is blown into the furnace through nozzles. The furnace products are a zinc-bearing

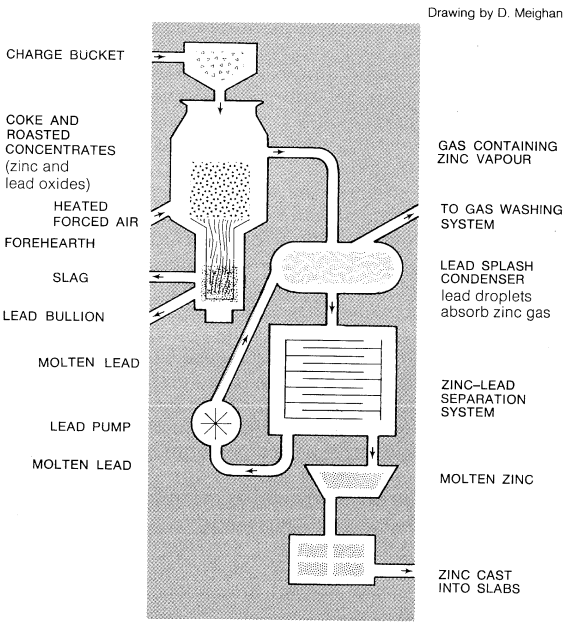


Figure 25: Stages in the production of zinc in a zinc-lead blast furnace.

gas and liquid phases, which separate at the bottom of the furnace and are tapped periodically. The liquids consist of molten lead, reduced from the lead oxide content of sinter and containing recoverable copper and silver, and the gangue content of the load in the form of a molten oxide slag.

Gas containing 6 percent zinc, 11 percent carbon dioxide, and 18 percent carbon monoxide leaves the upper shaft and passes to a lead splash "condenser"; *i.e.,* a chamber in which an intense shower of lead droplets is thrown up by rotors revolving in a pool of molten lead. The zinc is absorbed into the liquid lead, from which it later separates as a liquid layer on cooling.

A typical shaft furnace with a cross-sectional area of 17 square metres (185 square feet) has an annual production capacity of some 60,000 tons of zinc and 30,000 tons of lead, though one furnace at Avonmouth (Bristol, England) has a 50 percent higher capacity.

Electrolytic process. First operated commercially around the end of World War I, the electrolytic process accounts for about half the world's zinc production. The basic steps in this process include, first, the preparation of a solution consisting of zinc oxide calcines (from the roasting of sulfide concentrates) and dilute sulfuric acid; and second, the electrolysis of the resulting zinc sulfate solution.

The theoretical voltage required to deposit zinc from zinc sulfate solution is about twice the voltage necessary to decompose water, and the electrolysis at that voltage should result in the production of hydrogen at the cathode. If the cathode is zinc, however, overvoltage prevents the release of hydrogen and, hence, zinc is deposited. A critical step in the process is the purification of the zinc sulfate solution before electrolysis. Impurities such as copper, cadmium, cobalt, and germanium, which would interfere with the electrolysis or contaminate the zinc product, are precipitated from solution by stirring in zinc dust. The electrolysis is performed in lead-lined concrete cells with anodes of lead containing 1 percent silver and cathodes of aluminum sheet. The zinc deposits are stripped from the cathodes every 24 hours and remelted in an induction furnace before casting into ingots or pigs.

Refining techniques. In the electrolytic zinc process, the purity of normal production is 99.95 percent. It is possible to make a special high-purity metal of 99.99+ percent purity suitable for die-casting alloy production. Such purity can be achieved by adjusting the conditions of electrolysis—temperature, current density, etc.—or by adding strontium carbonate to the electrolyte to reduce the amount of lead deposited.

Zinc produced by pyrometallurgy contains lead, cadmium, and iron as impurities and is refined by fractional distillation. The operation is performed in two fractionating columns, each consisting of a series of superposed rectangular trays made of bonded silicon carbide refractory material and arranged to allow a descending flow of liquid metal and an ascending flow of vapour. In the first column, a large part of the zinc is vaporized and freed from higher boiling-point impurities such as lead and iron. The distilled vapour is condensed and fed into the second column, where the liquid's remaining impurity, cadmium—with a boiling point lower than that of zinc—is distilled, and high-purity liquid zinc is run off from the bottom of the column. Fractional distillation

A continuous vacuum distillation integrated into a suitable blast furnace production line has made an intermediate grade of zinc which contains less than 0.1 percent lead.

Relative importance of extraction methods. Since about 1960, most new zinc-smelting capacity has been divided almost equally between the electrolytic process and the zinc blast furnace. The electrolytic process may be used in areas of cheap electrical power; the blast furnace can treat complex and mixed ores.

THE METAL AND ITS ALLOYS

The total world consumption of zinc is used primarily in the form of protective coatings for iron and steelwork, in die casting, in brass making, and in sheet zinc. The remainder is used in paints, pharmaceuticals, and other products.

Zinc coatings. Zinc coatings are used to protect a wide range of steel products including structural steelwork, sheet, tube, wire, and nails from corrosion.

There are five principal methods of application, with the largest tonnage consumed by galvanizing. In this process steelwork is first pickled in acid, treated with chemicals to prepare it for fusion with the galvanizing agent, and then dipped into a bath of molten zinc at about 440° C (820° F). A series of iron–zinc alloy layers is formed on the surface, ending with an outer layer of zinc. Galvanizing

Steel strip and wire are galvanized in a continuous process. A small aluminum content in the zinc bath reduces

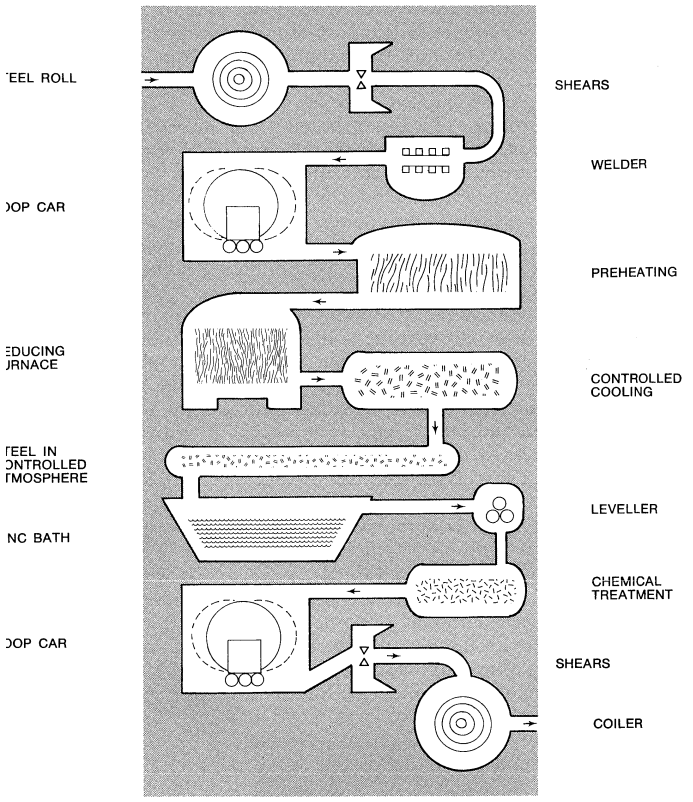


Figure 26: Stages in the production of galvanized steel strip.
Drawing by D. Meighan based on *Technical Notes on Zinc-Galvanizing*; Zinc Development Association

the thickness of the alloy layers and assists the formation of a highly ductile coating. Figure 26 is a diagram of a typical continuous galvanizing line for steel strip.

Galvanizing protects steel by providing a barrier between the steel and a corrosive atmosphere and, more importantly, it protects through its sacrificial electrochemical action whereby the zinc is corroded instead of the steel. Electrogalvanizing (or zinc plating) is an alternative method of coating steel and is used when controlled thinner coatings are required.

Zinc spraying. Atomized particles of molten zinc can be sprayed onto cleaned and roughened steel from a special gun fed with wire or powder. The process is useful for applying protective coatings to structures that are too large for the dipping process.

Sherardizing. Small articles may be heated in a revolving drum of heated zinc powder, in which case an alloy layer forms on the surface.

Zinc-rich paint. Paint formulated with a high content of zinc dust gives an electrically conducting dry film. The coating gives the same protection as galvanizing and is widely used in protecting steelwork and in repairing damaged zinc coatings.

Die castings. The pressure die-casting process is one in which molten alloy is injected under pressure into a steel die. High production rates can be achieved, and the resulting components have good dimensional stability and surface finish; they also can be plated to produce a highly decorative finish. The alloy used is based on high-purity 99.99-percent zinc to which 4 percent aluminum and 0.04 percent magnesium have been added. A slightly harder alloy contains 1 percent copper in addition. The automotive industry uses more than half the total weight of die castings made—as plated functional and decorative components and trim and also as carburetors. Zinc alloy castings also are used widely in domestic appliances, locks, door handles, bathroom fittings, toys, and a variety of engineering components.

Applications of zinc alloy castings

Brass. Alloys of zinc with copper are called brass. Alpha brasses—*i.e.*, those with a content of up to 37 percent zinc—have a single metallographic phase and are suitable for rolling, pressing, and drawing, and for small castings. The alpha–beta brasses, with 40 percent to 45 percent zinc, have two metallographic phases and are used in casting, hot pressing, and extruding. The brasses have high strengths, good corrosion resistance, and good electrical conductivity; they have wide domestic and industrial applications.

Rolled zinc sheet and strip. The major consumption of rolled zinc sheet and strip is in the dry-battery industry and in the building and printing trades.

Battery cells

Dry-battery primary cells normally are made from zinc strip from which circular or hexagonal blanks are stamped; the blanks are then impact-extruded in automatic presses into cylindrical cans, which form the anodes of Leclanché cells. Higher voltage cells are made in a layer-type construction of zinc sheets with carbon cathodes.

Zinc sheet is widely used in the building trade in Europe for roof cladding, gutters, rainwater pipes, and flashing. It is also used in the form of special plates in printing for photoengraving and lithographic work. In the 1960s the old production method for zinc sheet, consisting of hot-rolling a pack of two or more sheets, was first supplanted by a continuous process in which a continuous-strip casting machine is placed in line with rolling mills.

Zinc alloys. Wrought zinc alloy markets are still largely traditional, but other alloys based on "Zilloy" alloys have been developed by the New Jersey Zinc Co. (U.S.). These are zinc alloys containing approximately 1 percent copper and 0.1 percent titanium. On rolling, the alloy particles of titanium and zinc having the composition $TiZn_{15}$ are aligned in the rolling direction forming "stringers" that are the main strengthening feature of the alloys, restricting grain movement and providing much improved creep strength; *i.e.*, resistance to permanent deformation by the continuous application of a stress. Because of their fine grain size these alloys have good ductility at normal and low temperatures.

The alloy is amenable to a number of sheet-forming and strip-forming processes that require bending, stamping, cutting into blanks, or stretching. A wide range of components including electric motor brush boxes, coaxial cable clamps, TV antenna mounting plates, and automobile keys are in current production from zinc–copper–titanium alloy strip. Roll-forming processes to make complicated sections are also feasible.

In billet (*i.e.*, bar) form the alloy may be extruded, although the extrusion pressure is usually higher than for either brass or aluminum for an equivalent reduction ratio or billet size. Forging and cold-stamping operations also have been applied successfully to the alloy.

Another development has been an alloy of zinc containing 22 percent aluminum, which may attain commercial importance by virtue of its superplastic properties; *i.e.*, its ability to be formed as easily as a plastic at elevated temperatures while retaining normal metal properties at room temperature. Vacuum- and pressure-forming techniques have been applied to sheet material in a wide range of potential applications, from parts for the electronics industry to automobile body panels. Forming output may be lower than in conventional metal-pressing operations, but this disadvantage is offset by the low capital cost of thermoforming equipment. The superplastic alloy has offered a totally new way of fabricating metal parts.

Superplastic alloy

Table 6: Typical Properties of the Important Zinc Alloys

property	alloy		
	pressure die-cast Zn, 4.0 % Al, 0.04 % Mg	wrought Zn, 1.0 % Cu, 0.1 % Ti	superplastic Zn, 22.0 % Al
Tensile strength			
ton/in²	18.5	12.5	16.0
kg/mm²	29.1	19.7	25.2
Elongation (%)	15	40	50
Brinell hardness kg/mm²	83	55	70
Creep strain per hour under stress (%)*	0.05	0.005	1.0

*12,000 lb/in² (8.4 kg/mm²).

Table 6 gives typical properties of the pressure die-cast, wrought, and superplastic alloys.

CHEMICAL COMPOUNDS

Zinc oxide. Two main processes are employed for producing zinc oxide, a white powder. In the direct, or American, method of manufacture, zinc ores (or residues) are heated in air with coke or anthracite, and the resulting zinc vapour is subjected to controlled oxidation. In the indirect, or French, process, the zinc vapour for oxidation is obtained by boiling zinc. The purity of the zinc depends on the starting materials; the various grades have different uses. A major use is as an accelerator in the vulcanization of rubber (automobile tires contain up to 5 percent zinc oxide). It is also used in paints, toughening the film, preventing yellowing, and resisting mold growth. Zinc oxide is known to have semiconducting properties related to which is the specific property of light sensitivity or photoconductivity, use of which has been made in photocopying processes. Miscellaneous uses include incorporation in ceramics and enamels and in lubricants.

Other industrial compounds. Zinc sulfide in a suitably activated form—*i.e.,* with trace quantities of certain elements—can exhibit fluorescence, phosphorescence, and luminescence. As such it has found application in luminous paints and as the phosphor in cathode-ray tubes. Lithopone, which is a mixture of zinc sulfide and barium sulfate, is used as a white pigment in paints and mastics.

Zinc sulfate and zinc chloride are used in a wide range of comparatively small-scale applications. The former is used in agriculture as a weed killer and to give protection against pests. It is important in the manufacture of viscose rayon as a constituent of the precipitating bath. Zinc chloride has applications in the textile industry and as a flux constituent in soldering, aluminum refining, and galvanizing processes.

Zinc chromates are used as corrosion inhibitors and also as bright yellow pigments. Zinc phosphate, in addition to providing corrosion protection to iron and steel, is available in paint form as a pigment in an organic binder and as such is used as an anticorrosive primer on steel.

Medical compounds. Pharmaceutical-grade zinc oxide is a high-purity oxide manufactured by the indirect French process from high-purity zinc metal. It is used in the preparation of ointments, lotions, and cosmetics. Zinc carbonate is a mild astringent to the skin and as an ingredient of calamine lotion is used to alleviate skin irritations. Zinc peroxide finds application as an oxygen-liberating compound and in the form of an antiseptic ointment is applied to wounds and sores.

(A.W.R./Ed.)

Lead production

Lead has found a wide range of applications throughout history and in terms of world tonnage today remains one of the most important metals produced.

HISTORY

The Egyptians may have used lead as early as 5000 BC. Applications in the ancient world included use as weight standards, in coinage, for ornamental objects, for refining silver, as solder, and as a glaze for pottery. The Egyptians used red lead, an oxide of lead, as a cosmetic pigment. A lead figure dating from 3800 BC was found at the temple of Osiris, on the site of ancient Abydos, and is now located in the British Museum. About 2000 BC the Phoenicians were mining lead deposits in Spain, and it was at about the same period that the Chinese first employed lead in their monetary system. Later, when silver coins replaced lead, the mining of lead was prohibited in order to prevent counterfeiting, but lead became and remained the principal metal used to debase silver for counterfeiting purposes into modern times.

Lead was well known in biblical times and is used metaphorically in Exodus in noting that the pharaoh's host "sank as lead" in the Red Sea. The Hanging Gardens of Babylon, built about 600 BC, are said to have had floors sheathed with sheets of lead soldered together to hold the soil and moisture.

The lead mines at Laurium, Greece, flourished in the 5th century BC, and in the 3rd century BC the Romans operated mines in the Río Tinto region of Spain. The Romans used lead pipe extensively, and fabricated it in 15 standard-sized diameters, all three metres (10 feet) long. Many sections of pipe in almost perfect condition have been unearthed at Pompeii, Rome, and Roman sites in England. Indeed, the Latin word for lead was *plumbum,* denoting water conductors or spouts; from it is derived the word plumber. The Romans also made wide use of lead in soldering.

Lead in Roman times

Lead applications multiplied in the Middle Ages; lead "cames" held the fragments of stained-glass windows together, and lead gutters, drainpipes, and interior pipes were fashioned for churches, monasteries, castles, and houses.

Almost since the beginning of warfare, lead has been important in missiles because of its low melting point (which made it easy to fabricate) and its high density. Lead chunks were hurled with slings, and molten lead was poured on attackers from fortress walls. Lead bullets and shot have been used since the inception of firearms.

The medieval alchemist, who preceded today's chemist and metallurgist, considered lead the father of all metals, and its low melting point was a factor that led him to try to convert it to a substance that would transform other base metals into gold.

In the United States, now the world's major lead-producing country, lead mining and smelting began in Virginia in 1621. Lead ores from the Mississippi Valley were used from 1700 to 1867, when the first of the great western mines was opened.

ORES AND MINING

Principal lead ores. Galena (lead sulfide), by far the most important primary lead ore mineral, frequently contains as impurities zinc, copper, cadmium, bismuth, arsenic, antimony, and silver; often the silver value is greater than that of the lead, in which case it is a true silver ore. A primary mineral is formed by natural processes and unaffected by subsequent events, while a secondary mineral is affected in some way; for example, by weathering. Other primary lead minerals are clausthalite, altaite, jordanite, kobellite, boulangerite, aikinite, jamesonite, and bournonite. Carbonates and sulfates, formed by weathering, are the chief secondary minerals. The most common, cerussite (lead carbonate), is formed by the action of carbonate groundwaters on galena. Anglesite (lead sulfate) is derived from galena that has been contacted by sulfate solutions generated from the oxidation of sulfide minerals. Other oxide minerals of lesser importance are wulfenite, vanadinite, and pyromorphite. Lead is rarely found in its elemental form.

More than 95 percent of mined lead is from primary ores in which galena is the preponderant lead mineral. Galena is usually intermixed with other sulfides, chiefly sphalerite (zinc sulfide) and the sulfides of iron: pyrite and marcasite, less commonly with chalcopyrite (copper and iron sulfides). Lead ores constitute the chief source of silver, usually in solid solution with galena.

Mining techniques. The widely varying characteristics of lead ores have given rise to many mining techniques, of which underground methods predominate. Most lead deposits occur at moderate depths in narrow channels called veins or in deposits of roughly spherical shape (lodes). Both veins and lodes are of varying thicknesses. Holes are drilled for blasting, often by large mounted diamond drills or carbide percussion drills. After blasting, the broken ore is sorted and loaded onto cars or continuous belts for transportation to the surface. Most modern lead mines are highly mechanized.

ORE TREATMENT AND REFINING

Although galena contains 86.6 percent lead, it usually occurs dispersed with other minerals, and in some places, ores with as little as 3 percent to 30 percent lead are mined.

Preparation for roasting. Lead ore is prepared for roasting by grinding, flotation, and concentration. First, the

ore is ground and classified by particle size: approximately 70 percent is less than 0.043 millimetre (0.0017 inch) in diameter, and 90 percent less than 0.074 millimetre (0.0029 inch); some particles are less than 0.0076 millimetre (0.0003 inch). The ground and classified ore, which is a mixture of lead sulfide and gangue, passes to large tanks where flotation takes place.

Flotation separation process

Flotation, which has supplanted earlier methods of mechanical and gravity separation, takes advantage of the fact that water, with a suitable wetting agent added to it, wets gangue but not lead sulfide. In the flotation tank, the ground and classified ore is mixed with water and some flotation agent such as oil. The lead sulfide attaches itself to the oil, which is turned into froth by bubbling air into the mixture, and then is floated off the top. The wet gangue sinks to the bottom of the tank and is discharged.

Concentration involves aggregating the tiny particles of lead sulfide; this is done with the aid of an agent such as lime or alum. Filtered to reduce its water content to about 10 percent, the concentrate then passes to the smelter. Concentrates include 45 to 60 percent lead, along with varying amounts of impurities: 10 to 30 percent sulfur, up to 15 percent zinc, 8 percent iron, 3 percent calcium oxide, 3 percent copper, 2 percent antimony, and small amounts of arsenic, bismuth, gold, silver, and other elements.

Roasting. Before lead concentrates can be added to the furnaces, they must be treated to lower the sulfur content because the ordinary furnace does not reach a high enough temperature to reduce lead sulfide. Substances such as limestone, iron ore, silica, coke, etc., are mixed with ore concentrate; these fuse with the sulfur in the lead sulfide to form compounds. The ore mix is laid on a moving grate; air is blown through the mass; sulfur acts as the main fuel, forming sulfur dioxide, which is usually recovered for sulfuric acid manufacture. The mix is fused into a strong, homogeneous material called sinter, comprising primarily oxides of lead, iron, and zinc, with substantial amounts of metallic lead and a little sulfur. The porous sinter mass is broken into lumps as it is discharged from the moving grate. Sinter clinkers up to 125 millimetres (five inches) in size are passed over grates to remove the fine particles, which are stored for blending with other materials prior to reduction in the blast furnace. Reduction is the removal of oxygen from a compound; for example, reduction of lead oxide yields metallic lead.

Smelting. Smelting is the continuous reduction of ore constituents to metal, generally carried out in a blast furnace, which consists mainly of a water-jacketed shaft or column up to 7.2 metres (24 feet) high with a rectangular cross section and is loaded from the top. The load, or charge, consists of properly proportioned sinter, limestone, coke, silica, and other fluxes. Orifices called tuyeres, through which air is blasted, are positioned just above a crucible at the bottom of the shaft. The crucible collects molten lead and is arranged with a suitable siphon well so that the molten lead can be removed from the furnace without affecting the blast pressure within the furnace. Ports to discharge slag and metallic waste (matte) are located near the top of the crucible. Sight glasses that allow for visual inspection of the smelting reaction are located in the tuyeres. Air is injected to aid combustion and to complete the formation of metallic oxides, which are then chemically reduced to metal by the coke and carbon monoxide present. Nonmetallic waste forms a slag with the fluxing materials. A balanced smelting charge of iron, limestone, and silica yields a slag that is fluid. When the coke and sinter ratio is not correct, there is an improper reduction of ore to metal.

Blast-furnace operation

Reverberatory furnace smelting, in which a furnace charge is held in a basin-like vessel and heated by flames on the surface to effect the reduction, is rarely practiced in lead refining as it requires that the ore contain 70 percent or more lead.

The output of the blast furnace is lead bullion, which contains dissolved metallic and nonmetallic (oxide and sulfide) impurities. Slag, matte, and other products of the blast furnace are either further treated to recover metal values or scrapped.

Drossing. Lead bullion is maintained in a molten state just above its melting point in a drossing kettle, which contains up to a hundred tons. (The term dross refers to any solid scum floating on top of a metal bath comprising metal oxides, sulfides, metallic compounds, etc.) At temperatures of about 330° C (626° F), remaining impurities are brought to the surface by stirring, by the addition of small amounts of sulfur, and by agitation with submerged air lances and are removed by skimming.

Refining. The bullion is transferred to the refinery either as a pumped liquid or as large cast blocks, weighing as much as 10 tons.

Electrolytic refining. Electrolytic refining involves purifying the bullion by means of electricity. If electrolytic refining is to be practiced, the bullion is cast into anodes (or plates) weighing several hundred pounds. These are hung in electrolytic tanks where they dissolve; pure lead is deposited on a thin sheet of lead that serves as the cathode. Impurities left behind can be recovered by many complex operations.

Kettle refining. Kettle refining of bullion is complex, but the recovery of by-product silver and gold is simpler and faster. All of a number of kettle-refining processes or modifications have the following steps: softening, desilverizing, dezincing, and debismuthizing. Softening refers to removal of impurities that harden lead, such as antimony, arsenic, tin, zinc, and others. Large-capacity steel pots or reverberatory furnaces are used to hold the bullion at red heat. Blowing with air or applying suitable oxidizing reagents oxidizes impurities. The oxidation is selective, and oxide drosses or slags in which each element is concentrated can be removed by skimming.

Oxidation of impurities

Silver is removed from softened lead by adding zinc and raising the temperature sufficiently to melt it (about 540° C [1,000° F]). With stirring, the molten zinc reacts, forming compounds with gold and silver, plus traces of copper and nickel. These zinc compounds are lighter than the metallic lead; on cooling, they form a crust that is removed. The gold and silver are recovered in a parting plant. The lead is now treated for removal of the remaining zinc by a vacuum process. The zinc recovered from the vacuum retort, along with that from the parting plant, is reused.

If the bismuth content is too high, the softened lead is next debismuthized. The operation is similar to desilverizing, except that a combination of calcium and magnesium metals are used as the reagents to form high-melting-point intermetallic compounds with bismuth, which are skimmed off and treated for recovery of bismuth.

The refining processes yield pig lead of good marketable uniformity and purity. Large quantities of pig lead are available, 99.999 percent pure.

LEAD PRODUCTS

Lead is widely used both in the metallic form and in chemical compounds.

Storage batteries. The lead-acid storage battery is the largest single use for lead. Economy, reliability, and the ability to deliver large amounts of energy in a short time have made it an essential part of every automobile. Storage-battery grids require a metal with a low melting point, high corrosion resistance, and certain electrical properties. The modern automotive SLI (starting, lighting, ignition) battery consists of about one-third lead metal (grids, posts, and connectors) and two-thirds litharge (lead monoxide). The lead metal, 6 to 12 percent antimonial lead plus a small amount of tin, is almost wholly recovered by smelting secondary materials, mostly scrap storage batteries. Litharge mixed with sulfuric acid to a paste consistency for incorporation into the metal grid is usually made from primary (mined) lead because impurities such as silver and bismuth cannot be removed economically during refining of secondary (salvaged) lead.

Chemical lead. Lead is used extensively in the construction of chemical-plant equipment. If unalloyed lead is used it is generally of the grade commonly known as chemical lead, sometimes termed acid lead, or copper lead. The term chemical lead denotes a metal of very high purity: standards vary slightly from country to country. Chemical lead forms a tightly adherent, insoluble lead sulfate film

in sulfuric acid solutions. As sheet, pipe, rod, etc., it is readily welded (burned) into complex structures for holding and conveying sulfuric acid mixtures. Other materials, such as plastics, rubber, etc., have been used for similar operations, but lead's demonstrated corrosion resistance over a wide range of conditions, its ease of repair, and, after a long useful life, its high salvage value have enabled it to retain its position as a material of construction in the chemical industry. A 6-percent-antimony–chemical-lead alloy is used where greater strength or fatigue resistance is required. The use of 6-percent alloy is restricted to temperatures below 125° C (257° F) because of its lower melting point. Similar alloys are used with phosphoric acid, hydrofluoric acid, chromic acid, and other corrosive environments where insoluble protective films are formed. Each different smelter uses a proprietary alloy as insoluble anode in chromium plating, and electrowinning (recovery by electrolysis) of copper. To achieve a higher order of strength, lead-lined and lead-covered steel pipe, fittings, valves, and vessels are used with lead metallurgically bonded to the substrate to maintain suitable heat transfer.

Other uses. The electric cable in industry is a major user of lead. Lead alloys, usually 1 percent antimonial lead, can be easily extruded, forming an impervious corrosion-resistant cover on telephone and power cables. When the cable is replaced, the sheathing can be recovered by a simple stripping operation, refined, and used to cover new cable.

Lead has a longer record of satisfactory performance than any other water-service pipe material. The flexibility of lead pipe allows adjustment to ground settlement without damage, and minimizes the number of joints. Lead is also used for caulking joints in cast-iron water mains and sewer and soil pipes. Because it withstands the effects of polluted air and water, lead is useful in drainage and venting systems.

Architectural applications

Sheet-lead roofing and flashing have long been used on buildings because of their durability under all types of climatic conditions. The development of special alloys capable of resisting buckling due to alternating high and low temperatures (thermal cycling) has caused a renewed interest in the use of lead in modern architecture. Special joint designs have also helped in the construction of essentially maintenance-free roofs, reflecting pool liners, etc., and soft lead cames are easily worked to form leaded art-glass windows.

Ammunition, both military and sporting, accounts for a large use of lead. Spherical shot for shotgun use is made by pouring molten 2 to 6 percent antimonial lead, with up to 1 percent arsenic added, through suitable sieves at the top of a 38-metre (125-foot) tower. The molten alloy, while dropping, forms a true sphere before solidifying near the bottom of its fall. The shot is collected in water, removed, dried, and sorted for size and sphericity. Military ammunition is formed into bullets by swaging (cold-shaping) extruded wire in suitable dies.

The densest of the common metals, lead is widely used for radiation shielding. It prevents the penetration of gamma rays and neutrons, and since it does not itself become radioactive, it may be used continuously. High-density glasses, containing as much as 80 percent lead oxide, are used as windows in radiation installations.

Due to its limpness and high density, lead is finding expanding use in the control of noise pollution. Thin-cast sheet lead is used for encasing noisy machinery, or in room partitions for keeping noise out. Jet airliners use a lining of an adhesive-backed layer of vinyl plastic compounded with lead powder to keep the noise level in the passenger compartment to a minimum.

LEAD ALLOYS AND COMPOUNDS

Lead alloys. The mechanical properties of pure lead are quite poor; therefore, when more strength is required, lead is alloyed with other elements.

Antimony. The most common metal used is antimony, which on additions up to 12 percent by weight not only increases the tensile strength of the alloy as much as four times but also lowers the melting point. As noted earlier, antimonial lead is used as a storage-battery grid metal.

Tin. Tin alloyed with lead in amounts from 2 to 70 percent by weight forms the solders widely used in industry. Not only is the strength of the alloy increased with higher tin content but the melting point decreases from 327° C (621° F) for lead to 183° C (361° F) for 63 percent tin–37 percent lead. Solders of 2 percent tin–lead are used for automatic soldering of tin-coated steel cans. Industrial solders, used for such applications as automotive radiators, contain from 5 to 30 percent tin. The electronic solders are generally of the lowest melting point (63 percent tin), so that the temperature-sensitive components will not be damaged.

Industrial soldering alloys

Casting alloys. Lead alloys containing antimony and tin form a series of low-melting metals with excellent casting properties and the ability to produce intricate castings with excellent detail. Linotype metal with 4 percent tin and 12 percent antimony, used for casting the finest printer's type, exhibits suitable strength for use on a press. Electrotype and stereotype metals used by printers also have excellent casting properties. Ornamental castings, which are usually plated with silver or other metals, are made with 11 percent antimonial lead, with about 0.5 percent arsenic added to increase the strength and to yield better detail and a smooth-surface finish. Lead alloyed with varying combinations and amounts of bismuth, tin, cadmium, antimony, indium, or gallium yields a series of metals that have very low melting points, culminating in one so low that it can be melted in the palm of the hand. Such alloys, rated as to melting point, are widely used for equipment of very high reliability.

Bearing metals. Lead-base bearing metals are usually antimonial lead with additions of tin or copper or both and arsenic to increase the compressive strength. The ability of these alloys to maintain a thin film of lubricating oil at relatively high loads and speeds prevents wear on the bearing-metal surface. The alloys are sufficiently soft so that even in case of lubrication failure no damage is done to the rotating shaft. The use of thin layers of bearing metal, with a stronger backing material reinforcing the lead alloy, increases the load-carrying capacity of such bearings.

Lead compounds. At least a third of all lead produced is used in chemical compounds. Most of the uses of lead discussed thus far are not consuming uses, in that the metal eventually returns as scrap for secondary refining and reuse. The chemical compounds, which generally use corroding lead (lead refined to a high degree of purity) as a base, are generally dissipated into the atmosphere or elsewhere so that recovery is impossible.

Gasoline additives. A major use of lead compounds in the 1960s was in gasoline additives, such as tetraethyl lead to form the antiknock, high-octane gasoline required by high-compression internal combustion engines. This use has declined since the 1970s because of air-pollution problems.

Paint pigments. The earliest known and one of the most widely used lead paint pigments is white lead (basic lead carbonate). A process for making white lead starts with metallic lead blown with steam or air to produce a fine powder, which is then treated with acetic acid, air, water, and carbon dioxide gas in long, revolving wooden cylinders. The process requires about 12 to 14 days for conversion. Recent modifications of the process use litharge as the starting reagent to shorten the production time. The outstanding characteristic of white lead is its chemical affinity for paint vehicles, becoming an integral part of the paint film. This reinforcement results in a tough, flexible film, unusually resistant to embrittlement.

Basic lead sulfate, also called basic sulfate white lead, is a pigment used in mixed paints with the same properties as basic carbonate white lead.

Red lead (Pb_3O_4) is manufactured in furnaces by heating litharge at a carefully controlled low temperature. Litharge takes up more oxygen to form red lead. A part of the litharge usually remains unconverted in the core of red lead particles. A 98 percent true red lead is used for paint.

Litharge. Litharge (lead monoxide) is one of the most important and widely used metallic compounds. It is produced by melting pig lead in the presence of air, or

occasionally by burning lead to a fume in the presence of oxygen. It is possible to produce a specific type of litharge for almost any use, and the various commercial grades are generally known by the uses to which they are put. Storage batteries account for the largest consumption of litharge. Battery litharge is very pure and may be blended with some red lead and other additives to form the active material of the plates. Glassmaker's litharge is a high-purity flake or powder, which must be free of iron or other impurities that would discolour the glass. Similarly, pottery maker's and enameller's litharge must be pure. Porcelain-enamelled aluminum sheets have litharge as a major constituent in the coating. Colourmaker's litharge is reacted with acetic acid to form lead acetate. Various compounds added to the lead acetate precipitate a number of pigments, such as chrome yellow, chrome red, and chrome green.

Other compounds. Lead arsenate, an insecticide, can be prepared by dissolving litharge in arsenic acid, with acetic or nitric acid as the catalyst, or litharge may be converted to lead acetate or nitrate and allowed to react with sodium arsenate.

Lead azide is an initial detonating agent for high explosives, such as TNT. Crystalline lead azide is prepared by the slow addition of a dilute solution of sodium azide to a constantly stirred dilute solution of lead acetate or nitrate. The precipitated azide is washed with water and stored under water or a 50–50 mixture of water and ethyl or methyl alcohol. By varying procedures it can be prepared as a colloid or treated with starchy compounds called dextrins to achieve certain properties.

Lead dioxide (PbO_2) is formed electrolytically by anodic oxidation of metallic lead. In neutral or alkaline solutions the alpha crystalline form of lead dioxide is formed. In acid solutions the beta crystalline form results. Both forms have been identified in the positive plates of storage batteries.

Lead compounds as vinyl stabilizers

Unless stabilizers are added, the widely used vinyl plastics (polyvinyl chloride) suffer a breakdown due to heat and light, rendering the plastic brittle and discoloured. Stabilizers take up hydrogen chloride released in the degradation. Theoretically, litharge would be a very efficient acid acceptor, but the nature of vinyl stabilization goes beyond mere acid acceptance, and various other chemical and physical actions must be incorporated into a desirable stabilizer or combination. A series of reaction products of litharge with various organic and inorganic acids are widely used as vinyl stabilizers.

Lead oxide was used as a major constituent of glass by the Chinese as early as the 6th century. The glass was strongly discoloured by impurities such as copper. The use of lead oxide in glass to achieve clarity was unknown until the invention by George Ravenscroft of English crystal, or flint glass, in 1675. Lead oxide yields a glass with a high index of refraction, which accounts for the brilliance of cut lead crystal and for its use in optical lenses. It also lends weight and resonance to the glass. (E.J.M./Ed.)

Production of other important industrial metals

As noted above, iron, aluminum, copper, zinc, and lead are the most important metals based on tonnage of metal refined. There are a number of other metallic elements with wide industrial application, some of the more significant of which are treated in this section. (Ed.)

CALCIUM

Calcium was so named by the English chemist Sir Humphry Davy because of its occurrence in chalk. Though this metallic element does not occur in the free state, its compounds are widely distributed. It is the fifth most abundant element in the Earth's crust; it constitutes 3.63 percent of the igneous rocks and 3.22 percent of the entire crust of the Earth.

Occurrence. Calcium is found in gypsum and related minerals. Forms of calcium occur as fluorite and (with calcium phosphate) as apatite. Many minerals, notably feldspars and zeolites, contain calcium silicates. Asbestos is composed of calcium magnesium silicate. Calcium phosphate occurs as the mineral phosphorite. Calcium carbonate (calcite) occurs in limestone, chalk, marble, dolomite, eggshells, pearls, coral, stalactites, stalagmites, and the shells of many marine creatures.

Production. In 1808 Davy showed that lime is a compound of oxygen and a metal, which he named calcium. While calcium was formerly produced by electrolysis of anhydrous calcium chloride, practically all commercial production now is by the reduction of lime by aluminum in heated retorts under low pressures. Calcium distills out of the reaction mass and is collected in a cool section of the retort or in a condenser. Calcium metal is costly and production is small.

Major uses

The metal. Calcium is used in making alloys of aluminum, copper, lead, magnesium, and other base metals. It is an important deoxidizer for chromium–nickel, chromium–nickel–iron, and related resistance high-temperature alloys and for nickel, steel, and tin bronzes. It is an evacuating agent (getter) to remove residual gases in the production of high-vacuum products such as electron tubes. It has been employed as a reducing agent in the preparation of chromium, thorium, uranium, zirconium, and other metals from their oxides; as a dehydrating agent for organic liquids; a desulfurizer for petroleum fractions; and a decomposing agent for thiophenes and mercaptans.

Calcium alloys. Calcium–silicon, made in the electric furnace from lime, silica, and carbon, is a deoxidizer and degasifier for iron, steel, and high-tensile-strength cast irons. It is added either to the furnace in the final stages of processing or at the time of casting. Both calcium and silicon are active deoxidizers and their reaction products are a slag with a low melting point, giving a clean steel ingot with improved fluidity and freedom from holes (blowholes) caused by captured gases.

Calcium–manganese–silicon, also made in the electric furnace, similar to calcium–silicon but with manganese ore as an addition, is sometimes preferred to calcium–silicon for the additional effect of the manganese.

Calcium compounds. Calcium carbonate ($CaCO_3$, or calcite) occurs in all limestones, other rocks and minerals, and in some naturally hard waters. Precipitated chalk is calcium carbonate prepared by mixing solutions of calcium chloride and a soluble carbonate. Tooth powders, white polishes, and cements are made with calcium carbonate.

Lime, calx, or quicklime is calcium oxide (CaO), a strongly caustic ingredient used for making mortars and plasters. Calcium hydroxide [$Ca(OH)_2$], slaked lime, is important in the production of mortars, plasters, and cements. In these materials, which set as they dry, the hydroxide combines with carbon dioxide in the air to form calcium carbonate. Lime has many uses in industry as a cheap alkali and in agriculture as an agent for removing excess soil acidity.

Calcium peroxide (CaO_2) is obtained by the addition of hydrogen peroxide to limewater; the resulting calcium hydrate ($CaO_2 \cdot 8H_2O$) loses water on being heated to 130° C (266° F) and forms the peroxide—a pale, buff-coloured powder sometimes used for bleaching and antiseptic purposes.

Calcium phosphide (Ca_3P_2), a highly flammable substance, is obtained by passing phosphorus vapour over strongly heated lime. It is used in marine signal lights because it decomposes on contact with water to produce spontaneously inflammable hydrogen phosphide.

Calcium phosphate [$Ca_3(PO_4)_2$] occurs as phosphorite [$Ca_3(PO_4)_2$] in Florida, Tennessee, parts of the western United States, and northern Africa. The acid salt [$CaH_4(PO_4)_2$] is obtained by evaporating a solution of the normal salt in hydrochloric or nitric acid; it is very soluble and, with calcium sulfate, is a superphosphate plant food, produced by treating mineral phosphates with sulfuric acid. In medicine calcium phosphate is used as a source of phosphorus. It is used as a polishing powder in dentifrices. Calcium phosphate is used in manufacturing milk glass, porcelains, pottery, and enamelling. It is a mineral supplement in animal feeds.

Calcium sulfate ($CaSO_4$) occurs in nature as gypsum, alabaster, anhydrite, satin spar, selenite, and terra alba. In water it is one of the salts causing permanent hardness; *i.e.*, hardness not removed by boiling. Calcium sulfate, a white

polishing powder, is used to make plaster of Paris, which is a semihydrate ($CaSO_4 \cdot \frac{1}{2} H_2O$), and cements, particularly Keene's cement. It is a white pigment in paints. In paper manufacturing it is used as filler, size, and glaze.

Calcium sulfide (CaS), erroneously called sulfurated lime, is produced by heating the sulfate with charcoal or by heating lime in a current of hydrogen sulfide. It has been used medically to treat acne and boils. Luminous calcium sulfide, called Canton's phosphorus, used in luminous paints and varnish, is made by igniting a mixture of calcium carbonate and sulfur with tiny amounts of bismuth or manganese salts.

Calcium chloride ($CaCl_2$), water soluble and moisture absorbing (deliquescent), occurs in many natural waters and is a by-product of several industrial processes. The crystals may contain two, four, or six molecules of water for each gram molecular weight of the salt. The hexahydrate and dihydrate are used for antifreeze and refrigerating solutions and for laying dust on roads. The anhydrous chloride, obtained by heating the crystals above 200° C (392° F) in a current of hydrogen chloride, is used as a drying agent. Bleaching powder is usually a mixture of the basic chloride [$CaCl_2 \cdot Ca(OH)_2(H_2O)$] and the basic hypochlorite [$Ca(OCl)_2 \cdot 2Ca(OH)_2$]. It is made by exposing thin layers of slaked lime to chlorine at 30° to 40° C (86° to 104° F).

Calcium fluoride (CaF_2) occurs as the mineral fluorite and can be prepared as a white precipitate by mixing solutions of calcium chloride and sodium fluoride. It is used to increase the fluidity of slag in ferrous metallurgy. Single pure crystals are used in spectroscopy.

Calcium hydride (CaH_2), sometimes called hydrolith, is obtained by heating the metal in a current of hydrogen. This compound is easily transported and on treatment with water yields hydrogen. It is used as a portable source of hydrogen. Calcium hydride is used to reduce rare metals from their oxides and as a drying agent.

Calcium silicide, calcium boride, and calcium carbide are used in metallurgical work as gas-removing agents, as reducing agents, and as sources of calcium for alloys. Often, however, in (highly specialized) alloy work, examples of which are the nickel–chromium and nickel–chromium–iron materials for high-temperature service, these compounds may introduce undesirable materials from which the metallic calcium is free. (C.L.M./Ed.)

CHROMIUM

Chromium is unusual among metals in that its ores were used extensively long before the pure metal was prepared. As early as 1800, they were used to make pigments and chemicals for leather tanning. In 1879 they were successfully used as refractories. In 1893 Henri Moissan obtained pure chromium through reduction of chromic oxide by coal in an electric furnace, and in 1898 Hans Goldschmidt obtained carbon-free chromium through the reduction of this oxide by aluminum. By 1910 the importance of chromium ores in metallurgy was established.

Preparation of pure chromium

Ores and ore treatment. Chromium is never found in nature in the free state. Its chief ores are made up of the mineral chromite, $FeO \cdot Cr_2O_3$. The actual composition of high grade ores varies between 42–56 percent Cr_2O_3 and 10–26 percent FeO, together with varying amounts of MgO, Al_2O_3, and SiO_2.

The majority of chromium ores can be concentrated with only hand sorting. This process can yield a concentrate containing about 50 percent Cr_2O_3, but the chromium-to-iron ratio is often not affected.

Chromium ores are divided into three groups: metallurgical, refractory, and chemical, depending on their end use. The metallurgical ore must be hard and lumpy, contain a minimum of 48 percent Cr_2O_3, and have a chromium-to-iron ratio of 3 to 1 for maximum usefulness. The ores used for refractories must be high in Cr_2O_3 plus Al_2O_3 and low in iron. Magnesium oxide is added in excess to these ores to form magnesium silicates on firing. For chemical purposes, ores should be high in Cr_2O_3 and low in Al_2O_3 and SiO_2; the specifications for this type are determined by price, availability, and experience.

Production of chromium metal. Chromium metal is produced commercially by two methods: by electrolysis, or by the reduction of purified chromium compounds, usually Cr_2O_3.

Basic methods

Reduction processes. Aluminum has been used to reduce Cr_2O_3 to produce large amounts of chromium metal by the reaction

$$Cr_2O_3 + 2Al = 2Cr + Al_2O_3.$$

This reaction is self-sustaining and is carried out by igniting a mixture of oxide and aluminum powder in a refractory-lined container. A 97–99 percent chromium metal is thus obtained, with aluminum, iron, and silicon the chief impurities. Such a metal also contains about 0.03 percent carbon, 0.02 percent sulfur, and 0.045 percent nitrogen.

Silicon is also used to reduce Cr_2O_3 and produce chromium metal:

$$2Cr_2O_3 + 3Si = 4Cr + 3SiO_2.$$

This reaction is carried out in an electric furnace since it is not self-sustaining. The resulting metal may contain up to 0.8 percent silicon.

Carbon may also be used as a reducing agent for Cr_2O_3 to produce chromium metal:

$$Cr_2O_3 + 3C = 2Cr + 3CO.$$

This reaction is carried out by briquetting a mixture of finely divided oxide and carbon and heating the briquettes to 1,200°–1,400° C (2,200°–2,500° F) in a refractory vessel at a reduced pressure of 280 to 315 micrometres of mercury for reduction at 1,400° C. Low pressure causes some chromium volatilization. The product contains 0.015 percent C, 0.001 percent N_2, 0.04 percent O_2, about 0.02 percent Si, and less than 0.03 percent Fe.

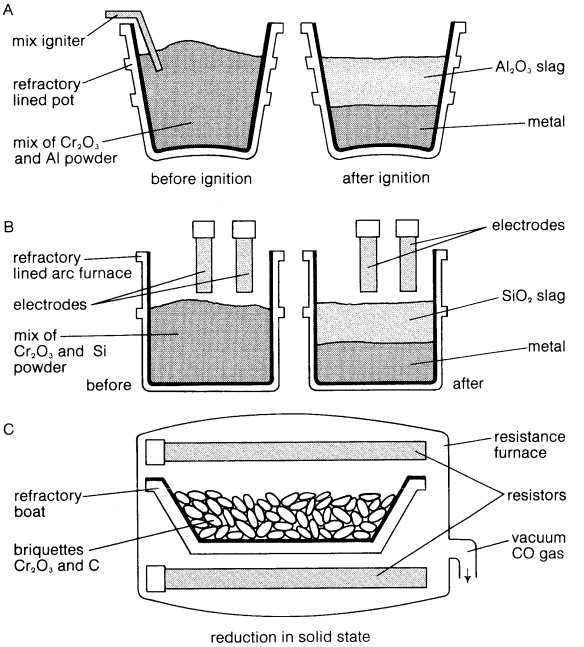


Figure 27: *Stages in the production of chromium metal by reduction processes.*
(A) Aluminothermic, (B) silicothermic, (C) carbothermic.

Electrolytic processes. There are two commercial processes for the electrolytic recovery of chromium metal. One uses a chrome–alum electrolyte and the other a chromic acid electrolyte. Although the chrome–alum electrolyte may be prepared from chrome ore, the difficulties encountered in preparing a pure solution make this process uneconomical. A commercially available product, high-carbon ferrochromium, is used as a starting material.

Crushed ferrochromium is slowly fed to a brick-lined steel tank in which it is leached (treated with a solvent) by a mixture of reduced anolyte (anode electrolyte), chrome–alum mother liquor, and sulfuric acid. A heating coil in the tank keeps the temperature near the boiling point, and a condenser controls the resulting spray. Leaching and digestion time is 48 hours. The slurry (liquid mixture)

is then transferred to a holding tank and cooled to 80° C (176° F) by the addition of cold mother liquor. The silica and undissolved solids are filtered from the solution. By heating at elevated temperature for a few hours, the chromium in the filtrate is converted to the non-alum-forming condition. The temperature is then lowered to about 5° C (41° F), and crude iron-sulfate crystals are formed that are removed from the mother liquor by a vacuum drum filter. The crude iron sulfate is dissolved by the liquor that results from washing the leach residue. Ammonium sulfate is added, and the solution is conditioned at elevated temperatures for several hours. The ferrous ammonium sulfate formed is recovered and sold for fertilizer or other purposes.

The mother liquor, free from the crude iron sulfate, is clarified in a filter press and aged at about 30° C (86° F). About 80 percent of the chromium is stripped as ammonium chrome–alum. The slurry is then filtered and washed, and the filtrate is sent to the leach circuit while the chromium–alum crystals are dissolved in hot water to produce cell feed.

To obtain good cell operation, close pH (hydrogen ion concentration) control of the catholyte (cathode electrolyte) is essential, as is the presence of divalent chromium at the cathode. A diaphragm cell is used to prevent the chromic and sulfuric acids formed at the anodes from mixing with the catholyte.

The flow of the catholyte through the diaphragm into the anolyte compartment controls the pH. The sulfuric acid in the anolyte is held at 250–300 g/l (grams per litre) by the addition of water.

In 72-hour cycles the cathodes are removed from the cells, the adhering salts removed by hot water, and the brittle chromium deposit (3.2–12.7 millimetres [0.125–0.5 inch thick]) is stripped by hand. This material is crushed, washed with hot water, sealed in stainless-steel cans, and heated to remove water and hydrogen.

An overall current efficiency of 45 percent can be obtained, which includes the low efficiencies prevailing during cell start-up. A pH of 2.1–2.4 contributes to lower efficiencies but gives a safe operating latitude.

The process results in a metal that contains 99.8 percent Cr, 0.14 percent Fe, 0.01 percent C, 0.001 percent Cu, 0.002 percent Pb, 0.5 percent O_2, 0.004 percent H_2, $\leqq$ 0.01 percent N_2, and no Si, P, Al, or Mn.

The second way of producing electrolytic chromium is from a chromic acid bath operating at an efficiency of about 6.7 percent. Current efficiency is inherently low because six electrons are necessary to reduce the hexavalent ions to chromium metal and is further reduced by hydrogen evolution at the cathode. This metal has a lower oxygen content than that from the chrome–alum process. It contains 0.02 percent oxygen, 0.0025 percent nitrogen, and 0.009 percent hydrogen.

The metal and its alloys. Two principal classes of metallic chromium are available to industry: chromium metal and ferrochromium.

Chief classes of metallic chromium

Pure chromium. Pure chromium metal may be arc cast by consumable or nonconsumable electrode techniques or formed by powder metallurgy methods. The product possesses a certain degree of hot workability.

Chromium surfaces are produced on other metals by electroplating and chromizing. There are two types of electroplating: decorative and "hard." Decorative plate varies between 0.000254 and 0.000508 millimetre (0.00001 and 0.00002 inch) in thickness and is usually deposited over nickel. "Hard" plating is used because of its wear resistance and low coefficient of friction. For these types of plating, solutions of CrO_2 are used.

In one method of chromizing, chromium is condensed on the surface from the vapour and is diffused into the metal by heating. In another method a chromium layer is fused on the surface and diffused in. Electron beam deposition of chromium onto the surface followed by diffusion into the metal has also been used. Salt-bath chromizing using $CrCl_2$ has been tried.

Ferrochromium. Ferrochromium is an alloy added to steel to increase corrosion and oxidation resistance and strength (see above *Steelmaking*).

High-carbon ferrochromium (4–6 percent C, about 68 percent Cr) is used to produce steels in which chromium and carbon must be present. The chromite ore used for ferrochromium production contains a minimum of 48 percent chromic oxide and has, preferably, a ratio of chromium to iron of three to one. To produce high-carbon ferrochromium, ore and coke are fed into the top of an open-top, submerged arc furnace, and the alloy is tapped from the furnace into chill molds. After cooling, it is broken into lumps and graded.

For steels in which carbon is detrimental, low-carbon ferrochromium is produced. A silicon reduction of chromite is accomplished in a two-step process. In the first stage, a high-silicon ferrochromium, substantially carbon-free, is produced in a submerged arc furnace. This product is treated in an open arc furnace with a synthetic slag containing chromic oxide. The alloy produced ranges in carbon from 0.03 percent to 0.25 percent.

A ferrochromium of very low carbon content is made by heating high-carbon ferrochromium with ground quartzite or oxidized ferrochromium in a vacuum. Carbon is removed as carbon monoxide. This ferrochromium has a carbon content of about 0.01 percent.

Other chromium alloys. Up to 3 percent chromium is added to steel to improve the physical properties and increase the heat treatability. Steels containing chromium in combination with other elements, such as molybdenum, nickel, manganese, and vanadium, are used for springs, roller and ball bearings, dies, rails, and high-strength structures. Steels containing 5–6 percent chromium have increased corrosion and oxidation resistance and are used in the form of tubes in the oil industry. Stainless steels, which have a high resistance to oxidation and atmospheric corrosion, contain between 10 percent and 18 percent chromium. The low-carbon varieties (0.1 percent carbon or lower) cannot be heat-treated if the chromium is more than 13 percent. The higher chromium (17–18 percent Cr) steels are used as automobile trim and for equipment to handle nitric acid. Steels with about 13 percent chromium and 0.3 percent carbon are used for cutlery. High-carbon (1–2 percent) steels of this type are used when hardeners and abrasion resistance are desired. Steels containing 25–30 percent chromium have excellent oxidation resistance and are used for products subjected to high temperatures such as furnace parts, burner nozzles, and kiln linings.

Stainless steel

Nickel and manganese can be added to high-chromium steels to form the austenitic types, of which the 18-percent-chromium–8-percent-nickel variety is probably the best known. In addition to their resistance to oxidation and corrosion, austenitic steels maintain their strength at high temperatures better than do the plain chromium steels. Sometimes molybdenum, tungsten, columbium, or titanium are added to improve strength and corrosion resistance or to stabilize the carbides present.

Chromium is also added to cast iron to increase its tensile strength and resistance to wear and heat.

Chromium is added to cobalt alloys in amounts up to 25 percent to obtain corrosion resistance and hardness. Cobalt–chromium–tungsten alloys are used for cutting tools and hard facings.

Nickel base alloys with up to 20 percent chromium and sometimes a little iron are used for high-temperature electric resistance applications.

Chromium is added to niobium base alloys to improve their strength and their oxidation resistance. It is also used in the so-called superalloys for high-temperature applications.

Refractory applications. The use of chromite as a refractory is next in importance to the metallurgical applications of chromium. A typical analysis of a chromite suitable for refractory purposes is 38 percent to 48 percent Cr_2O_3, 12 percent to 24 percent Al_2O_3, 14 percent to 24 percent Fe_2O_3, 14 percent to 18 percent MgO, and less than 10 percent SiO_2. The usefulness of chromite as a refractory is based on its high melting point, moderate thermal expansion, the stability of its crystalline form at elevated temperatures, and its neutral chemical behaviour.

The melting point frequently given for chromite is 2,180° C (3,956° F). Chromite refractories are among the heav-

iest produced; the specific gravity varies between 4 and 4.6. Thermal stability is one of their most valuable properties; there are no polymorphous forms to cause sudden volume changes on heating. Because of the silicates present, chromite exhibits some shrinkage on heating. Cuban ore at 1,500° C (2,700° F) shows a linear shrinkage of 1.5 percent.

Chrome refractories are available in the form of molded bricks and shapes, plastic mixtures consisting of moistened aggregates that are rammed in place, castables composed of dry aggregates and a binder that after mixing with water can be poured like concrete, and mortars and cements for laying brickwork. Crushed raw ore is also sold for patching or filling holes.

Bricks of 100 percent chrome ore have been largely replaced by bricks composed of mixtures of chrome ore and magnesia for greater refractoriness, volume stability, and resistance to spalling. One of the refractories used in the fused-cast condition is composed of 80 percent alumina and 20 percent chromite. This product is highly resistant to the corrosive action of a variety of fluxes, slags, and glasses.

Chromium chemicals. The third major area of chromite applications is the production of chromium chemicals. For this purpose the ore should contain a minimum of 45 percent Cr_2O_3. The iron content may be higher than for the other applications. The production of sodium dichromate is usually the first step in making chemicals. This reagent is produced by heating the ore with soda ash and then leaching out soluble chromate, which is then converted to the dichromate by treatment with sulfuric acid.

Pigments account for about one-third of the primary production of chromium chemicals. Chrome oxide green, which is nearly pure Cr_2O_3, is the most stable green pigment known. It is used for colouring roofing granules, cements, and plasters. It is also employed as a fine powder for polishing. Chromium yellow varies greatly in the shades available and is essentially lead chromate. This pigment makes an excellent paint for both wood and metal. Zinc yellow, a basic zinc chromate, is used as a corrosion-inhibiting primer on aircraft parts fabricated from aluminum or magnesium. Molybdate orange is a combination of lead chromate with molybdenum salts. Chrome green is a mixture of lead chromate with iron blue. This pigment has excellent covering and hiding power and is widely used in paints.

The textile industry accounts for about 10 percent of the primary production of chromium chemicals. They are used in mordanting, in textile-printing compounds, and in aftertreatments. Reagents employed include basic chromic acetate, basic chromic chloride, chromic fluorides, bisulfites, lactates, bromates, oxalates, and thiocyanates.

Chrome tanning

About 25 percent of the chromium chemicals produced go into chrome tanning of leather. This process uses chrome reagents in the form of basic chromic sulfates that, in turn, are produced from the dichromate.

The insolubilizing action of chromium chemicals makes them useful in hardening photographic films and in photoengraving.

More than one-fourth of the production of primary chromium chemicals is used in metal-surface treatments and corrosion control. Such applications include chromium plating, chromizing, anodizing of aluminum, and treatment of zinc and magnesium. Chromium chemicals are used in dips for iron, steel, brass, and tin and also as inhibitors for brines and for recirculating water systems.

These chemicals also find important application in organic oxidations such as the production of synthetic dyes, saccharin, benzoic acid, anthraquinone, camphor, and synthetic fibres. They are used for bleaching, for the purification of chemicals, as analytical reagents, for inorganic and electrochemical oxidations, and in the manufacture of sponge rubber. They also have applications in electric dry cells, slushing compounds, phosphate coatings and catalysts. (F.E.B./Ed.)

COBALT

Ores containing cobalt have been used since antiquity as pigments to impart a blue colour to porcelain and glass. It was not until 1742, however, that a Swedish chemist, Georg Brandt, showed that the blue colour was due to a previously unidentified metal, cobalt.

In 1874 the output of cobalt from European deposits was surpassed by production in New Caledonia; in about 1905, Canadian ores assumed the leadership. Since 1920 the dominant world producer has been the Belgian Congo, now Zaire. Other important producers are the Soviet Union, Zambia, Australia, Canada, Finland, Cuba, and West Germany.

Trends in consumption

Prior to World War I, most of the world's production of cobalt was consumed in the ceramic and glass industries. The cobalt, in the form of cobalt oxide, served as a colouring agent. Since that war, increasing amounts have been used in magnetic and high-temperature alloys and in other metallurgical uses; about 80 percent of the output is now employed in the metallic state.

Mining, refining, and recovery. Because nearly all cobalt is derived from ores of copper, copper–nickel, or nickel, conventional underground or open-pit mining techniques are employed.

In the copper–cobalt ore bodies of Central Africa and the Soviet Union, cobalt occurs as sulfides (carrollite, linnaeite, or siegenite) and as the oxidized minerals heterogenite (hydrated cobalt oxide), asbolite (mixture of manganese and cobalt oxides), and the carbonate sphaerocobaltite. In the copper–nickel–iron sulfide mines of Canada, Australia, the Soviet Union, and other regions, cobalt is present as a substitute for nickel in many minerals.

Cobalt arsenides, such as smaltite, safflorite, and skutterudite, with the sulfoarsenide cobaltite and the arsenate erythrite, are mined in Morocco and on a much smaller scale in many other countries.

Huge nickel-containing deposits found in New Caledonia, Cuba, Celebes, and other countries contain a small quantity of cobalt in the form of oxidized minerals, such as asbolite.

A few pyrite (iron disulfide) deposits mined for their sulfur content contain enough cobalt to warrant the extraction of the latter from the roasted residue. Cobalt sulfides occasionally occur in lead–zinc deposits in quantities sufficient to justify their recovery.

Cobalt ores, like those of most metals, usually require beneficiation before production of a refined product. Such concentrating processes may include hand sorting, separation by gravity or magnetic means, and a variety of techniques based on heating and leaching.

In copper–nickel–iron sulfide ores, cobalt accompanies nickel to the refinery. In an electrolytic nickel refinery, the cobalt is precipitated out of the impure solution surrounding the anode, is purified, and finally electrolyzed. In the purification processes, the impure cobalt solution is treated with sodium hypochlorite, precipitating the cobalt as cobaltic hydroxide and leaving nickel in solution. If nickel is being refined by the carbonyl process (see below *Nickel*), the residues are roasted and then leached, a process in which a liquid is allowed to percolate through the roasted residue. Copper and iron are removed from the solution, and cobalt is separated by sodium hypochlorite. Another refining process for nickel sulfide concentrates consists of leaching with acid or ammonia solutions at high temperature and pressure, followed by successive precipitation of nickel and cobalt from the solution by adding hydrogen.

For copper–cobalt ores, a sulfide concentrate is roasted under controlled conditions to transform most of the cobalt sulfide to sulfate while minimizing the change of copper and iron to their water-soluble states. The product is leached, and the resulting solution is treated to remove copper and iron, and the cobalt is finally recovered by electrolysis. If the copper and cobalt ores are in the oxidized state, copper can be removed by electrolysis in sulfuric acid solution, and the cobalt precipitated from the spent electrolyte by adjustment of the acidity of the solution. Cobalt is again eventually obtained in the metallic state by electrolysis.

Concentrates of cobalt arsenides are smelted with coke and limestone in a blast furnace. After several specialized treatments of the impure mass, cobalt is precipitated by

sodium hypochlorite. Arsenical ores can also be treated by leaching and precipitating the cobalt with hydrogen as with nickel sulfide concentrates.

Recovery from production of sulfuric acid

In several countries, cobalt is recovered from the residues that remain after the roasting of pyrite for the production of sulfuric acid. The roasting operation may be controlled to yield a maximum amount of cobalt sulfate with a minimum quantity of iron sulfate; subsequent treatment gives a hydrated cobalt oxide.

Some cobalt is still obtained from high-grade oxide ores and slags of copper–cobalt smelting operations by reduction with coke and limestone in an electric furnace to produce a copper–cobalt–iron alloy. The latter is dissolved with sulfuric acid, the copper remaining insoluble. Iron is precipitated by sodium chlorate and soda ash, and cobalt finally precipitated as hydroxide by sodium hypochlorite.

Cobalt and its alloys. *Cemented carbides.* In the production of a so-called cemented carbide, such as tungsten carbide, a briquetted mixture of tungsten carbide and cobalt powder is heated at a temperature above the melting point of cobalt. The latter melts and binds the hard carbides, giving them the toughness and resistance to shock needed to make carbides of practical value for machine tools, drill bits, dies, and saws. Cobalt is the most satisfactory matrix metal for this purpose and may be present in amounts from 3 to 25 percent by weight.

Cobalt-60. A radioactive form of cobalt, cobalt-60, prepared by exposing cobalt to the radiations of an atomic pile, is useful in industry and medical science. Cobalt-60 is used in place of X-rays or radium in the inspection of materials to reveal internal structure, flaws, or foreign objects. It is used in cancer therapy and as a radioactive tracer in biology and industry. The advantages of cobalt-60 over radium lie in lower cost, more homogeneous gamma radiation and softer beta radiation, which can be easily filtered out, in the absence of contamination, and in its ability to be machined or shaped in any form before irradiation to fit special requirements.

Magnetic alloys. About 25 percent of the world's cobalt output goes into magnets. Cobalt is one of three elements that exhibits magnetic properties at room temperature. (The others are iron and nickel.) It has the highest Curie point (the temperature at which the magnetic properties disappear) of any metal or alloy. For the past 50 years, the best permanent magnets have always contained a substantial quantity of cobalt.

Cobalt steels, containing 2–40 percent cobalt, are used extensively as magnets. Iron–cobalt–vanadium alloys, such as the so-called Vicalloys, are employed for ductile permanent magnets, as are the Cunicos—*i.e.,* alloys of copper, nickel, and cobalt. Another magnet alloy is iron–cobalt–molybdenum, typified by Remalloy or Comol.

Use in Alnico magnets

Since 1930 it has been known that a series of alloys, the Alnicos, containing 6–12 percent aluminum, 14–30 percent nickel, 5–35 percent cobalt, and the balance largely iron, with small quantities of copper and titanium, make the best permanent magnets. Cobalt is also used in some soft magnetic alloys such as the Perminvar alloys of nickel–iron–cobalt, the Permendur alloys of iron–cobalt, and the cobalt ferrites. Nearly all magnetic alloys are used in electrical equipment and electronic devices.

High-temperature alloys. An important use for cobalt, accounting for about 25 percent of the world's output, is in the field of high-temperature steel alloys. Required in gas turbines, jet engines, and similar applications, such alloys retain their strength above 650° C (1,200° F); these alloys contain from 5 to 65 percent cobalt.

Cutting and wear-resistant alloys. The addition of 2–12 percent cobalt to tool steels enables them to be used more effectively on hard materials for which deep cuts and high speeds are required.

Cobalt improves the wear resistance of many alloys. Hard-facing materials contain 10–65 percent cobalt, and abrasion-resistant die steels usually have 0.4–4 percent cobalt.

Glass-to-metal sealing alloys. Alloys, such as Kovar and Fernico, containing 15–25 percent cobalt, are used extensively in glass-to-metal seals because they have expansion characteristics similar to those of certain glasses.

Dental and surgical alloys. Materials containing 28–68 percent cobalt, with chromium, nickel, and molybdenum, are used in dentistry and bone surgery. They have excellent resistance to tarnish and abrasion, compatibility with mouth tissues and body fluids, high strength and stiffness, good casting properties, and low cost in comparison with precious metals.

Electroplating. For certain applications requiring smooth, bright films that are hard but relatively ductile, an alloy of cobalt–nickel is deposited instead of the conventional nickel-plating. The plating alloy may contain 1–18 percent cobalt, and the electrolyte contains both cobalt and nickel salts.

Other alloys. Cobalt-base alloys having 40–50 percent cobalt are often used as springs. Beryllium copper, used for a multitude of applications besides springs, generally contains 0.1–2 percent cobalt.

An alloy of 54 percent cobalt, 9.5 percent chromium, and 36.5 percent iron changes dimensions very little with changes in temperature. Cobalt-base alloys of 56–63 percent cobalt, with iron and chromium or vanadium, have the advantage in certain applications of very little variation in elasticity regardless of temperature.

Another alloy, a stainless steel containing 18 percent nickel and 8 percent cobalt, also has some industrial applications.

Important cobalt compounds. *Cobalt oxide.* This substance, usually prepared by heating the cobaltic hydroxide that is precipitated from cobalt-containing solutions by sodium hypochlorite, has a number of important uses in the glass and ceramics industries.

Applications in glass industry

Cobalt oxide additions of from 140 to 4,500 grams (five ounces to 10 pounds) per ton of glass are made to impart a blue colour to structural glass, bottles, and optical filter glasses. To neutralize the yellow tint of iron in plate and window glass, small quantities of cobalt oxide, 1–45 grams (0.04–1.6 ounces) per ton of glass, are added. In the proportion of about 454 grams (one pound) per ton of dry clay, cobalt oxide is also employed to neutralize the iron colour in pottery, sanitary ware, and tiles, and in larger quantities to add blue colour. A rich blue is obtained by adding 5 percent cobalt oxide to a glaze of high lead content. Thenard's blue, a turquoise, is characteristic of cobalt aluminate, whereas cobalt silicate gives a unique violet-blue shade. Cobalt oxide in white enamels neutralizes yellow caused by iron; larger amounts give a blue or black colour. In quantities of 0.2–2 percent, used in enamel coats on steel, it promotes adherence of the enamel to the metal.

Cobalt salts. Cobalt salts are usually made by the action of the appropriate acid on cobalt metal or oxide. A number of cobalt salts, particularly organic compounds, are excellent driers of paints, varnishes, and inks. Cobalt linoleates, resinates, oleates, stearates, tallates, and naphthenates, containing 4–12 percent cobalt, are used.

Cobalt, usually in the form of a cobalt–thoria–kieselguhr catalyst, is used in the synthesis of liquid hydrocarbons. Many other cobalt catalysts have been used for a wide variety of chemical reactions.

In many parts of the world, the content of cobalt in the soil and herbage is too low to maintain the health of cattle and sheep. The addition of a small quantity of a cobalt compound to the ration, water, salt lick, or fertilizer has become a well-established practice.

Vitamin B_{12} contains 4.3 percent cobalt; it is the only vitamin known to contain a heavy metal. (R.S.Y.)

GOLD

Gold possesses several outstanding qualities that have made it exceptionally valuable throughout history. It is attractive in colour and brightness, durable to the point of virtual indestructibility, highly malleable, and usually found in nature in a comparatively pure form. Examples of elaborate gold workmanship survive from ancient Egyptian, Minoan, Assyrian, and Etruscan artists, many in nearly perfect condition. The development of a more complex economic system gave gold a new major function as high-denomination currency and later as backing for paper-currency systems, a function it has retained in

many countries to the present. So precious was gold in the Middle Ages that the foundation of chemistry derived largely from the vain experiments of medieval alchemists seeking to convert other metals into gold.

Occurrence. Gold is widely dispersed throughout the Earth's crust. Naturally occurring metallic, or native, gold usually contains variable amounts of silver, copper, platinum, palladium, or certain other elements mixed with it. Gold exists in association with most copper and lead deposits, and although the quantity of gold present is often extremely small, it is readily recovered as a by-product in the refining of those base metals.

Large masses of rock rich enough to be called ores are unusual. Those that do exist are usually quartz lodes or veins. Ores may also be found in deposits which were brought to their sites from veins, such as in river gravels and gold-bearing quartz conglomerate beds (called blankets or reefs) of the Witwatersrand system in the Transvaal and Orange Free State in South Africa. The mineral most commonly associated with gold, other than quartz, is pyrite, FeS_2, or "fool's gold," the yellow disulfide of iron.

The gold mined during great gold strikes in California, Colorado, and Alaska was removed by the placer method (see below *Placer mining*) from alluvial deposits in which metallic gold occurred in the form of gold dust, larger grains, irregularly shaped masses, or occasional nuggets dispersed through the sand or caught in rock crevices.

For the most part, the gold found in rock is invisible to the naked eye, although it does sometimes occur as grains or flakes large enough to be seen, and more rarely as specimen rock. Crystals an inch or more across have been found in alluvial deposits in California. Boulders of pure gold were found in the early days of the Australian gold rush of 1851. Compounds formed with the element tellurium, known as gold tellurides, are found principally in Western Australia and Colorado: the principal ores are calaverite, a bronze-yellow telluride that contains 40 percent gold, and sylvanite, steel gray in colour and containing up to 28 percent gold combined with some silver.

Gold is also found in seawater. Though the total amount in the oceans is estimated at billions of tons, the concentration of it is less than six parts per 1,000,000,000,000 parts of seawater.

Mining. Gold is recovered by placer mining of alluvial deposits, and by lode or vein mining. It is also recovered as a by-product of base-metal mining. Placer mining, the oldest method, entails exploiting the high density (or heaviness) of gold to separate it from the much lighter siliceous material with which it is found. The alluvial deposits mined by placer methods are the gold-bearing sands and gravel that have been deposited by rapidly moving streams and rivers at places where they widen or for some other reason lose speed. As the current slows, the sediment being carried downstream settles to the bottom.

Placer mining. Although the basic principles of placer mining have not altered since early times, the methods have improved considerably, chiefly in mechanical procedures. In the great American gold strikes in California, Colorado, and Alaska, placer mines were almost exclusively the source of gold, and the panning method was the technique utilized by the individual miners. The miner used a pan or a batea (a pan or basin with radial corrugations) in which he placed a few handfuls of dirt and a large amount of water. Swirling the pan washed the siliceous material over the side and left the denser material in the centre of the pan. After many washings, only gold and the other heavy minerals were left. At this point, if the gold particles were large enough, it was comparatively easy to separate them from other materials. If not, further concentration was needed.

Panning

The cradle or rocker was an improvement over the pan and the batea. The cradle, named for a child's cradle, which it resembled, could process larger quantities of ore. Gravel was shovelled onto a perforated iron plate, and water was poured on. The finer material dropped onto the apron, which distributed it across riffles, pieces of wood or iron perpendicularly fixed to the bottom and sides of the cradle. The entire apparatus was rocked, and as the material moved through the rocker, the gold was caught by the riffles. When enough gold was accumulated, the riffles were cleaned.



Figure 28: Miners panning for gold in the Klondike, Yukon Territory, 1897.
The Bettman Archive

Although the cradle or rocker is largely obsolete, its riffles remain in use in sluice boxes and corduroy tables. The name riffle is applied to any strip, bar, or groove placed at right angles to the flowing stream to provide a protective spot where the gold can settle. The corduroy table consists of wide, sloping plates with shallow sides that hold a coarse corduroy cloth. Periodically, the corduroy is removed and washed by hand in boxes partly filled with water to recover the gold-rich concentrates.

Hydraulic mining. In California, thick beds of gravel on the hillsides were worked by hydraulic mining. Powerful jets of water at pressures of hundreds of pounds per square inch were passed through giant swivel-mounted nozzles to break down the gravel banks and wash the material through lines of sluices. Although great volumes of water and many miles of pipes and flumes were required, the cost of treatment was only a few cents per cubic yard, which made it possible to work even poor ground at a profit. The millions of tons of tailings (discarded residue) that were washed into the Yuba and Feather rivers, however, had such an adverse effect on farming downstream that an injunction was obtained against hydraulic miners in 1880, and the work thereafter was strictly limited.

Dredging. In the early 1900s, dredging became the most important type of placer mining and has become probably the single most prevalent technique. The dredge generally used the world over is similar to that employed to deepen harbours and rivers. Invented in New Zealand, the gold-mining dredge achieved its greatest popularity on the rivers there and in California. Dredging is the major technique used today in the U.S.S.R.

Paddock dredging, a later development in the western United States, makes it possible to work placer deposits even if they are not in or near a riverbed. The paddock dredge floats in a pond that is continuously extended by the digging equipment at one end of the dredge while simultaneously being filled in by the waste or tailings at the other end. In this way, the dredge moves across country taking its pond or reservoir with it. By piling more gravel around the reservoir and increasing the water level, the dredge can be made to work its way uphill. In 1910 there were 72 operating dredges in California. After World War II, however, only a few remained in operation.

Lode mining. Many of the methods used in the underground mining and exploration of gold lode or vein deposits are similar to the shaft- and pit-mining methods used for other metals. Tremendous tonnages of gold ore are treated throughout the world since most gold-mine ores contain an extremely low percentage of gold. For example, in one area three tons of ore are processed for every ounce of gold obtained. Ordinarily, a vertical shaft is sunk to allow the miner to get to the lode. Horizontal

working levels then follow the vein. In the Homestake mine in South Dakota (one of the largest producers of gold in the U.S.), the working depth lies 2,070 metres (6,800 feet) below the surface, and since 1877, some 320 horizontal kilometres (200 miles) have been worked.

Recovery and refining. Gold particles wetted by mercury adhere to each other and to copper plates that have been coated with mercury, a process called amalgamation. It forms the basis for a technique for recovering gold from ore. For many centuries this was the method used in treating massive ores. The percentage of gold recovered varied with the type of ores, and considerable gold was lost.

Amalgamation

Cyanide process. The cyanide process, introduced in South Africa in 1890, effected a vast improvement over amalgamation and other earlier methods, and has been extensively used ever since. In this process, the gold in finely ground ore is dissolved by treating it with a very dilute solution of sodium cyanide or the less expensive calcium cyanide plus lime and oxygen from air. The mixture is held for some hours in large tanks equipped with agitators. The chemical reaction yields a water solution of gold cyanide and sodium cyanoaurite. This solution of gold is treated to remove oxygen, then clarifed and mixed with zinc dust to precipitate the gold and the other metals, such as silver and copper, that were dissolved by the cyanide. The precipitate is then treated with dilute sulfuric acid to dissolve residual zinc plus most of the copper. The residue is washed, dried, and melted with fluxes (materials used to promote fusion of the gold and silver and to dissolve the remaining copper). The operation may be repeated to flux off more base metal. The remaining gold and silver alloy, called doré, is then cast into molds for assay.

The cyanide process may constitute the whole recovery process or may follow amalgamation. At some mines in the United States amalgamation is used to recover about 60 percent of the gold, after which cyanidation achieves an overall recovery of 95 to 96 percent. The cyanide process has been used to treat the tailings from early operations, and fortunes have been made in reworking old dumps. The process is employed almost exclusively in recovering gold from the South African mines.

Some South African mines reach depths of at least three kilometres (two miles). In contrast, open-pit mining has opened in northeastern Nevada. The Carlin mine there is one of the largest U.S. producers. Deposits there are sedimentary and originally contained some organic material, which prevented satisfactory cyanide extractions. Weathering, however, destroys the organic material and makes the use of cyanide effective. An electrolytic method also has been devised for oxidizing the deeper unoxidized deposits and destroying the troublesome organic material.

The dominant position of South Africa in world production of gold is a tribute to the ability of modern mining methods to cope with the tremendous rock pressures and high temperatures encountered at great depths.

In addition to the gold from placers and hard rock mines, considerable gold is recovered from copper and copper–nickel ores in the course of electrolytic refining, a process in which a metal is refined by being deposited electrolytically much as in electroplating.

Recovery from scrap. Considerable amounts of gold are recovered from scrap. The composition of scrap varies widely and dictates the process to be used. The scrap is largely produced in the manufacture of gold jewelry and electrical contacts, and as electroplating residues. Melting the scrap under oxidizing conditions will volatilize some of the zinc and remove some of the iron in the slag. The metal may then be formed into pellets, and if the gold content is about 20 percent it can be treated with nitric acid to remove the base metals. If the pellets are low in lead, treatment with hot sulfuric acid may suffice. Retreatment will be required, however, to yield gold of acceptable purity if the sulfuric acid process is used. These methods are appropriate for scrap that is free of platinum metals. Lower grade scrap may be melted and cast into anodes that are electrolyzed in a sulfate solution. The gold will remain in the anode mud and can be recovered by first converting it to gold chloride and then precipitating the gold by adding ferrous sulfate.

Refining. In the Wohlwill process, doré gold from the usual mining and smelting operations is electrolyzed in a chloride solution using direct plus alternating currents. The gold in the doré anode dissolves and is deposited on the cathode as very pure gold. The silver is converted into chloride of which some clings to the anode and must be removed from time to time. This material also contains considerable gold and must be reworked. Any platinum or palladium in the anode will dissolve and is recovered by treating the electrolyte. This extensively used process yields gold at least 99.95 percent pure.

Wohlwill process

Because the Wohlwill process is rather slow, and for other reasons, it has largely been replaced by the Miller process in which chlorine is bubbled through the molten doré, converting the base metals into chlorides, some of which volatilize. The silver also is converted to silver chloride, which is molten and can be poured off and recovered. After this treatment, the remaining gold usually has a purity of 99.5 percent or above.

Assaying. *Fire assay.* Assaying—a process for determining the content of gold in ores, scrap, and refinery products—demands great skill in collecting a proper sample as well as in carrying out all the subsequent steps. A sample of an ore, for example, must be finely ground, then mixed with a flux consisting of soda, borax, silica (if needed), lead oxide (litharge), and flour to reduce the lead oxide to metallic lead during fusion by heat. The reduced lead picks up the precious metals and settles to the bottom. After solidification the lead button is placed in a cupel (small porous cup) made of bone ash and heated in a special furnace provided with a strongly oxidizing atmosphere. The lead oxidizes and the resulting litharge soaks into the cupel.

When the oxidation is complete the appearance of the remaining metallic bead changes and the cupel is removed from the furnace and allowed to cool slowly. The bead, which consists of gold and silver, is removed, cleaned, hammered, and rolled into a thin strip, which is treated with nitric acid to remove the silver, an operation called parting. For proper parting, the silver content of the bead should be about three times that of the gold.

The gold strip, which should remain intact, is washed, heated, and weighed as gold. If platinum metals are present, they will enter the bead and alter its appearance. An examination of the arc spectrum of the bead will give both some clue to its content of platinum metals and guidance in selecting a method for further detailed analysis. If osmium is present initially, it may be lost during oxidation in the cupel and the wet analysis must begin with the lead button. Fire assay is excellent for laboratories that are set up to carry it out on a routine basis, but for occasional use in general laboratories it is quite unsuitable.

Examining arc spectrum

Assay of low-grade ores. A method has been devised for the accurate assay of low-grade gold ores. In it, the ore is ground fine enough to pass through a screen woven of 40 wires per centimetre (100 wires per inch), heated at 550° C (1,020° F) to destroy organic matter, dissolved in a mixture of nitric and hydrochloric acids, then dried to a paste. Slightly diluted hydrochloric acid is then added, followed with a measured amount of 4-methyl-2-pentanone, to extract the gold chloride. After this product is shaken and allowed to separate, small amounts of the ketone, which carry the gold, are fed into an ethyne air burner and the amount of gold determined by atomic absorption measurements.

The metal and its alloys. Apart from its use as a backing for currencies, the largest single use for gold is in the manufacture of jewelry. Use in coins is rare, but industrial applications are important, especially in electronics and related industries. In some countries, considerable amounts are used in dentistry.

Most gold used in jewelry is alloyed with silver, copper, and a little zinc to produce various shades of yellow gold, or with nickel, copper, and zinc to produce white gold. Most of this gold is of 14-karat quality, the karat representing a measure of purity in a scale of 1 to 24; thus 14-karat gold contains $^{14}/_{24}$, or 58.35 percent, gold. A considerable amount of jewelry is gold clad—that is, coated over with a thin cover of gold and is called "gold filled,"

or "rolled gold plate." Varying international designations describe the weight of gold cladding in accordance with government or industry regulations.

Electronics. In the electrical and electronics industries, gold is applied as a finish to electrical connectors, and as a thin coating over palladium in telephone relay contacts. It is also used in the manufacture of transistors and microelectronic assemblies, and to protect etched circuit boards during storage. Large quantities of less expensive jewelry are finished with an electrodeposit of nearly pure gold.

Dental use. Dental applications include wrought and cast gold alloys with a gold content of 60 percent or more along with silver, copper, and sometimes platinum; the rest of the alloy consists of palladium. Some of these dental alloys can be hardened by heat treatment; all possess good strength and a high resistance to corrosion. Gold solders also are used in some industrial fields in which strength plus corrosion resistance are required.

Other uses. Small amounts of gold are used in decorating china and glassware and to provide heat-reflecting surfaces for window glass. Extremely thin coatings on glass with a metal cost of about 54 cents per square metre (five cents per square foot) will reflect most of the Sun's heat rays but will transmit a reasonable amount of visible light. Some gold goes into heat-reflecting shields in aircraft, and small quantities find applications in the medical and chemical fields. Gold leaf is used for decorative purposes.

Gold compounds. Since gold is not attacked by any single acid and does not react with oxygen, sulfur, or the dry halogens (chlorine, bromine, fluorine), it remains bright after outdoor exposure. It is attacked by wet halogens and particularly by a 3:1 mixture of hydrochloric and nitric acids called aqua regia. It is also attacked when it becomes the positive electrode of an electrolytic cell in a hydrochloric acid solution. This fact is the basis for an electrolytic refining process.

The gold chlorine compounds $AuCl_3$, AuCl, and $HAuCl_4$ are involved in the electrolytic refining of gold. When gold is dissolved in aqua regia, $HAuCl_4$ is produced, and gold can be precipitated, after evaporation and resolution, by adding sulfur dioxide, formic acid, or other reductants. If palladium is present in the chloride solution, however, great care is necessary to insure complete precipitation of the gold to avoid the production of the explosive fulminating gold during the later recovery of palladium.

Potassium cyanoaurate $K[Au(CN)_2]$ is the basis for most gold plating baths (the solution employed when gold is electroplated) and is best made by passing current from a gold anode in a potassium cyanide solution to a cathode within a porous pot filled with a potassium hydroxide solution.

Gold forms a host of other inorganic compounds, most of them of little practical importance. Likewise, a large number of organic gold compounds exist. Only two will be mentioned here. The first forms the base of the "liquid bright gold" and is made by reacting gold chloride with a sulfurized terpene. The resulting compound is dissolved in certain organic solvents, mixed with fluxing materials and a trace of rhodium, and then applied to glass or ceramics. After drying and firing, a bright film of gold will remain, well bonded to the glass.

The second, disodium aurothiomalate, is used in the treatment of rheumatoid arthritis. The compound is administered in small intramuscular injections over a period of 24 weeks. Careful supervision of the patient is required because the amount of gold needed for a useful result is close to the point where undesirable side effects result.

(E.M.W./Ed.)

MAGNESIUM

The English chemist Sir Humphry Davy is said to have produced an amalgam of magnesium (an alloy with mercury) in 1808 by reduction of the oxide, using a mercury cathode. But the first free metallic magnesium was produced in 1828 by the French scientist A.-A.-B. Bussy. His work involved the reduction of fused magnesium chloride by metallic potassium.

In 1833 the English scientist Michael Faraday was the first to produce magnesium by the electrolytic reduction of magnesium chloride. Magnesium remained a laboratory curiosity until 1886, when I.G. Farbenindustrie of Germany undertook the manufacture of metallic magnesium on a production basis, using Faraday's electrolytic process. Faraday's electrolytic process

Until 1915 Germany was the sole producer of magnesium, but during World War I this source was shut off from most of the world. Operations to produce magnesium were initiated in the United States by several organizations, most successfully by the Dow Chemical Company of Midland, Michigan. The Dow process, involving the electrolytic reduction of magnesium chloride, originally utilized deep brine wells at Midland as the raw material source, but these have been superseded by the use of seawater.

Raw material sources of magnesium. A number of extractive methods are used in the recovery of metallic magnesium from its raw sources. These methods vary considerably depending upon the type of primary raw mineral, which may be found either in mineral deposits or in aqueous solutions.

Among the raw ore deposits, the most common minerals are the carbonates dolomite (a compound of magnesium and calcium carbonates, $MgCO_3 \cdot CaCO_3$) and magnesite, or magnesium carbonate ($MgCO_3$). The oxide mineral brucite, magnesium oxide associated with water ($MgO \cdot H_2O$), is less common, while the chloride mineral carnallite, a compound of magnesium and potassium chlorides and water ($MgCl_2 \cdot KCl \cdot 6H_2O$), is rarer still. Magnesium is recoverable from aqueous solutions of brine wells, but by far the largest source of magnesium is the oceans of the world, where it appears in its chloride form. While seawater is only approximately 0.13 percent magnesium, this proportion remains quite constant and represents an almost inexhaustible source of supply—approximately 5,500,000,000 kilograms of magnesium per cubic kilometre of seawater. Magnesium in seawater

The uniformity of magnesium content in seawater facilitates standardization in process recovery and offers certain economic advantages, which can vary with geographic location of the producer.

Refining and recovery. Magnesium is produced by two distinctly different methods. One is the direct reduction of the magnesium ores by strong reducing agents such as carbon, silicon, etc. The other is by electrolytic reduction of magnesium chloride. In the past, magnesium produced by direct reduction was of higher purity than that derived by electrolytic reduction, but because of process refinements, this is no longer the case. Where power costs are low, electrolytic reduction is the cheaper method; and, indeed, it accounts for much the greater part of world production.

Electrolytic reduction processes. Both the early German process and the Dow process use magnesium chloride as the basic source, differing primarily in the characteristics of the electrolytic cell employed. Magnesium chloride as raw material

The flow chart of the Dow process is shown in Figure 29 and represents the method used by the Dow plant near Freeport, Texas, where abundant natural gas is available for cheap power; nearby Galveston Bay provides an inexhaustible supply of lime in the form of oyster shells. In this process, seawater is mixed with calcium hydroxide in a processing unit called a flocculator. The calcium hydroxide is obtained by processing oyster shells. In the flocculator the calcium from the calcium hydroxide is exchanged with the magnesium from the magnesium chloride in the seawater, forming the insoluble precipitate magnesium hydroxide. The seawater containing the magnesium hydroxide precipitate flows from the flocculator to large settling tanks. After the magnesium hydroxide has settled on the bottom, it is pumped out as a slurry, filtered, and converted to magnesium chloride by reaction with hydrochloric acid. After drying, the magnesium chloride is fed to electrolytic cells where it is broken down into magnesium metal and chlorine gas.

Direct reduction processes. Among the several direct reduction processes for the production of magnesium, by far the most important is the ferrosilicon process (sometimes referred to as the Pidgeon process for its originator, L.M. Pidgeon of Canada).

In the ferrosilicon process, dolomite is calcined (convert-

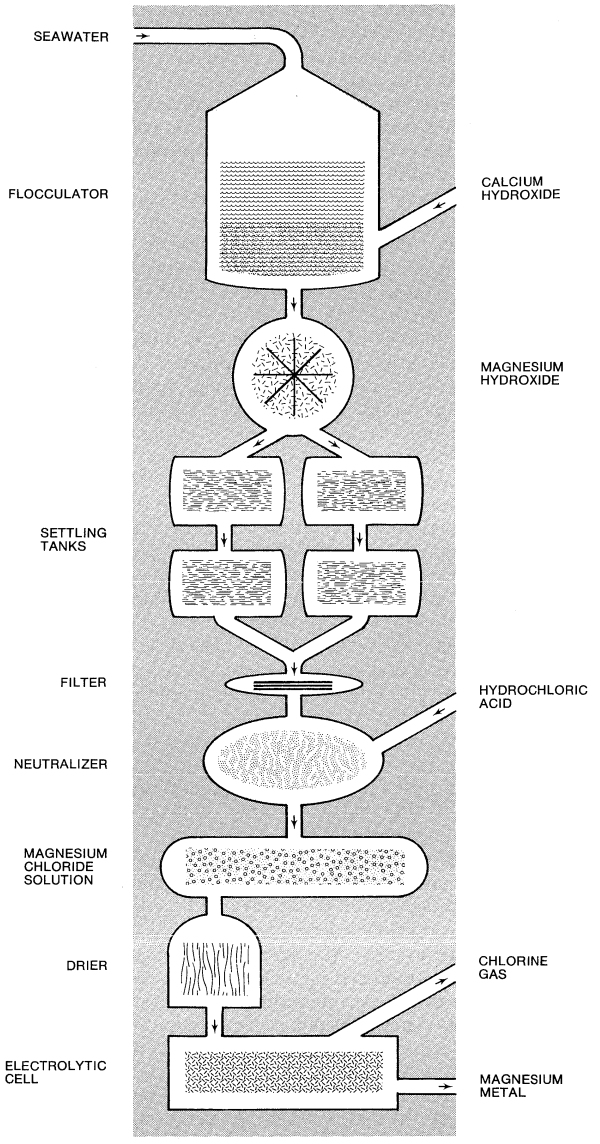


Figure 29: Stages in the Dow process for obtaining magnesium from seawater (see text).
Drawing by D. Meighan

ed to lime by heating) and mixed with finely pulverized ferrosilicon Si–Fe. This mixture is briquetted and then placed in retorts at approximately 1,200° C (2,192° F). The direct reduction takes place within these retorts under very carefully controlled process conditions, the pressure and temperature being especially critical. A flow chart for the ferrosilicon direct-reduction process is shown in Figure 30.

The reaction taking place in the retorts is as follows:

$$2\ MgO \cdot CaO + Si{-}Fe \rightarrow 2\ Mg + (CaO)_2\ SiO_2\ (Fe).$$

The dolomite combines with the ferrosilicon to yield metallic magnesium and a combination of calcium and silicon oxides and iron.

Several hours are required for the reduction cycle. The magnesium yield may run from 13 to 23 kilograms (30 to 50 pounds), depending upon retort size. For the process to compete economically with electrolytic reduction methods, plants must be located near sources of raw materials and power.

The metal and its alloys. Primary commercial magnesium metal is approximately 99.8 percent pure when produced by either of the methods. It is cast into small ingots or pigs, usually weighing from nine to 23 kilograms (20 to 50 pounds). Higher purity can be attained by further processing in which the metal is remelted in a vacuum.

Many nonstructural uses of magnesium utilize the unalloyed metal, while specific alloy compositions are required for structural uses.

Nonstructural applications. The use of magnesium as an alloying element in aluminum-base alloys constitutes its greatest metallurgical application. In these aluminum alloys, magnesium is present in amounts ranging from less than 1 percent to approximately 10 percent. The use of the magnesium enhances the mechanical properties, as well as the corrosion resistance of such alloys. Similarly, pure aluminum is used as an alloying element in many magnesium-base alloys and has a similar beneficial effect upon structural properties.

Magnesium is frequently used to protect steel against corrosion by a method known as cathodic protection. When magnesium comes into electrical contact with steel in the presence of water, the magnesium corrodes sacrificially, leaving the steel intact. Ship hulls, water heaters, storage tanks, bridge structures, pipelines, and a variety of other steel products are protected in this manner.

Magnesium has been used in military pyrotechnics for many years and has found numerous uses in incendiary devices, flares, etc. Magnesium, in the form of finely divided particles, has been used as a fuel component, particularly in solid rocket propellants.

Magnesium metal is used as an alloying element in zinc die castings to promote dimensional stability and to improve mechanical properties. It is also used as an oxygen scavenger and desulfurizing agent in the manufacture of copper and nickel alloys and as a reducing agent in the manufacture of titanium and zirconium.

The use of magnesium in making nodular cast iron is well established and is based upon the nodular form imparted to the graphite present. This, in turn, gives cast iron vastly improved mechanical properties.

The magnesium dry cell

The magnesium dry cell has achieved some degree of prominence, particularly for specialized applications in which long storage life and light weight are desirable. Such dry cells usually consist of magnesium plus either silver

Drawing by D. Meighan

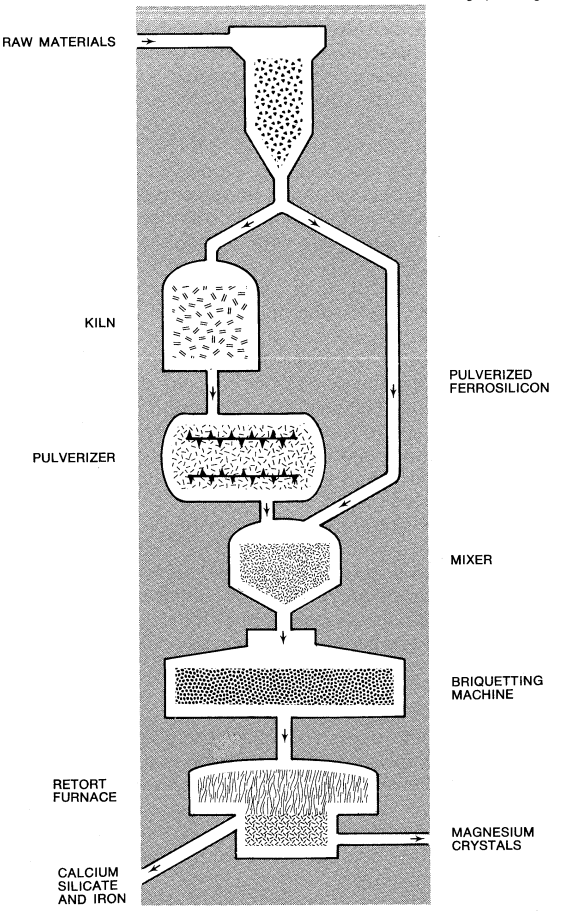


Figure 30: Stages in the ferrosilicon magnesium process.

chloride or cuprous chloride. When activated with water, these dry cells rapidly build up voltages of 1.3 to 1.8 volts per cell. Silver chloride–magnesium cells operate with constant potential in the range of 94° to −54° C (200° to −65° F). Cuprous chloride–magnesium cells exhibit similar output characteristics but with somewhat lower power output per unit of weight. Both cells use synthetic porous fibers for separation of the electrodes. The shelf life of both is almost indefinite, and the primary usage is in military applications.

The use of magnesium in photoengraving plates, a relatively new use of the metal, is based primarily upon its ease of etching by certain reagents.

Structural applications. Magnesium is useful as a structural metal when it is alloyed with small amounts of other metals. These alloying metals are aluminum, zinc, manganese, silver, zirconium, lithium, thorium, and the rare earth metals. In most cases, the alloying elements form intermetallic compounds that permit heat treatment for enhanced mechanical properties as well as other benefits, such as corrosion resistance, size stabilization, and improved fabricability.

Structural forms of magnesium include sand, permanent mold, and die castings as well as extrusions, forgings, and rolled sheet and plate.

Magnesium alloys are fabricated by most of the methods used for other metals and show exceptionally good machining characteristics. While magnesium alloys do require careful processing in machining, fire hazard is minimal if the metal is properly handled. Accumulation of large quantities of magnesium chips or fines is usually the only circumstance in which magnesium will burn in an industrial environment.

Forming at elevated temperatures

The forming of magnesium into structural shapes follows the pattern used for most common metals, such as aluminum and steel, except that it is usually done at elevated temperatures. These temperatures may vary from 160° to 400° C (320° to 750° F), depending upon the specific alloy and the degree of deformation. The need for elevated temperatures is ascribed to the hexagonal close-packed crystal structure of magnesium as compared with the cubic structure of the other common metals.

The welding of magnesium alloys is accomplished readily by either resistance welding or inert-gas-shielded arc welding. Normally the type of equipment used for the resistance welding of aluminum may be used for magnesium, whether it be spot welding or seam welding. For arc welding of magnesium, an inert gas shield (helium, argon, or a mixture) around the weld area prevents oxidation and contamination from the surrounding atmosphere. Depending upon the type of weld, the arc is produced by either a tungsten electrode or a magnesium wire. The latter serves as both electrode and filler rod.

Magnesium structures also are formed by the use of rivets and other mechanical fasteners, such as screws, bolts, inserts, and other special devices. When using these types of attachments, the danger of galvanic (electrochemical) corrosion is very real, particularly since the fasteners are not made of magnesium.

Because of the ease of fabrication, together with the attractive mechanical and physical properties, magnesium alloys have been widely accepted in structures where light weight and high strength are needed. The aerospace industry consumes large amounts of magnesium alloys in aircraft, rockets, and missiles. These applications include all types of castings, extrusions, rolled products, and forgings. Magnesium castings have found especially wide acceptance since high-integrity castings with excellent mechanical properties have been developed through improved foundry techniques. Gearbox housings, aerodynamic surfaces, and hardware chassis are typical applications. The environment of space makes possible the use of magnesium alloys that would not be suitable for similar applications in an earth environment. This is because the low-pressure, dry atmosphere does not sustain galvanic corrosion. Orbiting space satellites probably contain more magnesium than any other metal.

Magnesium alloys in outer space

In the late 1950s special compositions of magnesium alloys were developed for use in high temperature applications. These alloys contain thorium, silver, rare earth metals, or zirconium or combinations of all these in various amounts, enabling magnesium to be used at temperatures up to approximately 320° C (610° F), as compared with 150° C (300° F) for conventional alloys.

One of the newer magnesium alloys, containing 14 percent lithium and 1 percent aluminum, has excellent damping (vibration-suppression) characteristics. It has found numerous uses in structures where resonant vibration problems existed.

The varied commercial uses of magnesium structural alloys include luggage frames, hand trucks, ladders, business machines, camera bodies, power tools, tooling fixtures, and race-car wheels. Magnesium has found substantial use in tools used by masons and plasterers because of its light weight and resistance to alkalies. Magnesium die castings are extensively used, notably in the manufacture of automobiles.

Magnesium compounds. The compounds of magnesium form an important group of chemicals that are widely used in industry, medicine, and agriculture. Some of these chemicals are formed as by-products of primary magnesium metal production while others are synthesized by specific chemical processes.

Epsom salts

The best known medical compounds are milk of magnesia, $Mg(OH)_2$, and the hydrous magnesium sulfate popularly known as Epsom salts, $MgSO_4 \cdot 7H_2O$.

Magnesium has had a traditional place in the field of organic chemistry since the discovery of Grignard reagents in 1900. This group of reagents is represented by RMgX, in which R is an organic radical and X is one of the halides: chlorine, bromine, or iodine. Organomagnesium compounds are important in many organic chemical syntheses such as conversion of aldehydes and ketones to alcohols.

Magnesium carbonate, $MgCO_3$, is one of the most widely used magnesium compounds in industry. It is used in large quantities for refractory (high-temperature) bricks and linings of open-hearth furnaces used in steel manufacture. In this application, magnesite and dolomite ores are the principal sources of supply.

While basic magnesium carbonate serves primarily as a component of insulating compounds, the technical (high-purity) grade is widely used for fireproofing and as a filler for paper, plastics, paints, and varnishes. The principal pharmaceutical uses are as a bulking compound in powder formulations and as an antacid.

The most important use of magnesium chloride is in the recovery of magnesium metal, but it is also used in wallplaster compositions and as a fireproofing agent for wood.

In addition to its medical applications, magnesium sulfate is used as a tanning agent, a conditioning agent for cotton, a mordant for dyeing wool, and as a component of fertilizers. (J.H.R./Ed.)

MERCURY

Chief mineral source

Mercury, also known as quicksilver, and its principal ore mineral, cinnabar, have been known and used for more than 2,300 years. In early times, mercury was used in religious ceremonies, in medicinal ointments and cosmetic preparations, and in the recovery of gold and silver by the process of amalgamation. Mercury readily forms alloys (amalgams) with gold, silver, and many other metals. Since the 17th century the importance of mercury has steadily increased with advancing science and industry.

Ores and mining. Of the more than 25 known minerals that contain mercury, the principal ore mineral is cinnabar—a soft red to reddish-brown mercury sulfide. Mercury is found in the elemental state in some deposits. Cinnabar has been found in all types and ages of rock, usually alone, but sometimes in association with antimony, gold, iron, and zinc.

Air samples taken in the vicinity of sulfide and precious metal ores have shown a mercury content several times greater than that found in samples taken over nonmineralized areas, and such air sampling for mercury content promises to become an important tool in exploration for mercury and associated minerals.

Mercury deposits are small and irregular, usually veinlets

but sometimes small disseminated deposits, a situation that precludes large-scale, highly mechanized mining methods.

Most of the ore comes from underground mines in which the ore is broken by conventional drilling and blasting and is removed by scrapers or is mechanically loaded into cars. Surface mining is also carried on.

Despite the widespread occurrence of mercury, only a few commercial-size deposits have been found. The largest reserves are reportedly in Spain at the El Entredicho Mine, located about five kilometres (three miles) southwest of the Almadén Mine, which first began production about 400 BC. The Idria Mine in Yugoslavia, where mercury was first discovered around 1470, continues to be a major producer, as does the Monte Amiata district in Italy, where production began about 1868. The Santa Barbara Mine in Peru, which began operations in the 16th century, also continues to be a significant producer. The New Almadén Mine in California helped make the United States the world's largest producer for a few years following 1875, though today one-quarter to one-half of U.S. needs are supplied by imports from Spain, Algeria, and Japan. Mercury is also produced in China, Czechoslovakia, Turkey, and the Soviet Union.

Secondary mercury—that is, mercury that has been recovered after primary uses—is a significant source in most developed countries.

Recovery and refining. The recovery of the metal from the ore is one of the more simple of the metallurgical extraction processes. Mercury is separated by roasting the ore as mined or as a concentrate at temperatures near or above 580° C (1,080° F), the temperature near which mercury sulfide sublimes (changes from solid to vapour directly). The mercury sulfide vapours react with oxygen in the air to form mercury and sulfur dioxide and are then passed through a dust collector and into a series of condenser pipes.

Because concentrates and some ores contain relatively large amounts of sulfur, various substances are added to assist in oxidizing the sulfur and to prevent the formation of a synthetic cinnabar or free sulfur in the condenser pipes. Brine, iron filings, and scrap iron are some of the substances used.

Cinnabar is roasted in a variety of furnaces. The most common is a retort or tube, ranging in size from 25 to 50 centimetres (10 to 20 inches) in diameter and 1.8 to 3.6 metres (six to 12 feet) in length, supported by some type of masonry construction with a firebox below or beside the retort. A number of other types of furnaces are used to provide greater capacity and continuous operation.

The upper end of the furnace is connected to a condenser system. The typical condenser system consists of a series of vertical cast-iron pipes connected at the top with cast-iron or glass fibre return bends and at the bottom with cast-iron hoppers. The gaseous mercury condenses and collects in the hoppers under a water seal.

Mercury recovered in this manner, pure enough for most commercial uses, is of prime virgin grade averaging not less than 99.9 percent bright and clean mercury.

Mercury is packaged in cast-iron, wrought-iron, or spun steel bottles or flasks 10 to 18 centimetres (four to seven inches) in diameter and about 30 centimetres (12 inches) high. The net weight of one flask of mercury is 34.5 kilograms (76 pounds), the commercial unit of world trade.

Most mercury is recovered by furnace treatment of ore that has come directly from the mine without processing other than crushing. Ore from some mines may be low grade or may contain impurities that create problems in roasting. In such cases the mercury content may be upgraded by the ore flotation process, which takes advantage of mercury's nonwetting characteristic to separate it from other substances. The concentrates are then treated by the roasting process described above.

Although it is not done in current mining operations, mercury can be recovered by leaching with a solution of sodium hydroxide and sodium sulfide. Metallic mercury is precipitated from solution by adding aluminum.

Generally neither flotation concentration prior to furnacing nor leaching of mercury can compete economically with direct furnacing of the ore.

Laboratory tests have shown that electrochemical oxidation in a brine slurry is a promising method of recovering mercury from low-grade ore.

The metal and its alloys and amalgams. Perhaps the best known use for mercury is the "mercury" thermometer, which is based on mercury's property of constant volume expansion over its entire range in the liquid state. Mercury is also used in many measurement and control instruments, including barometers, manometers, flow meters, and pressure gauges. These instruments utilize mercury's high surface tension, high density, and low vapour pressure.

Electrical applications

But by far the greatest usage of mercury is in electrical applications. The largest single application is in the production of chlorine and caustic soda by the electrolysis of brine in a mercury cell. This process makes use of mercury's electrical conductivity, as well as its ability to amalgamate with other metals. As a moving cathode, mercury amalgamates with the liberated sodium, which immediately reacts with water to form metallic mercury and sodium hydroxide. Chlorine is recovered at the anode. Investigations have shown the feasibility of using mercury cathode cells in the recovery of uranium and vanadium by electrolysis. Mercury is used in an electrolytic process to remove impurities from metals.

The next largest use is the mercury dry-cell battery. These batteries operate under conditional high temperature and humidity, have long life, deliver the same ampere-hours of service at the rated current range whether operated continuously or intermittently, and have large power output per unit size. The mercury battery is used in hearing aids, guided missiles, communications equipment, and other applications in which a constant output of electric current is required.

Mercury-vapour arc lamps are compact and have high wattage and high lumens-per-watt output. They are used particularly for industrial floodlighting and street lighting. Other uses for mercury lamps are in motion-picture projection, photography, dental examination, and heat therapy. Electrical applications also include rectifier bulbs, oscillators, power-control switches, and home-type wall switches.

In precision casting, patterns have been made of frozen mercury, a technique used particularly for the casting of complex parts. A possible space application is a mercury-filled flywheel for spacecraft stabilization.

Amalgams

Mercury amalgamates readily with many metals. Gold and silver are recovered by amalgamation. Amalgams of sodium and potassium are used as reducing agents. A mercury–thallium amalgam with a melting point of −60° C (−76° F) is used as a low-temperature thermometer. Mercury–silver amalgams are used in dentistry for repairing cavities.

Although prime virgin grade mercury is satisfactory for most uses, it may have to be purified to meet the special requirements of dental and medical instrument and battery applications. Mercury is purified by redistilling or treatment with nitric acid.

Mercury vapour boilers have been tried for power generation but were found uneconomical.

Chemical compounds. Mercury compounds are used in the chemical industry, in pharmacology, in the paint industry, and in agriculture. Mercury sulfate is used as a catalyst in the production of glacial acetic acid, methyl styrene, and vinyl chloride. Bichloride of mercury, mercurochrome and ointments of metallic mercury, yellow mercuric oxide, and ammoniated mercuric chloride are used as skin antiseptics. Mercurous chloride (calomel) has been used as a diuretic and a cathartic.

Phenylmercury compounds, predominately the acetate, are used for treating and storing seeds, for spraying fruit trees, for weed control, and as textile preservatives. Though mercury compounds were once widely used for slime control in the production of pulp and paper, this use has declined considerably in recent years because of the potential health hazard in paper used for food products.

A growing use for mercury is in mildew-proofing paints. The use of mercury for antifouling paints—*i.e.,* paints resistant to marine growth—has declined in recent years. Mercury as a solid solution of mercury sulfide and cad-

mium sulfide is used in producing heat- and light-fast red pigments.

Mercury fulminate, resulting from a reaction of alcohol and mercuric nitrate, explodes on impact and is used in percussion caps as a detonator for other explosives.

(J.E.S./Ed.)

NICKEL

Nickel was used industrially as an alloying metal almost 2,000 years before it was isolated and recognized as a new element. Although it is best known for its use in coinage, nickel has become much more important for its many industrial uses, particularly as a constituent in alloy steels.

As early as 200 BC, the Chinese made substantial amounts of a white alloy from a copper–nickel ore found in Yunnan province and zinc. The alloy, known as *pai-t'ung,* was exported to the Middle East and even to Europe.

Later, miners in Saxony encountered what appeared to be a copper ore but found that processing it yielded only a useless slaglike material. They considered it bewitched and ascribed it to "Old Nick." Thus, it became known as kupfernickel (Old Nick's copper). It was from this ore, studied by A.F. Cronstedt, that nickel was isolated and recognized as a new element in 1751. In 1776 it was established that *pai-t'ung,* now called nickel-silver, was composed of copper, nickel, and zinc.

Demand for nickel-silver was stimulated in England around 1844 by the development of silver electroplating, for which it was found to be the most desirable base. The use of pure nickel as a corrosion-resistant electroplated coating developed a little later; both these uses continue, with nickel plating much the more important.

Ores and mining. Small amounts of nickel were produced in Germany in the mid-19th century. More substantial amounts came from Norway, and a little from a mine at Gap, Pennsylvania. A new source, New Caledonia in the South Pacific, came into production about 1877 and dominated until the development of the copper–nickel ores of the Copper Cliff-Sudbury, Ontario, region in Canada, which since 1905 has been the world's largest source of nickel. Several producers are active in Canada.

Canadian ores are sulfides containing nickel, copper, and iron. The most important nickel mineral is pentlandite $(Ni,Fe)_9S_8$, followed by pyrrhotite, usually ranging from FeS to Fe_7S_8, in which some of the iron may be replaced by nickel. Chalcopyrite, $CuFeS_2$, is the dominant copper mineral in these ores, with small amounts of another copper mineral, cubanite, $CuFe_2S_3$. Some gold, silver, and the six platinum-group metals also are present, and their recovery is important. Cobalt, selenium, tellurium, and sulfur also may be recovered from the ores.

Other important classes of ore are the laterites, which are the result of long weathering of peridotite initially containing a small percentage of nickel. Weathering in subtropical climates removes a major portion of the host rock, but the contained nickel dissolves and percolates downward and may reach a concentration sufficiently high to make mining economical. The nickel magnesium silicate, garnierite, is the richest in nickel. The New Caledonian deposits are of this type, and numerous other laterite deposits are scattered around the world, presenting a wide range of mining, transport, and recovery problems. The nickel content of laterites varies widely: at Le Nickel in New Caledonia, for example, the ore delivered to the smelter in 1900 contained 9 percent nickel; currently it contains 1 to 3 percent, but production from this source is increasing substantially.

Laterite deposits

The search for new ore bodies is an essential part of a mining enterprise. Though prospectors using visual surface indications found the earliest ore bodies, many sulfides, particularly in Canada, are covered with clay and give no surface clues. Geophysical methods are thus required, and magnetic and electrical methods have been employed. Nickel ore bodies may contain magnetic components that are reasonably good conductors of electricity. Much of the surface, however, is wooded, rough, or boggy, so efforts have been made to devise airborne prospecting devices. An air magnetometer developed during World War II as an antisubmarine device was thus employed after the war by International Nickel Company. A second method, developed by the same company, is the airborne conductor sensor or electromagnetic device that detects changes in the conductivity of the Earth by inductive methods. It opened a new era in prospecting, not only for the nickel–copper sulfides but also for other sulfide ores and is now used throughout the world.

Refining and recovery. The Sudbury sulfide deposits were originally identified as copper deposits, and the Canadian Copper Company was organized in 1885 to develop the mines. But the new ore, which contained 2 percent of nickel, along with copper and iron, baffled copper refiners. The owner of a copper smelter at Bayonne, New Jersey, who contracted to buy 100,000 tons of the new ore, found that it could not be handled by normal copper smelting. After extensive experimentation, it was discovered that by adding sodium sulfide and allowing the molten mixture to cool slowly, two layers, which could be broken apart, formed. The top layer contained most of the sodium sulfide and copper sulfide; the bottom was largely nickel sulfide. If the bottom was again melted with more sodium sulfide and slowly cooled, the separation was improved. After several repetitions, a nickel sulfide of good purity resulted. This could be roasted to remove the sulfur, then reduced to yield nickel of adequate purity for use. The process, known as "Orford tops and bottoms," was used until 1948, when it was supplanted by a scheme first proposed in 1932, when it was determined that the low-temperature mutual solubility of copper sulfide (Cu_2S) and nickel sub-sulfide (Ni_3S_2), which constituted the mixture of sulfides (the matte), was very small and could be selectively floated as mechanical mixtures of the two sulfides. Though first efforts in selectively floating solidified matte containing these two components were not successful, good separation was achieved after much finer grinding of the solidified matte was employed.

In the present practice, the ore is crushed and ground with water to liberate the sulfides from the less dense gangue. Density differences as well as magnetic properties are utilized in this step. The sulfide minerals can then be separated from the other minerals by selective flotation, in which chosen minerals attach themselves to air bubbles and form a froth. This process has been refined to achieve good separation of the copper-containing chalcopyrite from the nickel-containing fractions. The nickel-containing fraction of pyrrhotite-pentlandite from the flotation process, plus material recovered by magnetic concentration, is roasted to remove a portion of the sulfur, and the material is then smelted to remove much of the iron and silica. The resulting matte consisting of nickel, copper, and iron sulfides is further oxidized in a Bessemer converter to yield essential copper and nickel sulfides. These are separated by slowly cooling the matte, grinding it, and selectively floating it to yield essentially nickel sulfide and copper sulfide. In subsequent operations, the two sulfides are reduced to metal and converted into anodes for electrolytic refining to yield high-purity nickel and copper. Special methods have been devised to recover nickel from pyrrhotite and to produce a marketable iron ore.

Selective flotation

At another Canadian mine, impure nickel sulfide is cast into anodes that are heat-treated to render them less brittle and are then decomposed electrolytically in a sulfate solution, a process that yields nickel cathodes of good purity. The sulfur, precious metals, and copper are also recovered.

A unique method for recovering high-purity nickel from matte is employed at a nickel plant at Clydach, Wales, where the impure sulfide is roasted to remove all of the sulfur and then processed to form finely divided metal, which is treated with carbon monoxide gas at about 55° C (131° F) that reacts to produce volatile nickel carbonyl, $Ni(CO)_4$. This in turn is decomposed at about 230° C (446° F) to yield pure nickel in the form of small pellets.

Another version of the carbonyl nickel process is used at Clydach to produce very fine nickel powder for many powder-metallurgy applications. In this process, the reduced nickel reacts with carbon monoxide at about 120° C (248° F) at a pressure of 21 kilograms per square centimetre (300 pounds per square inch). The resulting $Ni(CO)_4$ is condensed, fractionated to remove iron carbonyl, and de-

composed at high temperature to yield high-purity nickel powder of sizes and shapes required for various end uses, including the nickel–cadmium storage battery.

Refining laterite ores is difficult because of the impossibility of concentrating the nickel fraction by gravity or flotation. In processing laterites by pyrometallurgy, all the diluent material must be removed by means of a flux, a process requiring large amounts of fuel or electrical energy. In some cases, sulfur is added in the form of gypsum to produce a matte that can be further processed in the manner employed for sulfide ores. The copper content usually is low, however, so processing is somewhat simpler. When sulfur is not added, the nickel plus considerable iron may be reduced with carbon or silicon to yield a ferronickel alloy.

An alternative to pyrometallurgy is chemical extraction of the nickel. With one type of laterite, the extraction of nickel, plus some iron and cobalt, can be accomplished by treatment with a sulfuric solution under high temperature and pressure. Alternatively, extraction can be effected by treatment with an aqueous solution containing ammonia or ammonia plus carbon dioxide.

Chemical extraction

It is difficult to generalize about the worldwide methods used to refine nickel since the methods depend upon the ore, the cost of electric power, and the market for the various end products that are possible. In the U.S.S.R., most of the production comes from the north, Norilsk in western Siberia and the Pechenga area of the Kola Peninsula. The ores involved are the copper–nickel–iron–sulfide type similar to those of Canada. The methods employed for processing them are similar to those used in Canada, including selective flotation of the matte. The remainder of Russian production comes from the southern Urals. The various oxide ores of this area are smelted with gypsum to provide enough sulfur to produce a matte, the same method long used in New Caledonia.

Nickel and its alloys. Pure nickel possesses a useful combination of properties including corrosion resistance, good strength, and high ductility, even at extremely low temperatures. It also possesses useful electronic properties and special magnetic properties. Nickel is a particularly good catalyst for reactions involving hydrogen, and substantial quantities of it are used in hydrogenating natural oils. Nickel catalyzes the addition of hydrogen to unsaturated compounds in vegetable, animal, and fish oils, converting them from liquids to solids. Natural oils treated in this way are used in such products as shortening, oleomargarine, and soap.

The white colour of nickel is attractive, and most of its alloys with copper are substantially white. Its ability to form strong, ductile alloys with many metals, including iron, chromium, cobalt, copper, and gold, is utilized in industry.

Nickel is resistant to corrosion by fluorine, alkalies, and a variety of organic materials. It remains bright on indoor exposure but tarnishes outdoors, although its corrosion rate is very low. Its low corrosion rate, coupled with its resistance to corrosion by sodium chloride and other chlorides used on roads during the winter, makes it essential as an undercoat on chromium-plated automotive trim. About one-half of the nickel employed for electroplating in the United States is used in this way. Nickel containing some phosphorus can be deposited by chemical reduction; this "electroless nickel" is finding new uses.

Heavy nickel plating is employed as a lining for tank cars and as a coating for the inner walls of large pipes and similar equipment in the chemical industry.

Nickel is essential as the base for oxide-coated cathodes used in all television tubes and all but the largest radio power tubes. Alloyed with about 2 percent tungsten plus a trace of magnesium, nickel is used as the cathode base in amplifiers for submarine cables that are expected to function for 20 years without attention.

Nickel also is an essential component of white-gold alloys widely used for jewelry. These alloys also contain nickel, copper, and zinc, all of high purity.

The addition of copper to nickel provides a series of useful alloys. Monel metal, 67 percent nickel and the balance essentially copper, is stronger than nickel and has broad corrosion-resisting applications. Extremely resistant to rapidly flowing seawater, it has many marine uses. The addition of a small percentage of aluminum and titanium renders it precipitation hardenable; this high-strength version is widely used for propeller shafts. Increasing copper to 55 percent produces an electrical resistance alloy known as Constantan, which is used as a thermocouple in conjunction with pure copper.

Monel

The 30 percent and 10 percent nickel–copper alloys, usually containing 0.5 percent and 1.5 percent iron, are widely used in the form of tubes for heat interchangers and condensers. Their resistance to seawater corrosion makes them important in desalination plants. Copper-base alloys containing a small percentage of nickel become precipitation hardenable if 5–8 percent of tin or a smaller amount of silicon or phosphorus is added. These have special uses.

The ancient Chinese alloy *pai-t'ung,* now known as nickel-silver, contains 10–30 percent nickel with the balance copper plus zinc. This alloy continues as a favoured base for silver-plated ware. It is also used as a spring material for relays and has numerous other applications.

An alloy of 25 percent nickel and 75 percent copper is essentially white in colour and was adopted for coinage by Belgium in 1860 and by the U.S. five years later. More recently it has been employed in the U.S. as the outer layer of copper-centred 10- and 25-cent coins. Pure nickel was adopted by the Swiss for coinage in 1881; this use has spread to many other countries.

The fact that nickel changes in length as it is magnetized makes it useful as an ultrasonic transducer in various underwater defense devices. Alloying nickel with about 21 percent of iron has a spectacular effect in producing alloys with extraordinarily high magnetic permeability in weak fields. This type of alloy, known as Permalloy, discovered at Bell Telephone Laboratories in 1916, has had a great value in long-distance telephone transmission, including undersea cables. Other alloys of about 45–50 percent nickel, and the balance iron, have been developed for magnetic uses at higher field strengths.

Invar, an alloy containing 36 percent nickel, with the balance iron, is notable for its extremely small thermal expansion. Discovered in 1898, it has, along with later developed nickel alloys, many applications ranging from thermostats to balance wheels for watches, metal-to-glass seals essential to electric lights, and radio tubes.

A remarkable group of nickel-containing permanent-magnet alloys was developed, beginning in Japan in the early 1930s. An early example contained 25 percent nickel, 12 percent aluminum, and the balance iron. More powerful versions, such as Alnico V (containing 8 percent aluminum, 14 percent nickel, 24 percent cobalt, 3 percent copper, balance iron), developed in The Netherlands, were heat-treated in a magnetic field. These materials had a profound effect on the design of many electrical devices, including magnetic separators, DC motors, and automobile generators.

The first major market for nickel was in the production of nickel and nickel–chromium steels for armour plate, an application based on the work of James Riley of Glasgow in 1889 and tests by the U.S. Navy in 1891 on armour plate from a French steel producer. Military demands supported the industry for many years, but with the development of steam-turbine power plants, the automobile, agricultural machines, and aircraft, a whole new group of high-strength steels containing from 0.5 to about 5 percent nickel along with chromium, molybdenum, etc., were developed. More recently, with a demand for steels for ultralow-temperature use with liquefied gases, steel of 9 percent nickel and alloys of higher nickel content have come into demand. These steels rely on carbon for hardening by heat treatment. The nickel toughens the steel and slows the hardening process so that larger sections can be heat-treated. A carbon-free iron alloy known as maraging steel has been developed. It contains 18 percent nickel, plus cobalt, titanium, and molybdenum. This alloy can be heat-treated to provide a tensile strength of some 21,000 kilograms per square centimetre (300,000 pounds per square inch), coupled with an elongation of 5 percent to 10 percent.

High-strength steels

Nickel is resistant to oxidation at high temperatures and to electrical erosion. For these reasons, alloys that are high in nickel, such as the 4 percent manganese alloy, are used for spark plug electrodes in automobiles and for other types of ignitors. The addition of 15–20 percent of chromium to nickel vastly improves oxidation resistance so that such alloys are used for electric resistance heaters. Alloys containing 15–20 percent chromium, plus various amounts of iron, and an alloy containing 35 percent chromium, 20 percent nickel, with the balance iron, find extensive industrial use where high strength and corrosion resistance over a wide range of temperatures are essential. The addition of small amounts of aluminum and titanium permits the alloy that is high in nickel and chromium to be further strengthened by precipitation treatment. Alloys of this general type made the jet-aircraft engine possible. Gas turbines require the same alloys and are growing in importance for industrial power uses.

A large group of alloy steels ranging from 18 percent chromium, 8 percent nickel, and 25 percent chromium, 20 percent nickel, to 20 percent chromium, 35–40 percent nickel are employed where corrosion resistance is a major requirement. The stainless steels, of which the 18-percent-chromium–8-percent-nickel variety is the best known, are widely used where stain and corrosion resistance must be coupled with high strength. The largest single use of nickel is in the production of stainless steel.

Nickel compounds. The compound nickel sulfate hexahydrate, $NiSO_4 \cdot 6H_2O$, is employed in the electrolytic refining of nickel as well as in most nickel electroplating baths. Nickel chloride hexahydrate, $NiCl_2 \cdot 6H_2O$, is often used in conjunction with the sulfate in plating baths; while the nickel sulfamate, $Ni(SO_3NH_2) \cdot 4H_2O$, and the nickel fluoborate, $Ni(BF_4)_2$, are employed in some of the newer types of electroplating baths.

Nickel dimethylglyoxime is an insoluble salt useful in analytical chemistry in precipitating nickel. Nickel carbonyl, $Ni(CO)_4$, a liquid at room temperature, is employed in a carbonyl-nickel-refining process. Like all other carbonyls, it is poisonous. Nickel sub-sulfide, Ni_3S_2, is the nickel component of matte involved in pyrometallurgy. Nickel oxide, NiO, is involved in refining processes and also may be an end product.

Pentlandite, $(Ni,Fe)_9S_8$, is the most important sulfide mineral; but pyrrhotite (FeS to Fe_7S_8) with some nickel replacing iron also is important in nickel and iron-recovery processes. Millerite, NiS, is a rare mineral. In the laterites, garnierite, $(NiMg)_6Si_4O_{10}(OH)_8$, is the richest in nickel; but nickeliferous limonite, $(Fe,Ni)O(OH) \cdot nH_2O$, constitutes a major portion of the laterites.

PLATINUM-GROUP METALS

Platinum and palladium are far more important commercially than the other members of the platinum-group metals, which also include iridium, osmium, rhodium, and ruthenium. The platinum metals are notable for their wide range of applications, many of which are based on their high melting points, resistance to corrosion, and catalytic properties.

Industrially useful properties

The first platinum to reach Europe came from placers in the Chocó district of what is now Colombia, South America. Gold also was recovered from the same placers, which are still being worked. The new material was brought to England by Sir Charles Wood in 1741 and aroused widespread scientific interest as its complexities were gradually unravelled. A method for separating quite pure platinum from the ore was devised, and the metal was found to be ductile.

The high melting points of these metals were a limitation to their use, but researchers in England, France, Germany, and Russia devised methods for consolidating and working platinum into useful forms.

Ores and mining. Platinum ores usually contain more platinum than the other platinum metals, except for the relatively rare osmiridium.

Where deposits containing platinum minerals have been subjected to weathering, the high density of platinum and its corrosion resistance have led to its concentration in riverbeds. In a number of locations, gold also is present in the same placers, but the original sources differ. The Chocó placers mentioned earlier were overshadowed by platinum from the Urals of Russia, which dominated the world supply from 1825 until about 1914.

Platinum was long known to exist as an arsenide, sperrylite, in Canadian copper–nickel–iron–sulfide deposits, and some recovery was made beginning in 1900. Not until the adoption of electrolytic methods for refining copper and nickel in Canada and the carbonyl process in England, however, did substantial production result. Platinum and palladium dominate in these deposits, but all of the platinum metals are present and are now recovered.

Before 1930, palladium was relatively rare, being derived from various platinum ores and from the electrolytic refining of copper, a situation that changed in 1930 when production by the International Nickel Company of Canada, Ltd., sharply increased and the metal became available in quantities adequate for industrial use. Production of palladium as well as platinum from copper–nickel deposits in the U.S.S.R. began to increase in the 1960s.

Production processes. *Recovery from South African pyroxenite deposits.* In working alluvial deposits, the high density of the platinum minerals is relied upon for recovery, as is the case with gold. In treating platinum ore from the pyroxenite deposits in the Merensky horizon in South Africa, now the largest sources of platinum, the ore is crushed, then finely wet-ground in ball mills. Then the pulp flows over corduroyed tables that recover free metallic particles and coarse platinum-bearing minerals. The recovered material is further concentrated and shipped to a refinery in England. The tailing from the table separation is thickened to remove much of the water and is then sent through a flotation cycle that recovers a concentrate of nickel, copper, and iron plus the remaining platinum minerals. The recovered material is dewatered, filtered, dried, and pelletized. The pellets are fed to blast furnaces, then converted into a copper–nickel matte containing substantial amounts of platinum metals, which can be treated to separate the copper and nickel sulfides and ultimately produce electrolytic copper and nickel. The platinum metals are recovered from the anode slimes from the nickel electrolysis. Alternatively, the Wise–Vines process can be employed, in which a small amount of silicon is added to the molten matte, followed by slow solidification. A nickel-rich bottom, which carries most of the platinum metals, separates and can be worked up promptly in a platinum metals refinery.

The Wise–Vines process

Recovery from Canadian copper–nickel ores. At the operations of the International Nickel Company of Canada, Ltd., the ore is selectively floated to yield fractions containing copper and nickel. The nickel fraction, after blast furnace and converter treatment, yields a nickel-rich matte, is cooled, crushed, ground, and passed over a magnetic separator to recover nickel-rich metallics, which contain most of the platinum metals. The remaining material is given further treatment and again selectively floated to yield a low-copper and a high-copper sulfide. The latter and the earlier copper concentrates are refined by the usual electrolytic methods, as is the nickel sulfide concentrate. Gold and silver are recovered from the copper electrolysis; any remaining platinum metals are recovered from the nickel electrolysis.

Some nickel matte is converted to oxide and shipped to Clydach, Glamorgan, where it is treated by the Mond carbonyl process, in which the nickel is converted into gaseous nickel carbonyl. The platinum metals remain behind, and this material, along with platinum metal concentrates from Canadian operations, is treated at Acton, England, to isolate the individual platinum metals.

In refining, it is essential that all the gold be precipitated and that no trace redissolves, as any gold will form the explosive fulminate when ammonia is added in the palladium recovery operation. Even a small amount of fulminate can cause a serious explosion when the palladium precipitate dries. It also should be noted that the rhodium compound $(NH_4)_2NaRh(NO_2)_6$ will explode if heated. Osmium is easily oxidized, forming volatile osmium tetroxide, OsO_4, which will seriously damage the eyes.

Pure platinum metals and their alloys. The mechanical

properties of the six platinum metals differ greatly. Platinum and palladium are rather soft and very ductile; these metals and most of their alloys can be worked hot or cold. Rhodium is initially worked hot, but cold-working can be done later with rather frequent annealing. Iridium can be worked hot, as can ruthenium, but with difficulty. Neither metal can be cold-worked appreciably.

Hardening by alloying

Pure platinum and pure palladium are too soft for many applications and so are hardened by alloying, generally with one of the other metals of the platinum group, usually iridium or ruthenium. Rhodium is a less potent hardener for platinum but confers special advantages for high-temperature and certain catalytic uses. Platinum–rhodium, platinum thermocouples are standard for the measurement of high temperatures. (A thermocouple is a junction of dissimilar metals, in this case pure platinum and a platinum–rhodium alloy. A voltage is generated when this junction is heated.) All of the platinum metals can be electrodeposited, palladium being the easiest to process. The deposited metal is much harder than the wrought metal.

Platinum. Platinum alloyed with 3.5 percent of rhodium is the standard material for crucibles in the laboratory, where they are subject to a variety of corrosives, including molten salts. A large-scale use for platinum and platinum with 10 percent and 20 percent rhodium for lining tanks has developed in the production of high-quality glass. Even more important is the use of the platinum–10-percent-rhodium alloy for the bushings through which molten glass flows in producing glass fibre, a new industry made possible by platinum.

Rayon fibre is produced by forcing a cellulose xanthate solution through hair-fine holes in a spinneret made of platinum–rhodium or a gold–platinum–palladium alloy.

Insoluble platinum–iridium anodes are essential to the production of perchlorates and persulfates, while platinum-coated titanium is used as an insoluble anode in some electroplating operations and in cells for producing chlorate as well as in cathodic corrosion protection systems. Titanium coated with ruthenium dioxide is used for chlorine production.

Platinum–10-percent rhodium in the form of gauze, operating at bright-red heat, serves as the catalyst in the reaction between ammonia and air to produce nitrogen oxide and ultimately nitric acid, a standard process throughout the world. By feeding in some methane along with the ammonia mixture, hydrocyanic acid can be produced.

Use by petroleum industry

Platinum has long been employed as the catalyst in many other processes, but the most important new application is in the petroleum industry to produce improved gasoline and other hydrocarbon products. In this and many other applications, most of the platinum in the spent catalyst can be recovered. Plant expansion accounts for most of the demand for the metal.

The use of platinum for fine jewelry began about 1906. It continues to be used in jewelry, but industrial demand has provided the impetus for opening new mines.

Platinum alloys have long been used as spark-resistant electrical contact materials, but for some years the largest electrical use was for lead-ins for incandescent lamps. Because of its resistence to spark erosion, platinum was used in spark plugs for combat aircraft in World War II.

Palladium in relay contacts

Palladium. Pure palladium was found to be very desirable for contacts in the relays of telephone and other wire communications systems, because it provided long life and a high level of reliability. This use continues to be the major one.

Hydrogen diffuses rapidly through palladium and some of its alloys at moderately elevated temperatures. Industrial application is made of this property by utilizing tubes with hair-thin walls of a palladium–25-percent-silver alloy to separate high-purity hydrogen from mixed gases.

Palladium hardened with ruthenium is used in certain types of jewelry, and a number of palladium-content brazing alloys are in use in the aircraft industry.

The range of corrosives that can be handled with titanium can be widened by alloying it with small amounts of palladium; alloys containing about 0.2 percent palladium are of commercial importance.

Other platinum metals. Rhodium is the whitest of the platinum metals, a property responsible for its extensive use as an electroplate in applications ranging from the finishing of jewelry to the production of 152-centimetre (60-inch) searchlight reflectors. Rhodium has useful catalytic properties that are receiving attention. It is also suitable for electrical contacts and is particularly useful as an electroplated surface on rubbing electrical contacts. The use of rhodium as an alloy with platinum for high-temperature use is important, and the platinum, platinum–10-percent-rhodium thermocouple is standard for measurement of temperatures up to about 1,500° C (2,700° F). For higher temperatures, a thermocouple made of platinum with 7 percent and 30 percent rhodium is available. Platinum–rhodium alloys also are used for the heating windings in laboratory furnaces. Pure rhodium as well as pure iridium crucibles are used in growing crystals for laser and related applications.

Iridium, the most corrosion-resistant of the platinum group metals, is a very effective hardener for platinum. Its extremely high melting point renders it useful in special situations where oxygen can be limited, as in growing special crystals. In addition to the small percentage of iridium found in most platinum deposits, natural alloys of iridium with osmium are found in a few locations around the world. In addition, a few thousand ounces of osmiridium are recovered yearly from the gold ores of South Africa. Iridium has outstanding resistance to lead compounds at high temperatures, one of the reasons for its use in ultralong-life spark plugs.

Sources of iridium

Ruthenium is a good hardener for platinum, the 12-percent alloy being used for contacts in voltage regulators, thermostats, and magnetos. It is used to harden platinum and particularly palladium for use in jewelry. Ruthenium-content alloys are also used for tipping the nibs of fountain pens. The electrodeposition of ruthenium is possible, and its hardness recommends it in instruments where a low-pressure rubbing contact is required.

Osmium is the rarest and hardest of the group and has the highest melting point, but its ready oxidation is a limitation. It has been employed in the synthesis of cortisone.

Purity. Traces of impurities in refined platinum metals can be estimated by the spectrography of the arc emissions. In the case of platinum, the measurement of the thermal electromotive force (EMF) against a sample of extremely pure wire often is employed, as is a measurement of the temperature coefficient of electrical resistivity between 0° and 100° C (32° and 212° F).

Within the temperature range of −183° C (−297° F [the boiling point of oxygen]) and 630° C (1,170° F), the platinum resistance thermometer defines the temperature on the International Practical scale. The range from 630° C to 1,064° C (1,947° F) is covered with the platinum, platinum–10-percent-rhodium thermocouple. In view of the importance of both the temperature coefficient and the thermal EMF against a sample of extremely pure platinum, a detailed study has been made of the effect of single impurities on these properties. (E.M.W.)

SILVER

Silver is in the precious-metal group because of its scarcity and price. Although it is widely distributed in nature, the total amount in the Earth's crust is quite small compared with other metals. As a consequence it costs about 500 times more than iron and about 50 times more than copper. Among the noble metals, however, it is relatively abundant and costs about 1/20 as much as gold and 1/50 as much as platinum. Noble metals are not easily changed to an oxide form by air or oxygen and consequently can be found in the elemental form or as an alloy in nature. They can be used for electrical contacts at which arcing (sparking) occurs, and they are excellent conductors of electricity. Silver is the best conductor of electricity and heat among all metals and is used as the electrical and heat-conductance standard against which other metals are measured in percentage terms. Copper, for example, has a relative electrical conductance of 95 percent and a thermal conductance of 73.7 percent.

Because of its resistance to atmospheric oxidation, as

well as its beauty and value, silver has been used since antiquity for ornaments, jewelry, and coins. It is the whitest of all metals and has excellent ductility and malleability; pure silver, however, can be easily scratched and tarnishes quickly in the presence of certain sulfur compounds. Silver objects have been found in ancient tombs unaffected by air pollution but have tarnished rapidly in areas in which coal containing sulfur is burned. The reflectivity of silver, particularly to that part of the spectrum visible to the human eye, is one of the highest among metals. When properly polished and under favourable conditions, about 95 percent of the light striking the surface is reflected.

Silver ornaments, vessels for ceremonial services, and decorations have been found in royal tombs dating back to 4000 BC. The Egyptian Code of Menes, about 3000 BC, decreed that "one part of gold was equal to two and one-half parts of silver in value," perhaps the first gold standard. Records show that gold and copper were the first metals to be worked by man, with silver third. The Romans, who probably advanced the science of metallurgy further than any other people up to their time, had large plants to produce silver for ornaments and decoration. The Roman process, in which the silver was separated from the ore by heating in a furnace, was used throughout the Middle Ages, when silver–copper mines were worked in central Europe. In the 16th century the Spaniards discovered enormous silver deposits in South and Central America, and during the 19th century, a so-called patio process, involving grinding silver ore in the presence of salts, water, and mercury to make an amalgam (alloy), was developed in Mexico. The amalgam (mercury and silver and gold if present) was heated in a retort to drive off the mercury, leaving the silver and gold; the mercury was recovered by condensation. Spanish America supplied the world with silver until the Comstock Lode strike of 1859 in the Sierra Nevada area in the United States.

The cyanide process (see below), developed in the latter part of the 19th century, is still used today for ores containing only gold and silver values. About 1910 a froth flotation process was introduced. In this process, particles of sulfides of copper, lead, and zinc containing small amounts of silver and gold are made to adhere to rising air bubbles, which are skimmed off. The froth flotation process today accounts for about 75 percent of the world production of silver.

Silver ore associations

Occurrence. Since the early 1970s, most of the silver produced from ores has been associated with the sulfides of lead, copper, and zinc. These so-called argentiferous ores may contain from a trace to several thousand troy ounces of silver per ton of ore. Precious metals are generally sold by troy ounces, one troy ounce equalling 31.1 grams, or 1.097 avoirdupois ounces.

Unlike gold, silver is found in many naturally occurring minerals. The most important is galena (lead sulfide), followed by argentite (silver sulfide), cerargyrite (or chlorargyrite; silver chloride), polybasite (silver antimony sulfide), proustite (silver arsenic sulfide), pyrargyrite (or stephanite; silver antimony sulfides), and others. Deposits of native silver (generally an alloy) are commercially important and widespread. Sylvanite (gold and silver telluride) ores are of commercial value in some areas. Arizona, Utah, and Colorado in the United States are large producers of silver as a by-product from copper, lead, and zinc ores. Low-grade porphyry copper ores mined in Arizona assay about 0.6 percent copper and about 0.1 ounce of silver per ton of ore. In Canada, silver is found in the Canadian Shield around Sudbury, Ontario, as part of the nickel–copper–cobalt ores, in British Columbia, and in the Yukon Territory. Large silver deposits at Cerro de Pasco, Peru, have been worked since the 17th century. Mt. Isa and North Broken Hill in Australia are becoming large producers.

Refining and recovery. Many methods are employed for the extraction of silver or silver and gold from ores, scrap, alloys, and used photographic materials. Most of the silver produced today is a by-product from the treatment of copper, lead, and zinc ores, and consequently any large increase in silver implies an increase in copper, lead, and zinc or vice versa. Silver occurring with copper ores is first extracted with the copper and recovered as a silver–gold anode mud after electrolysis. Silver occurring with lead ores is recovered from the lead bullion as a silver–gold dross (see above *Lead production*). Silver occurring with zinc ores, which also contain lead, is recovered with the lead bullion as a by-product (see above *Zinc production*).

The cyanide process is generally used on ores that contain only silver or silver and gold as the valuable constituent. The ore is pulverized and leached with a dilute sodium cyanide water solution. The silver and gold form water-soluble sodium–silver–gold complexes. The resulting mixture is filtered and the solids discarded. The filtrate is generally treated with finely divided zinc dust that causes the silver and gold to precipitate from the solution. This precipitate is filtered off, melted, and cast into bullion bars.

Silver is recovered from used photographic film by burning the film and dissolving the silver in the ashes with nitric acid. The silver is precipitated with sodium chloride and the resulting precipitate filtered off and cast into bars. Silver is also recovered from photographic processing solutions by replacing the silver with another metal, such as zinc, by electrolysis or by chemical precipitation with chemicals such as sodium hydroxide or sodium sulfide.

The method used in refining the crude silver or silver–gold bullion obtained from these metallurgical processes depends upon its purity and the amount of bullion to be refined. For bullion containing a large amount of impurities, smelting in a small furnace with lead oxide, fluxes, and a reducing agent produces slag that is further processed in a lead smelter. The resulting lead bullion, containing the silver and gold, is melted in the presence of air. The oxidized lead (a liquid oxide) is evaporated or absorbed by a cupel (holding vessel), a process called cupellation. The end product is an alloy of silver and gold called doré, a high-purity alloy. The cupel material (the liquid lead oxide absorber) can be cement or bone ash.

Electrolysis and parting

The two most important methods for separating silver and gold in silver–gold bullion are electrolysis and parting. In electrolysis, a silver nitrate water solution is employed, with silver, gold bullion, or doré serving as the anode (positive electrode). Pure silver is deposited at the cathode (negative pole), which can be a pure thin sheet of silver or stainless steel. The resulting cathode silver or silver stripped from the stainless steel cathode is melted and cast into bars assaying about 999.0+ fine of silver. Gold is recovered from the anode mud.

Silver and gold alloys are stated in terms of fineness, given by the parts of silver or gold per 1,000 of alloy. Sterling silver contains 92.5 percent silver and 7.5 percent of another metal, usually copper as a hardening agent; such silver is therefore 925 fine.

In the parting, or wet, method the doré or silver–gold bullion is placed in a bath of hot concentrated sulfuric acid or strong nitric acid, and the silver is allowed to dissolve. Gold is recovered from the residue. The clear solution contains a soluble silver sulfate or nitrate; it is treated with ferrous sulfate to precipitate the silver, which is filtered off and melted into bars assaying 995+ fine of silver. This method is applicable to doré or very high purity silver–gold bullion containing two or more parts of silver to one part of gold. Additional silver may be added to achieve this ratio, as ratios of less than two to one will not dissolve in the acid.

The metal and its alloys. *Pure metal applications.* By the late 20th century coinage consumed less than 5 percent of the world's silver production. The rest went to industrial production, with photography accounting for the largest proportion of this total. Photographic silver is used in the form of halides, such as silver bromide. Television cassettes contain film using silver halides. Electrical and electronic equipment also accounts for a large percentage of silver consumption. Silver solders are used extensively in refrigeration units, air conditioners, automobiles, and electrical appliances. Brazing, soldering, and welding use substantial amounts of silver. Silver–copper alloys are used in welding and, in some cases, as contacts in relays for computers, tabulators, and other applications requiring high strength at elevated temperatures. Inexpensive though intricate electrical circuits for radios, television, computers, and other electronic devices may be made by stamp-

ing or printing a given circuit on insulating panels with silver ink. Silver is used in certain cold epoxy solders to bond microminiature electronic devices without heat and in some electronic adhesive applications. Silver–magnesium–nickel alloys are used for high thermal-conductivity applications and relay switches operating at high temperatures. Silver coatings on ceramics, glass, mica, and other insulators are widely used in electronic devices.

Silverware, plated and sterling, takes another large share of the production. An average sterling teaspoon contains about one ounce of silver. Electroplated ware also takes a considerable amount of silver. Thin plates of silver for bone replacement in the human body, silver wire to prevent movement of broken bones, and drainage tubes of silver are useful in medicine because the silver exhibits germicidal properties. These properties may result from the ability of silver to absorb oxygen. Liquid silver, for example, can absorb about 20 times its volume of oxygen. Silver water purifiers developed to make water bacteria-free for space explorers are now also used by campers and sportsmen. High-silver batteries, requiring less space, maintaining higher voltages, and lasting longer than conventional batteries, are used for such applications as hearing aids. Because of the higher power per unit of weight, silver batteries are used in space travel, submarines, and portable television cameras.

Manufacturers of airplanes and diesel locomotives have been using a silver-base-alloy bearing containing 3 to 5 percent lead, which is much less subject to seizure (freezing) than Babbitt metal and has excellent oiliness under severe conditions. Because of its excellent reflectability, silver is used extensively as a coating on glass and other materials for making mirrors. Coatings may be applied in many ways, but generally the process uses an ammonium hydroxide–silver nitrate solution, which is applied to the surface, and a reducing agent, such as sugar, Rochelle salt, or formaldehyde, which is added to precipitate a layer of metallic silver. The process is hazardous because of the formation of ammonium nitrate; explosions may occur on the slightest disturbance.

Silver coatings

Silver in colloidal form (very fine particles) is used as a catalyst in certain vapour-phase organic chemical reactions, such as the manufacture of various types of alcohols. Cesium–silver alloys are used in some types of photoelectric cells. Dental amalgams are alloys of silver–tin–mercury that can be molded at temperatures of the mouth and harden quickly to give a polished surface.

Silver is also used in the manufacture of jewelry and art objects. Jewelry silvers are alloys containing around 80 percent silver and 20 percent copper.

Chemical compounds. Silver is easily tarnished by sulfur, sulfur oxides, and hydrogen sulfide. The black tarnish can be removed by household silver polish, which generally contains fine abrasive material, or by an electrolysis kit that removes the tarnish with no loss of silver. An antitarnish dip of 1.5 percent tin chloride water solution, used for five to 30 seconds and followed by a water rinse, may prevent tarnishing. Silver dissolves rapidly in strong or dilute nitric acid to form water-soluble silver nitrate. Silver nitrate solutions are used as fungicides in treating certain plants for disease control and as a starting point to make other materials, such as silver for photographic film and plates and for mirrors. The halogen compounds of silver are by far the most important; silver bromide makes possible the modern photographic industry. In photography the glass plate or film merely serves to hold the light-sensitive emulsion, generally silver bromide crystals, in suspension in some type of a gelatin. Silver chloride, a white crystal material practically insoluble in water, is made by adding hydrochloric acid to a silver nitrate solution. Silver iodide, a yellow crystal material, is made in a similar manner. In 1946, testing began on the use of finely divided silver iodide, which has a crystal structure similar to ice, for cloud seeding to produce rainfall, dissipate fog, and reduce the violence of hurricanes and hailstorms. Its effectiveness is open to some question, but one ounce of silver as silver iodide under the right conditions might produce as much as 11,350,000 litres (3,000,000 gallons) of water as rain.

Halogen compounds of silver

Argentic fluoride (AgF_2), a dark brown solid formed when fluorine reacts with argentous fluoride (AgF), is a powerful oxidizing agent and a good fluorinating agent. Silver sulfadiazene is used in preventing infection resulting from burns because it combats bacteria by attacking the wall and not the nucleus of the cell. (S.L.Sm./Ed.)

SODIUM AND POTASSIUM

Sodium and potassium are both soft, silvery-white metals. When freshly cut they have a brilliant lustre, but since they are among the most reactive of the metallic elements, they tarnish in contact with air, turning a dull, gray colour. Because they are extremely reactive, sodium and potassium are not found in nature in the free state (as metals) but are widely distributed in combined form as compounds. The most familiar compound is sodium chloride—ordinary salt. Other salts of sodium and potassium are found in nearly all soils and terrestrial waters and in many rocks. The rocks and soil are richer in potassium, but the oceans contain only about $^1/_{30}$ as much potassium as sodium. Sodium and potassium are the sixth and seventh most abundant elements in the Earth's crust, of which they constitute 2.8 percent and 2.6 percent, respectively.

Some of the earliest civilizations knew and used materials containing compounds of both sodium and potassium, and as these compounds were similar, they did not distinguish between them. Records originating in lower Mesopotamia and dating from the 17th century BC mention saltpetre (potassium nitrate) for making glazes. The Egyptians used native sodium carbonate as early as the 16th century BC for making glass.

Early methods of preparation

Both sodium carbonate and potassium carbonate were termed alkali and were prepared from early times by passing water through the ashes of burned plants and evaporating the water from the solution. The term soda was applied at first to either alkali and later only to the ashes of sea plants; the term potash, however, was restricted to the ashes of land vegetation, and the term originated from the use of large pots from which the solutions were evaporated. The two alkalies, soda and potash, were successively designated as natural and artificial and as mineral and vegetable; but finally, in the 18th century, it was found that the so-called mineral alkali occurred in the ashes of sea plants and the so-called vegetable alkali occurred in a number of minerals; then the terms soda and potash were assigned to specific substances.

In 1807 the brilliant English chemist Sir Humphry Davy decomposed both alkalies, obtaining first the metallic potassium and then sodium. He coined the names potassium and sodium, latinized versions of potash and soda. Among the strongest reducing agents known, sodium and potassium cannot be isolated by ordinary chemical reduction, so it is not surprising that they were only prepared after electric current was discovered in 1800 and electrolytic methods were developed.

Occurrence. *Sodium.* Sodium occurs in the form of halides, silicates, carbonates, sulfates, and nitrates. The principal source of sodium, of such importance that it outweighs all others, is sodium chloride, particularly rock salt deposits, but also salt evaporated from brine from underground salt deposits and salt obtained from seawater by solar evaporation.

Sodium is disseminated throughout the rocky crust of the Earth in the form of soda feldspar (*i.e.*, as silicate of sodium) and in various other sodium-containing rocks. Sodium carbonate is also widely dispersed in nature, forming constituents of many mineral waters and occurring as the principal saline component in natron or trona (hydrated sodium carbonate, sodium bicarbonate) lakes, as efflorescences, and as solid crusts found at the bottom of certain salt lakes. Gaylussite is a double salt, $Na_2CO_3 \cdot CaCO_3 \cdot 5H_2O$, occurring in Colombia.

Sodium sulfate occurs in an anhydrous condition as thenardite and in a hydrated condition, Glauber's salt. Glauberite is a double sulfate of sodium and calcium. Cryolite is a fluoride of aluminum and sodium, mined extensively for industrial purposes.

Potassium. Potassium occurs in many silicate rocks and minerals. The most important source is salt deposits,

but a small percentage is obtained from plant and animal sources. Only a small fraction of the forms of potassium are economical to recover, those compounds that are water-soluble, viz., potassium chloride and potassium sulfate; they are frequently found in mixtures with magnesium salts, both as dry mineral deposits and in brines. Their solubility allows them to be easily extracted, and they are readily amenable to chemical operations. Although nearly as abundant as sodium, potassium is less accessible; the relatively few workable deposits contain saline residues formed by the evaporation of inland seas of past geological eras.

Plants as potassium sources

All plants are potential sources of potassium compounds. The proportion of potassium present in organisms varies widely; but in higher animals and most vegetation, except some marine plants, the potassium content is greater than that of sodium. Sugarcane and sugar beets contain significant quantities of potassium salts, 90 percent of which is extracted with the sugar. Soluble potassium salts are drawn into the roots of plants from fertile soil and accumulate in the plant structure, where some of them are metabolized into organic salts, such as tartrates and oxalates. If, however, the plants are burned, the organic potassium salts are converted into potassium carbonate.

Recovery of the metals. When Davy first isolated potassium and sodium in 1807, he did it by electrolysis of the fused hydroxides. A modification of his method, called the Castner process, was long used commercially to prepare sodium, but it was not entirely satisfactory for the more difficult potassium reduction. Most sodium is now prepared by the Downs process, in which a fused mixture of sodium chloride and either calcium chloride or sodium carbonate is electrolyzed to yield chlorine as well as the metal. Pure molten sodium chloride is not used because at the high temperature required to melt it (801° C [1,474° F]) the vapour pressure of sodium metal is dangerously high, and sodium is appreciably soluble in the molten salt. With the fused mixture, a lower temperature (in the vicinity of 600° C [1,100° F]) may be used.

Potassium metal was obtained in 1808 by chemists Joseph-Louis Gay-Lussac and Louis-Jacques Thenard of France by reduction of the hydroxide with iron; they packed the hydroxide and iron wire or turnings into a gun barrel, which was coated with clay and heated to white heat. During the same year, François-René Curaudeau prepared potassium by reducing potassium carbonate with carbon. During the following 50 years, various chemists studied this reaction to devise methods of separation and rapid cooling of the products; such methods were necessary to prevent the explosive compound $K_6C_6O_6$ from forming.

Reduction methods using metals

The processes that evolved were the forerunners of reduction methods using metals, such as magnesium and aluminum, or using compounds, such as calcium carbide, hydride, or silicide. After high-vacuum techniques were developed, calcium could be used to reduce potassium to the pure metal in a vacuum; sodium vapour has been used similarly as a fairly successful reducing agent for potassium.

Electrolytic methods of preparing potassium entail decomposing the fused hydroxide, the cyanide, or a potassium oxysalt dissolved in a fused potassium halide.

The metals and their alloys. *Sodium.* Metallic sodium is used to manufacture sodium chemicals of technical importance, particularly sodium peroxide, sodium cyanide, sodium amide, sodium hydride, sodium azide, and sodium borohydride. Sodium is also used extensively in synthetic chemistry; *e.g.,* as a reducing agent and a drying agent.

Because sodium exhibits the photoelectric effect to a marked degree, emitting electrons when exposed to light, it has been used in the manufacture of photoelectric cells. Because of its relatively high heat capacity and high heat conductivity, sodium is used as a heat-transfer medium. In some types of nuclear reactors, molten sodium carries heat from the core of the reactor to a boiler for generation of steam. Metallic sodium is also being investigated as an electrical conductor, when properly protected from air and moisture.

Sodium-vapour lamps, in which gaseous sodium is made to glow inside a tube by an applied voltage, give a yellow-coloured light of good illumination characteristics and are extensively used for street and road lighting.

Potassium. Metallic potassium has few uses, the most important being the synthesizing of potassium compounds, such as potassium superoxide, KO_2, used in protective breathing equipment. In the rebreather gas mask, the superoxide reacts with exhaled air to liberate oxygen and remove carbon dioxide and water vapour.

During the first half of the 19th century, metallic potassium was used to produce aluminum by the reduction of the chloride; magnesium, boron, and silicon were produced by a similar method. In 1854 it was found that sodium, which is cheaper, serves as well as potassium, and sodium was used for these purposes until electrolytic processes were developed. Sodium also replaced potassium in many other industrial processes. Potassium, however, still plays a role in certain organic syntheses.

A radioactive isotope of potassium, potassium-40 (^{40}K), has been used to estimate the age of the Earth and some of the Moon rocks.

Alloys. Alloys of sodium and potassium serve as heat-transfer media, such as are described above for sodium alone. An alloy of 77.2 percent potassium and 22.8 percent sodium has been suggested as a liquid coolant for nuclear reactors. A potassium–sodium alloy takes part in the reduction of titanium tetrachloride to the metal.

Sodium–lead alloy often is substituted for metallic sodium as a reducing agent or a drying agent and is also employed in the manufacture of tetraethyllead, a common constituent of antiknock gasoline.

Chemical compounds. *Sodium compounds.* Sodium chloride (common salt) is the most abundant form and the most important source of sodium. It is the main salt present in seawater, of which it comprises about 80 percent of the dissolved matter. Throughout the world there are extensive deposits of sodium chloride, which probably originated when prehistoric seas evaporated. It is a white, cubic crystalline salt, very soluble in water; the large crystals are transparent. Present in all body fluids, sodium chloride is essential to human and animal nutrition, and it is a food preservative and seasoning. Sodium chloride is important in the manufacture of soaps, dyes, ceramic glazes, leather, and other sodium-containing chemicals.

Abundance of sodium chloride

Sodium hydroxide, or caustic soda, the most important industrial caustic, is a white, deliquescent (moisture-attracting) solid that may be obtained from a solution of sodium chloride by electrolysis. It is highly corrosive to animal and vegetable tissue. The alkaline solutions it forms when dissolved in water neutralize acids in various commercial processes: in petroleum refining, it removes sulfuric and organic acids; in soapmaking, it reacts with fatty acids. In the making of cellophane, paper, viscose rayon, and other products, its alkaline solutions enter into the treatment of cellulose and into the manufacture of many chemicals.

Sodium carbonate is a crystalline white salt soluble in water; in the technical grade, the anhydrous form is known as soda ash, and the decahydrate is known as washing soda. It was first manufactured on an industrial scale by the Leblanc process, which since has been supplanted by the Solvay, or ammonia-soda, process. (For descriptions of these processes, see INDUSTRIES, CHEMICAL PROCESS: *Chemical industry.*) A general cleaner, sodium carbonate has many industrial applications, and in medicine it is used to treat skin inflammations and to cleanse body orifices.

Sodium bicarbonate (sodium hydrogen carbonate), household baking soda, is a white crystalline powder or granular salt soluble in water; it is prepared principally from sodium carbonate. A source of carbon dioxide gas, sodium bicarbonate goes into the manufacture of effervescent beverages and baking powder and is a component of fire extinguishers. Medically, it counteracts urinary and stomach acidity, cleanses mucous membranes, and dresses burns.

Sodium sulfate in water-soluble white powder or crystalline form is usually called mirabilite or thenardite when obtained from deposits or brines; the name Glauber's salt refers to the decahydrate $Na_2SO_4 \cdot 10H_2O$. Its uses include the dyeing and printing of textiles and the manufacture of glass and paper.

Sodium nitrate's wide range of uses

Sodium nitrate, a soluble crystalline white salt, occurs naturally in the impure form of Chile saltpetre (caliche) or is manufactured from sodium carbonate and nitric acid. It goes into the manufacture of chemicals, fertilizer, glass, pottery, enamels, matches, explosives, and dyes. It also serves as a meat-pickling agent.

Sodium peroxide, a yellowish-white powder made by heating metallic sodium in dry air, is an oxidizing agent for bleaching paper, cloth, wood, ivory, and sponges, and for purifying air in confined spaces, such as submarines and airplanes. Although not explosive in itself, sodium peroxide is dangerous and may react explosively with combustible substances, particularly if moisture is present.

Sodium sulfide, prepared by heating sodium sulfate with coal, is important in the manufacture of sulfur dyes and rubber. It is an intermediate compound in the production of other chemicals and a laboratory reagent.

Sodium fluoride is made from sodium carbonate and hydrofluoric acid. An insecticide and pesticide, it is poisonous and may be fatal to human beings if ingested; but when added in minute amounts to municipal water supplies, sodium fluoride reduces the incidence of dental caries.

Sodium alkyl sulfates are important intermediate chemicals that may be derived from petroleum; they are essential in the synthetic-detergent industry, in the dye industry, and in the processing of a number of different oils.

Sodium thiosulfate (sodium hyposulfite) is in the "hypo" solution with which photographers fix developed negatives and prints; it acts by dissolving the unchanged silver salts. It is also used to remove traces of chlorine from bleached paper and textiles.

Water glass is sodium silicate of varying composition; it may be in the form of dry powder, an amorphous mass, or a syrupy liquid. It is a binder in the manufacture of abrasive wheels, grindstones, glass, and porcelain. As a water solution, it was formerly used to preserve shell eggs. In the paper and textiles industries, water glass is used as a protective coating and fireproofing agent.

Sodium amide (sodamide) is a reagent used in the chemical industry to introduce amino groups into molecules, particularly in dye production. Made directly from sodium and ammonia, sodium amide is used in turn along with carbon at red heat to prepare sodium cyanide.

Uses of sodium cyanide

Sodium cyanide solution readily dissolves gold and silver metals in the presence of air and is used to extract these metals from their ores. A violent poison, it is also used as an insecticide and as a source of the fumigant hydrogen cyanide. Sodium cyanide is an intermediate compound in the preparation of other useful chemicals, particularly for the manufacture of dyes, pigments, and nylon.

Sodium hypochlorite is a salt that is unstable, except when it is mixed with sodium hydroxide. Used in solution, it is a bleach, being used in the manufacture of paper, textiles, and other products, as well as in the household.

Potassium compounds. Potassium chloride, a white crystalline solid or powder, is the most abundant of the naturally occurring salts of potassium and closely resembles sodium chloride. Potassium chloride is used in various pharmaceutical preparations, as a fertilizer, and as a starting material in the manufacture of many important potassium compounds, including the hydroxide and the carbonate.

Potassium hydroxide, the caustic potash of commerce, is a white solid that can cause severe burns to human tissue. It dissolves in water to give a strongly alkaline solution. Deliquescent (moisture-attracting), it sometimes serves as a drying agent. Potassium hydroxide may be produced from a solution of potassium chloride by electrolysis. Formerly, it was prepared by reacting wood ashes (potassium carbonate) with slaked lime (calcium hydroxide), and from this primitive alkali, soap was made. Essentially, this was also the standard industrial method for manufacturing potassium hydroxide until the end of the 19th century, when it was superseded by the electrolytic process. Because it is more expensive than sodium hydroxide, the use of potassium hydroxide is somewhat limited, but it does go into the manufacture of soap, particularly liquid soap, and chemicals.

Potassium carbonate (potash) is essential to the production of Bohemian glass, soft soaps, fertilizers, and many inorganic chemicals. The first industrial-scale process for manufacturing this salt was an adaptation of the Leblanc sodium carbonate process, since replaced by a number of modern processes.

Potassium bicarbonate (potassium hydrogen carbonate), a white soluble salt, like sodium bicarbonate, is a source of carbon dioxide gas. It is used in baking powders and fire extinguishers and medically as a buffer for stomach and urinary acid.

Potassium sulfate is a constituent of a number of saline minerals. The colourless or white, hard rhombic crystals are used in fertilizers, in the manufacture of glass and alums, and as a laboratory reagent.

Methods of preparing saltpetre

Potassium nitrate, also called nitre, or saltpetre, is a soluble crystalline white salt. As the quantity obtained from natural deposits is limited, it is prepared either from synthetic nitric acid and potassium hydroxide or from naturally occurring potassium chloride and sodium nitrate. It is a valuable fertilizer and has many other uses: in the manufacture of fireworks, matches, blasting powders, explosives, fluxes, and glass, and in the food industry for curing meats and pickling.

Potassium bromide, a white crystalline salt, is used in photography to make gelatin papers and plates, and in medicine as a sedative and an antiepileptic.

Potassium dichromate is a soluble, bright orange-red crystalline salt prepared from chrome iron ore. It is an important oxidizing agent in leather tanning, and in the chemical, paint, and dye industries. A corrosive poison internally, its dust is caustic to human tissue.

Potassium permanganate, a deep purple crystalline salt, dissolves in water to give beautiful purple oxidizing solutions, that are used as bleaches, deodorizers, dyes, and reagents in laboratory analyses. Medically, it is used as a topical astringent, antiseptic, and antidote for some poisons. (W.L.J./J.B.Pa.)

TIN

Relatively rare despite its worldwide use, tin is found in the ore cassiterite, or tinstone, the oxide of tin, either in lodes or alluvial deposits. The chief producing countries are Malaysia, the Soviet Union, Thailand, Indonesia, Bolivia, and China. Tin exists in two crystalline forms; common or white tin (beta tin), which is stable at temperatures above 13.2° C (55° F), and gray tin (alpha tin), the transformation product. When white tin disintegrates into a gray tin in powder, the condition is known as tin pest or tin disease. Although the transformation temperature is moderate (13.2° C), the change from white to gray tin does not occur unless the metal is of great purity and has been exposed to temperatures well below 0° C (32° F). Small amounts of bismuth, antimony, and lead, the impurities normally present in commercial grades of tin, retard such transformation. Thus, the spontaneous appearance of gray tin at low temperatures is a rare occurrence.

Tin, in the form of its alloy bronze, was widely used in ancient times. Bronze implements with tin contents of 10 to 18 percent dating from about 3500 BC have been found in Ur (Iraq). Tin mining was reported in Caucasia as early as 1500 BC, and later in Persia (now Iran), Great Britain, France, and Spain. The Phoenicians are believed to have played an important part in spreading the early bronze culture by their trade in tin, which their ships brought to the eastern Mediterranean from Great Britain and Spain at least as early as 1100 BC. The Cornish tin mines were famous by the time the Romans conquered Great Britain in the 1st century AD. Tin-producing areas gradually opened up in other parts of the world, in Malaya (Malaysia) and China in the 9th century and in Indonesia and Thailand in the 18th century. Bolivia, Zaire, and Nigeria became important sources by the early 20th century.

Early trade in tin

Pliny the Elder in AD 79 described two alloys of tin and lead, now commonly called solder. The Romans are known to have used solder and tinned-copper vessels. Tinned-iron vessels appeared in Bohemia in the 14th century. Tinned-steel sheet (tinplate) was manufactured in England and Saxony by the middle of the 17th century.

The development of modern high-speed machinery and rail transportation was facilitated by the invention in 1839 of Babbitt metal, an alloy containing 89 parts of tin, 9 of antimony, and 2 of copper; the antimony combined with the tin to form hard particles in the otherwise soft matrix. The metal was developed to provide bearings for steam engines.

The 20th century has witnessed a proliferation of applications of tin in both pure and alloy form. Its principal uses, in addition to tinplate, include piping, wire, tubes, sheet, foil, and anodes for electroplating.

Ores and mining. The mineral cassiterite, or tin dioxide, is the major tin ore of the world. In lode deposits, as found in Bolivia, tin sulfide minerals such as stannite, teallite, cylindrite, and canfieldite are found with the cassiterite in granite and quartz.

Because cassiterite is both heavy and chemically resistant, it forms large placers, or residual concentrations, from which lighter, less stable materials have been removed by various natural forces. Alluvial placers, by far the largest in number and most productive, are deposited in streams where the water flow is slack. Almost 60 percent of the tin production in Southeast Asia comes from alluvial deposits, which yield about 0.24 kilogram of tin metal per cubic metre (0.4 pound per cubic yard).

Residual placers form on top of a bedrock source of cassiterite as chemical decay removes other minerals. In Zaire, residual placers grade downward into weathered lodes and are mined either as placers or open-pit lodes.

Eluvial placers occur at the bottom of slopes after the gravity separation of cassiterite from the decayed tin-bearing rock debris that has been washed or wind blown down the slope. Marine placers lie beneath seawater. Seagoing dredges work such placers off the coasts of Thailand and Indonesia.

Alluvial deposits lend themselves to mining by water-washing methods and gravity separation of the cassiterite. Lode deposits require the more complex operations, including rock blasting and crushing prior to gravity separation.

Placer tin is mined either by marine methods, using bucketline dredges and gravel pumps, or through land-mining techniques, especially open-pit methods. Bucketline dredges are quite similar to the floating dredges that are used to mine other types of placer deposits (see above *Mining*). The tin-bearing mud is excavated at depths down to 43 metres (140 feet) below the water level. In either method, final concentration and removal of associated minerals is carried out on shore using a variety of techniques. Final concentrates, ready for direct smelting, contain almost pure cassiterite (between 70 and 77 percent tin).

In gravel-pump and hydraulic mining methods, water jets break up the tin-bearing material, and the material is then conveyed by various methods to the washing apparatus, or launders. In gravel-pump mining, a pump elevates the watery mud up to a wooden trough, or palong. In hydraulicking, a natural head of water flushes the ore down the gently sloping palong, trapping the heavy tin oxide particles behind wooden slats, or riffles. The preliminary rough concentrates are collected periodically and transferred to dressing sheds for final concentration. In open-pit operations the material is excavated dry by power shovels and transported via conveyor belt to a washing plant and dressing shed.

Underground lode deposits can be found at high altitudes (3,700 to 4,600 metres [12,000 to 15,000 feet] above sea level in Bolivia) or low ones (down to 400 metres [1,300 feet]) below sea level, as in Cornwall, England). The lodes are reached by lateral passages driven into the side of a mountain or by vertical shafts. The ore is broken up by blasting and drilling and then taken above ground for further crushing and grinding. The finely divided ore is then concentrated by the same types of separations used for alluvial deposits.

Ore processing. *Beneficiation.* Concentrates from alluvial deposits are smelted without further beneficiation, or further concentration through refinement. Lode deposits of only 40 to 60 percent tin, on the other hand, must be roasted and acid-leached (see below) to remove sulfur and other metallic impurities.

Typically, the crushed concentrates are loaded into the top of a six-hearth so-called Herreshoff furnace and are moved down to succeeding lower hearths by rotating rakes (rabble arms) and the force of gravity. Temperatures inside the gas-heated furnace range from 315° C (600° F) in the lowest discharge hearth to 760° C (1,400° F) in the middle hearths. Most of the roasting is done in two-hour stages between 650° C (1,200° F) and 760° C (1,400° F). This step of the process removes most of the sulfur and some of the arsenic; the sulfides of iron, copper, bismuth, and zinc are converted to oxide; and lead sulfide is oxidized to sulfate. Adding 1 to 5 percent salt removes some of the volatile chlorides of lead, bismuth, arsenic, antimony, and silver in the fume; the chlorides remaining are water soluble.

Some Bolivian ores contain up to 9 percent tungsten, which is recovered by mixing sodium carbonate with the roasted concentrate and heating it to 600° C (1,100° F) to form sodium tungstate, a water-soluble compound from which the tungsten can be recovered by evaporation.

The roasted concentrates are leached with hydrochloric acid to remove most of the iron, zinc, lead, silver, bismuth, and copper impurities. After washing and drying, the purified concentrate is ready for smelting.

Problems of tin metallurgy

Smelting. The metallurgy of tin is comparatively simple, though it does present a few problems: (1) the temperature required to reduce tin dioxide with carbon also reduces the oxides of other metals in the ore (reduced iron, for example, forms hard compounds with tin); (2) since tin at smelting temperatures is extremely fluid, it finds the most minute openings for escape and soaks into porous refractories; (3) tin reacts with furnace linings, thus producing slags that must be retreated.

Because of those problems and the resulting high tin content of the slag, smelting is accomplished in two stages: (1) primary smelting in either reverberatory or blast furnaces and (2) retreatment of slags (unwanted hard compounds of tin and iron) and refinery dross. Most large tin-smelting plants use reverberatory furnaces for both primary smelting and retreatment processes. Operations are more easily controlled, and cleaner slags are produced than in the older blast furnaces. Loss of tin dust is much smaller when the reverberatory furnace is used to smelt finely divided ores and concentrates.

The reverberatory tin-smelting furnace is a relatively new and special type of blast furnace, in which heat is reflected from the roof of the furnace onto the material. Thus, the material does not come into contact with the fuel. The hearth slopes toward a taphole that is near the centre and to one side and low enough to drain the furnace completely. The size of the hearth varies, the largest being nine to 12 metres (30 to 40 feet) long and four to five metres (12 to 15 feet) wide. The larger reverberatory furnaces will take an 8- to 15-ton load that contains tin concentrates, 15 to 20 percent anthracite, and smaller amounts of limestone, slag, sand, and by-products from the refining steps.

Batch operations

Tin smelting is a batch operation in which one load is smelted and tapped before the next is added. The furnaces are operated at 1,200° to 1,300° C (2,200° to 2,400° F); the time for each load is 10 to 12 hours. The batch is stirred at intervals. It is customary to limit the amount of reducing agent in the primary furnace and to permit a considerable amount of tin to flow out as slag. This practice produces metallic tin low in iron.

The completed batch is tapped into a settler, from which the slag usually overflows into cast-iron pots. The molten tin is cast from the bottom of the settler into slabs of about 34 kilograms (75 pounds) and is ready for further refining.

The slag, containing 10 to 25 percent tin, is resmelted in the same kind of furnace used for smelting the ore. But a higher temperature, up to 1,480° C (2,700° F), is required because the reduction of the tin silicate in the slag is more difficult than the reduction of the tin oxide in the ore. Fine anthracite coal serves as the reducing agent; limestone flux and metallic iron may be added to aid in the reduction. The aim is to produce a discardable slag containing no more than 2 percent tin. During the

resmelting considerable iron–tin compound is formed and returned to the system.

Electric smelting. Electric furnace smelting was first used in 1934 at Geomines in Shaba (formerly Katanga), now in Zaire. The modern three-phase Héroult electric-arc furnaces, rated at 1,000 kilovolt-amperes, are cylindrical shells lined first with insulating brick and then with firebrick. The graphite electrodes are 40 centimetres (15 inches) in diameter.

Refining and recovery. The crude tin produced by smelting must be refined by further heat or by an electrolytic process. Heat treatment includes two steps: liquation and boiling.

Liquation and boiling

In liquation, crude tin placed on the sloping hearth of a small reverberatory furnace is heated to just above its melting point. The fluid tin runs into a boiling kettle, while metals whose melting points are sufficiently higher than that of tin remain on the hearth as dross. Iron is largely removed, and arsenic, antimony, and copper are partly removed; but lead and bismuth, also with low melting points, flow with the tin into the kettle, where the final refining (boiling) is carried out.

The molten tin in the boiling kettle is agitated with steam, compressed air, or poles of green wood, which produce steam. The remaining impurities form a light scum, which is removed and recirculated through the smelting cycle. The pure tin is cast in iron molds, usually in the form of 45-kilogram (100-pound) ingots. The purity of grade-A tin is guaranteed to exceed 99.8 percent tin.

Electrolytic refining. Electrolytic refining produces a purer grade of tin than that obtained by fuel-fired heat refining; but costs are higher and the demand for pure grades (99.99+ percent) is low. One modern electrorefining facility has developed an economical method by processing low-grade concentrates. First it upgrades them by roasting, leaching, and fire refining, to give a crude tin of approximately 93 percent tin, 3 percent antimony, 2 percent lead, 1 percent arsenic, and 0.5 percent copper. The final electrolytic process produces a metal in the range of 99.97 percent tin.

Recovery from scrap. Several thousand tons of tin are recovered each year from metal scrap, mostly in the form of tin-bearing alloys and clean tinplate clippings. In the U.S. secondary tin metal recovered from scrap may average about 3,360 tons annually, while nearly 20,000 tons of tin are recovered in the form of tin alloys and tin compounds. Other industrial countries have similar programs.

Detinning tinplate scrap

In salvage operations, often called detinning, clean tinplate scrap is placed in perforated steel-plate drums that revolve in a vat of hot caustic soda solution, containing an oxidant such as sodium nitrate or nitrite. Steam coils maintain a temperature of 96° to 99° C (205° to 210° F). The tin dissolves, forming sodium stannate in solution, which can be treated in three ways: (1) evaporation of the solution to crystallize the sodium stannate; (2) electrolysis to recover tin metal; or (3) acidification by carbon dioxide, sulfuric acid, or acidic gases to precipitate tin hydrate that can be further processed to tin oxide, smelted to tin metal, or converted to tin chemicals. About 784,000 tons of tinplate scrap are treated in this way annually in the United States.

Tin-bearing alloy scrap is handled by secondary smelters as part of the production of primary metals and alloys. Lead refineries recover metals from solder and tin dross, Babbitt metal, and type metal. Copper refineries rework scrap tin, bronze, and tin-containing brass.

Tin metal and its alloys. Smelters control purity of tin between 99.8 and 99.98 percent; about 90 percent of all tin produced is Grade A, which has a minimum purity of 99.8 percent. Ingots of 12, 25, and 45 kilograms (28, 56, and 100 pounds) are purchased from metal brokers by steel mills, copper and lead refiners, solder and bearing manufacturers, and metal and alloy dealers. Smaller consumers may purchase individual ingots of pure or alloyed tin from metal dealers; or they may purchase fabricated products such as sheet, tube, or powder. The primary industrial use of tin is for tinplate and tin coatings on iron, copper, and other metals; other uses are as alloys, chemical compounds, and pure tin products.

Pure metal. As a first step in the production of tin products, ingots often must be remelted. Tin, because of its low melting point, can be remelted in cast-iron or steel pots, which resist corrosion by molten tin. The tin may be poured, ladled, or pumped into open molds or hinged molds of cast iron, sand, plaster, rubber, or other materials. Shrinkage on solidification is about 3 percent. Cast bars or slabs are readily rolled into sheet and foil, extruded into tubes and pipe, or drawn into wire. Tin is pliable during cold-working and does not harden when worked.

Tin pipe

Molten tin is cast into a cylindrical tubing mold that forms the chamber of an extrusion machine. After the metal solidifies, a ram descends into the chamber and extrudes the metal through a die. The most common diameter sizes for tubes are from about 0.63 centimetres (0.25 inch) up to two centimetres (0.75 inch).

Tin piping is used for conveying distilled water and for draft-beer tubing from barrels to counter taps. Ancillary equipment for these operations requires tin-lined brass valves and fittings. In lining valves, the tin metal is cast into the fluxed, cleaned body, and the openings are machined out. Brass pipe can be tin-lined by inserting tin tubing into the pipe and drawing a steel ball through the tube to expand the tin against the wall in a tight bond.

Cylindrical billets with a diameter of 5.1 to 7.6 centimetres (two to three inches) are extruded as wire from a press at a diameter of about 0.48 centimetres (0.19 inch). This wire can be reduced on normal wiredrawing machines to as fine as 0.2 millimetre (0.008 inch). Tin wire is used for electrical fuses and other special purposes.

Collapsible tubes

Collapsible tubes of tin, in which shaving cream, toothpaste, and other pastelike materials are packaged, are formed by impact extrusion. Cone-shaped slugs or disks of tin containing about 0.2 percent copper as a hardener are oiled and fed into the die of the press, which is capable of exerting hundreds of tons of pressure per square inch. The ram of the press descends on the slug, squeezing tin out through the narrow annulus, or ringlike opening, of the die and forming a tube enveloping the ram. The tin-clad ram then returns upward out of the die, swings forward, and is simultaneously cleared of the tin by a blast of compressed air. The tubes are hand-stacked vertically over close-fitting pegs on a conveyor belt. The nozzles may be formed with their screw threads in this same operation. The tubes are printed and dried while still on the pegs.

Collapsible tubes are light; a body of 11.4 centimetres (4.5 inches) in length and 2.54 centimetres (one inch) in diameter weighs only 12 grams (0.026 pound). Wall thickness ranges from 0.13 to 0.25 millimetre (0.005 to 0.01 inch).

Tin-coated lead tubes, cheaper than all-tin, can also be impact extruded from slugs. A slab of lead is placed between two sheets of tinfoil of a thickness proportional to that desired on the wall of the tubes. Two roll passes weld the tin to the lead. The extrusion process is the same as for all-tin tubes.

Tin sheet and foil

In producing sheet and foil, tin is cast into slabs in either open or book molds. The slabs are rolled soon after casting because the grain size can enlarge on standing. The first pass through the rollers reduces thickness to one-half. The next pass, which sometimes is made across the direction of the first rolling, reduces the thickness by one-half again. Successive reductions to sheet or foil are about nine-tenths on each pass. On the way to the reelers, the foil passes over high-speed polishing rollers. A smear of pure olive oil serves as a lubricant in the early stages.

Tin anodes for electroplating processes can be cast, extruded, or rolled from billets to any shape or size. Connector hooks are attached by drilling and tapping. Large, 4,080-kilogram (9,000-pound) anodes for tinplate production are cast in iron molds.

Tin powder is produced by spraying molten tin from nozzles under slight pressure in a vertical tower in which an inert atmosphere is maintained. The tower is sufficiently high to allow the atomized powder to solidify and cool in its downward path. The fine powder is graded and classified by screening.

Clear, flat glass is produced by allowing molten glass to solidify while it floats on a pool of molten tin, avoiding the costly polishing and grinding required for plate glass.

Tinplate. Tinplate is a low-carbon steel sheet thinly coated with pure tin. The strength and formability of steel are combined with the corrosion resistance, solderability, and appearance of tin. Tinplate is ductile and can be drawn and folded into complicated shapes. The tin surface can be enamelled, lacquered, or painted. Tinplate is used for food cans because tin is nontoxic and resists corrosive attack by air and by food acids. Oxygen or dry air combines with tin to form an invisible and protective oxide film on the surface that thickens as the temperature is raised. In vacuum packing (in the absence of oxygen), tin resists a wide range of acids by retaining a film of hydrogen.

Electrolytic tinning

A coating of tin can be deposited on most metals in high-speed electroplating processes. Tin alloy coatings can be achieved by including salts of other metals such as copper, lead, zinc, cadmium, and nickel in the electrolyte. The world's first commercial electrolytic tinplate line was installed at the United States Steel Corp. plant in Gary, Indiana, in 1937.

In the halogen tin-plating process, using the halogen compounds stannous chloride and alkali fluorides with addition agents as the electrolyte, the sheet travels at 450 to 610 metres (1,500 to 2,000 feet) per minute horizontally instead of following vertical paths in individual plating vats or cells. One side of the sheet is plated on one level; the direction is then reversed, and the other side is plated on an upper level. An alkaline process is based on an alkaline solution of sodium stannate and sodium hydroxide that functions as the electrolyte. The plating speed (150 metres [500 feet] per minute) is lower than that in the other processes. The Ferrostan process uses a stannous sulfate bath containing phenol sulfonic acid and addition agents.

The cold-reduced steel is handled as a continuous band through all operations, including electrolytic cleaning, treatment with dilute acid (light pickling), electrodeposition of the tin coating, rinsing and drying, brightening of the dull matte coating by induction heating, chromate treatment, rinsing and drying, oiling with a thin film of dioctyl sebacate, and coiling. Shipped coils weigh up to 9,000 kilograms (20,000 pounds). Coils may be diverted to shears that cut the finished tinplate into sheets 51 × 36 centimetres (20 × 14 inches).

Electrolytic tinplate is produced in five grades, with tin-coating thickness ranging from 0.00038 millimetre (0.000015 inch) on both sides of 0.11-kilogram (0.25-pound) tinplate to 0.0015 millimetre (0.00006 inch) on both sides of 0.45-kilogram (one-pound) tinplate. Differential tinplate with a heavier coating on the surface that is to form the inner side of a container in which corrosive foods are to be packed is also produced.

Double-reduced tinplate, introduced in the 1960s, involves a base sheet that has been cold-rolled a second time to reduce its thickness to about 0.15 millimetre (0.006 inch) and bring out the inherent strength of the steel.

Hot-dip tinning

Molten tin wets and adheres to most metals and is highly fluid, draining rapidly from an article before solidification and leaving a thin, even coating that is bright, lustrous, and nontoxic. The coating protects the underlying metal from oxidation and chemical attack from weak acids, alkalies, and electrolytes; the coating also has lubricating properties and aids in forming and drawing operations; and the tin coating makes possible the joining of metals that are difficult to solder. The hot-dip process, though it is used to make less than 1 percent of the total tinplate produced in the U.S., is important in other countries.

A modern hot-tinning unit comprises a thermostatically controlled vessel of molten tin held at about 316° C (600° F). A submerged mechanism of rollers and guides conducts clean, cold-reduced steel sheets downward through a layer of zinc chloride flux into the tin and then upward and out of the tin through a thick covering layer of palm oil. The sheets are subjected to the squeegeeing action of immersed tinned-steel rollers that regulate coating thickness. The tinned sheets are cooled by an air blast and then cleaned and polished, usually by a combination of washing with alkaline detergent and dry cleaning.

Hot-dip tinplate is used primarily for packaging corrosive foods such as acid fruits. It has a minimum tin-coating thickness of 0.0004 kilogram per square metre (0.006 pound per square foot). Special grades with coating thickness up to 0.0009 kilograms of tin per square metre (0.014 pound per square foot) are produced for noncontainer uses such as gas meters, light engineering parts, kitchen utensils, and electrical equipment.

Other tin-coated articles

Fabricated articles made of cast iron, steel, copper, and copper alloys are given a smooth, bright tin coating by immersion in liquid tin. Coatings also may be electroplated, using either acid or alkaline electrolytes. The dull-plated tin can be brightened in hot oil, but proprietary bright-tinning processes are in commercial use. The coating thickness on fabricated articles is usually in the range of 0.0025 to 0.05 millimetre (0.0001 to 0.002 inch), depending on requirements and corrosion protection desired.

For tin-coated copper vessels and cast-iron bearing shells, liquid tin is poured onto the surface and wiped smooth with a cloth pad. Copper tubing and aluminum pistons can be protected from corrosion with a thin tin coating applied by immersion in a solution of tin salts.

Alloys and alloy-coated products. Tin is easily alloyed. Copper, antimony, bismuth, cadmium, and silver when alloyed to tin increase its hardness, mechanical strength, corrosion resistance, and workability. High-tin alloys are noted for their conformability, low coefficient of friction, retention of oil films on the surface, and good casting properties. When tin is a minor constituent in alloys, it can provide specific or unusual effects.

Solder

After tinplate, the largest use of tin is in solders for joining metals at temperatures below 427° C (800° F); tin also is used in filler metals (materials for filling in holes and dents in automobile bodies) that melt well below that temperature. Tin–lead alloy solders constitute the largest portion of all solders in use. Alloys ranging from 2 to 70 percent tin, with the balance lead, and melting between 183° and 312° C (361° and 594° F) are used for joining copper, brass, nickel–silver, iron, tinplate, and other coated metals. For special purposes, tin–antimony, tin–antimony–lead, tin–silver, tin–zinc, and lead–tin–silver alloys are used. Special low-melting alloys (fusible alloys) melting below 150° C (300° F) contain bismuth, cadmium, tin, lead, and indium.

Because of the large differences in the specific gravities of tin and lead (7.28 and 11.34, respectively), mechanical stirring is necessary throughout melting and casting to keep the lead from sinking to the bottom. Molten solder may be poured into separate molds to solidify into ingots or may be cast continuously by feeding the molten metal into a water-cooled graphite die at such a rate that billets are produced without interruption. A travelling saw divides the issuing billet into convenient lengths. Billets may be remelted to form solder strip that can be rolled to solder foil and then cut to any shape. Short billets may be extruded through a die, coiled, and drawn to form wire solder in sizes ranging in diameter from 3.25 millimetre (0.128 inch [10 standard wire gauge]) to 7.1 millimetre (0.028 inch [22 standard wire gauge]).

Hollow wire or tubes of various shapes, diameters, and configurations are manufactured containing a flux core that is introduced as a warm liquid during drawing and solidifies when cold. The flux can be noncorrosive or corrosive.

Solder also is made in powder form as paste solder, an alloy suspended in a mixture of emulsifying material and flux. It is particularly useful in mechanized soldering, in which premeasured quantities of solder and flux are applied automatically.

Fusible alloys

Alloys that melt at temperatures below the melting point of tin are formed mainly with metals of similarly low melting points, such as bismuth, cadmium, lead, and indium. They are termed fusible alloys because they melt rapidly when the melting temperature is reached. Two groups of alloys stand out. The first, containing 15.5 percent tin, 52.5 percent bismuth, and 32 percent lead, melts at 96° C (205° F); the second, containing 13.1 percent tin, 49.5 percent bismuth, 27.3 percent lead, and 10.1 percent cadmium, melts at 70° C (158° F).

Contraction of alloys during solidification results in castings of dimensions smaller than those of the mold. But

Table 7: Composition and Uses of Tin Alloys

trade name or type	composition (percent)											uses
	Cu	Sn	Pb	Sb	Ni	Al	Si	Ti	Zr	Cr	Fe	
Tin base and lead base												
Babbitt metal (grade 1)	4.5	91		4.5								turbines, centrifugal pumps
Babbitt metal (grade 2)	3.5	89		7.5								automotive bearings
Babbitt metal (grade 3)	8	84		8								
Babbitt metal (grade 4)	3	75	10	12								
Babbitt metal (grade 5)	2	65	18	15								bearings
Jeweler's white metal		92		8								costume jewelry
Pewter (casting)	0.5	92		7.5								plates, domestic utensils
Pewter (spinning)	2	91		7								tea and coffee service, trays, tankards
Pewter (soft)	2	95		3								
Britannia metal	1–3	90–95		4–8								castings, ornamental objects
Linotype		3–5	balance	11–12								type metal
Stereotype		6–8	balance	13–15								type metal
Monotype		7–8	balance	15–17								type metal
Foundry type		10–13	balance	20–25								type metal
Aluminum base												
Alcoa 80S	1	6.5			0.5	balance	1.5					solid bearings subject to high loads
Aluminum–Tin	1	20				balance						automotive bearings-connecting rod, main camshaft
Miscellaneous												
Alpha titanium		2.5				5		balance				structural material for aircraft and spacecraft
Zircaloy-2		1.5			0.05				balance	0.10	0.12	nuclear power reactor fuel cladding
Superconducting Nb_3Sn		33.4										solenoid magnets
Dental amalgam*		12										fillings

*Plus mercury, silver balance.

alloys containing bismuth and antimony may expand as well as contract during solidification, depending on the proportion of the metals present. Fusible alloys that are dimensionally stable with the mold are suitable for making foundry patterns and dies. Fusible alloys play a key role in safety devices that operate when the surrounding temperature reaches their melting points—*e.g.*, fire alarms, water-sprinkler devices, and fire-door closers. Other uses include: low-temperature soldering, to avoid damaging heat-sensitive materials; and tube bending, in which the alloy is poured into the tube, which is then bent and the alloy melted out.

Babbitt metal

Babbitt metal (tin and lead-base bearing metals or white-metal bearing alloys) serves as the lining for bearing shells of cast iron, steel, bronze, or gunmetal. Babbitt metal, which is white in colour and has very low friction, may be made of: (1) high-tin alloys that are substantially lead-free; (2) intermediate alloys in which some tin is replaced by lead; and (3) high-lead alloys. All are low-melting materials consisting of hard compounds in a soft matrix. The compounds found in each group are similar; it is in the composition and properties of the matrix that they differ. Typical compositions are given in Table 7. In compounding bearing metal, the tin is melted first; the copper and antimony strips are then dipped in tinning flux and lowered into the molten tin, where they dissolve below 400° C (752° F) without oxidation. Sometimes master alloys, 50 percent antimony and 10 percent copper, instead are immersed in the tin. When lead is added, the molten metals must be well stirred. Users of Babbitt metal usually buy ingots of the required composition and melt them in iron kettles from which they can ladle or pump the molten alloy.

The bearing shell or backing is precoated with tin for tin-base bearings or with lead–tin for lead-base bearings. While the coating is still molten, the white metal is cast onto the backing and allowed to solidify from the bond inward to prevent contraction cavities at the bond and to restrict the growth of intermetallic compound (alloy layer) at the interface between the bearing shell and the white metal alloy. Pouring temperatures are usually 50° C (122° F) above the point at which the metal is completely molten. The lining is machined to a mirror-bright finish and specified thickness.

High-tin alloy bearings are used for high unit loads and high operating temperatures. They hold oil films and readily adapt their shapes to conform to the shaft. Foreign matter not carried away by the lubricant is embedded below the surface and rendered harmless. Lead-base alloy bearings are prone to separate into elemental lead and tin and have a lower thermal conductivity than tin-base bearings.

Aluminum–tin bearing alloys

Aluminum–tin bearing alloys are of two types. The 6.5-percent-tin alloy, usually the material in sand or gravity die castings, is rarely bonded to a stronger backing. Its high fatigue strength enables it to carry higher fluctuating loads than either the tin-base or softer copper–lead alloys, but its ability to embed foreign matter is low. The finished bearings, usually electroplated with tin, serve as connecting rod and main bearings for tractors and heavy-duty gasoline and diesel engines.

The other type of aluminum–tin alloy (20 percent tin) is used for connecting rod, main, and camshaft bearings for passenger cars, especially in Europe. This type of bearing withstands high loads and fatigue at elevated temperatures. The aluminum–tin alloy in sheet form can be bonded directly to a soft steel backing. Both surfaces are scratch-brushed and rolled together in a single pass so that the total reduction is at least 45 percent. By annealing for one hour at 350° C (662° F), the tin is distributed throughout the aluminum matrix in discrete particles and provides a constant bearing surface.

Pewter

Pewter is an alloy of tin hardened with antimony and copper. Modern pewter compositions are given in Table 7. The best grades contain no lead.

Pewter played an important part in the development of civilization as a medium of artistic expression and as a material from which domestic utensils were fashioned. Pewter exceeds most other metals in ductility and ease of working. It may be stretched, compressed, and bent into almost any desired shape; and it does not harden sufficiently on cold-working with tools to make annealing necessary. Pewter is melted and cast like other tin-base alloys. Slabs for rolling can be cast in open or hinged molds, which may be of cast iron, plaster of paris, wood, papier-mâché, or rubber. Articles also can be cast in sections and assembled by soldering.

Pewter ingots are rolled to sheet, which is a starting

material for articles made by simple forming, spinning (forming while rapidly rotating), stretching, or hammering. Modern pewter can be given a mirror-bright finish or a soft satin sheen by polishing, and it retains its brightness indefinitely. It is used for tankards, domestic utensils, plates, trays, and tea and coffee services. Reproductions of objects from the 17th and 18th centuries are often made, some cast from the original molds.

Type metal compositions are shown in Table 7. Different kinds of typecasting and printing processes each require special alloy properties that are obtained by altering, within fairly narrow limits, the proportions of tin, antimony, and lead.

Tin as a stabilizer of cast iron

Alloyed cast iron, with a 0.1 percent tin content, is used in automotive engine blocks, transmissions, and heat-resistant applications. The tin acts as a stabilizer. The iron has a higher and more uniform hardness, better machinability, and better wear resistance with tin. The pure tin is added to the ladle in the form of shot or pieces of wire just before casting.

Alpha-type titanium alloys are particularly suitable for structural materials in aircraft and spacecraft because of their light weight, strength, weldability, and resistance to creep.

Zircaloy-2, an alloy of tin and zirconium, is technically feasible for use in all nuclear power reactors using pressurized water as a coolant and moderator. The low neutron absorption of zirconium makes it an attractive structural material and fuel cladding for nuclear power reactors, but it has low strength and highly variable corrosion behaviour. The addition of 1.5 percent tin overcomes these disadvantages.

Niobium–tin is important in the construction of high-field superconducting solenoid magnets. The alloys are excellent superconductive materials at low temperatures.

Tin-alloy-coated articles

Coatings of tin–lead alloys containing 63 percent or more tin can be applied to articles by hot dipping and by electrodeposition from a special electrolyte. The solderability of the coating is useful, particularly in the electronics industry, for parts made of copper, brass, or steel. Tin-alloy coatings on fabricated articles are generally harder, brighter, and more resistant to corrosion than pure tin coatings.

Tin–copper coated articles (12 percent tin) have the appearance of 24-karat gold. Tin–zinc coated articles (75 percent tin) are readily solderable and have excellent corrosion resistance. Tin–nickel plated articles (66 percent tin) have decorative and functional uses. The coating is used to resist etching of certain areas in the manufacture of printed circuits.

Terne metal

Terneplate is low-carbon steel sheet or strip coated with terne metal, a lead–tin alloy. Lead alone does not alloy with iron, but, when from 7 to 20 percent tin is added, the tin readily wets the steel and forms a solid solution with the lead. A composition of 80 to 85 percent lead and 20 to 15 percent tin is usually preferred. The clean, cold-reduced, mild-steel strip is coated by hot dipping, as in the tin-plating process, except at a higher temperature (370° C [700° F]). Its main uses are for roofing materials, automotive parts, casket linings, and packaging for paints and greases.

Tin compounds. Tin compounds have been known since 1605, when Andreas Libavius, German author of the first textbook on chemistry, described a method of preparing tin tetrachloride. This compound was widely used for weighting silk fabrics until the 1930s, when synthetic fibres supplanted silk. Large amounts of white tin oxide were produced to give opacity to pure white ceramic glazes on bathroom fixtures until cheaper substitutes, such as zirconium oxide, appeared.

In the early 1940s, production of inorganic tin chemicals surged with the invention of the continuous electrolytic tin-plating processes for can stock. In the same period, the organotin stabilizer was invented. Organotin compounds, in which a tin atom carries one, two, three, or four alkyl or aryl groups, are prepared from tin tetraalkyl or tetraaryl compounds. When about 2 percent of an organotin compound is added during the heat processing of clear polyvinyl chloride polymers, decomposition and darkening are prevented, and the polyvinyl chloride is stabilized against degradation by sunlight.

In the 1950s a Dutch chemist, G.J.M. van der Kerk, discovered that trialkyltin and triaryltin compounds had biocidal and fungicidal properties; he concluded that toxicity results when the tin atom combines directly with three carbon atoms and reaches a maximum when the total number of carbon atoms in the molecule is about 12. Hundreds of tons of tin compounds have come into use as industrial fungicides in wood preservation, papermaking, and textile manufacture; as disinfectants in hospitals; as antifoulants for paint; and as agricultural sprays for root crops, fruit, and nut trees.

About 60 percent of tin chemicals are by-products of the process of detinning tinplate scrap. The United States, Great Britain, West Germany, The Netherlands, and Japan are the principal producers of tin chemicals.

Inorganic compounds. About 15 inorganic tin compounds of industrial importance are produced from either pure tin or basic tin compounds—sodium stannate and hydrated tin oxide—obtained from detinning tinplate scrap. Sodium stannate and hydrated tin oxide react with hydrochloric acid to form stannous chloride, an important intermediate in making other tin compounds.

Uses of stannous chloride

Stannous chloride is made by dissolving granular tin in hydrochloric acid. When the solution is evaporated, crystals of hydrated stannous chloride are formed. The hydrate melts and decomposes at 38° C (100° F). Stannous chloride is a constituent of tin-plating electrolytes, the tin–nickel-alloy-plating electrolytes, and immersion-tinning processes. It is used as an activator for plating plastics and for silvering mirrors, an additive to drilling muds, an antisludge agent for oils, a stabilizer of perfume in soaps, and a catalyst in organic reactions. Careful neutralization of a solution of stannous chloride precipitates stannous oxide, a starting material for other tin compounds.

When stannous oxide is dissolved in fluoboric acid, stannous fluoborate is produced. This product, a salt, is sold commercially as a 47-percent solution for use in preparing the tin fluoborate electrolyte used in tin-plating.

Stannous fluoride is obtained by dissolving stannous oxide in hydrofluoric acid. The chief use of this salt is in toothpaste formulations for preventing demineralization of teeth.

Stannous oxide is prepared by precipitating the hydrate from a solution of stannous chloride with alkali carbonate. Its principal use is as an intermediate in preparing other stannous chemicals, such as stannous fluoride. It also is essential to the manufacture of gold–tin and copper–tin ruby glass.

Stannous chloride and sodium pyrophosphate combine to form the salt stannous pyrophosphate, which acts as a polishing agent in fluoride toothpastes and plays a part in electroplating with copper–tin alloy (bronze).

There are two methods of producing stannous sulfate: (1) from stannous oxide and sulfuric acid, or (2) by reacting granulated tin with an excess of sulfuric acid at 100° C (212° F) and evaporating in a vacuum. Stannous sulfate is the principal ingredient in the Ferrostan electroplating solution. It has found a large use in "liquor finishing," in which a copper–tin coating is applied to steel wire by immersion in a solution of stannous sulfate and copper sulfate. The coating acts as a lubricant in drawing the wire to thinner gauges and protects the wire from oxidation in storage.

The reaction of tin with bromine produces stannic bromide, which is thermally stable and can be purified by distillation at a boiling point of 202° C (396° F). Because of its low melting point (31° C [88° F]) and the high specific gravity (3.34 at 35° C [95° F]) of the compound it is used in heavy-media separation of minerals in metallurgical processes.

Stannic chloride

The commercial preparation of stannic chloride is by direct chlorination of tin under pressure. The chlorine must be dry and the temperature in the vessel not over 38° C (100° F). The liquid is purified by distillation and can be shipped in steel drums of special design to keep out moisture.

Stannic chloride (with pyrophosphates) acts as a replace-

ment for the gum extracted during the processing of silk (weighting silk). It acts as a colour fixative in the dyeing of silk. It has been widely used in the preparation of other inorganic tin compounds and organotin compounds. When sprayed on glass and fired, it forms a conductive coating and strengthens the glass. Minor uses are in stabilizing the perfume in soaps, as an additive to diesel oils and transformer oils, and in making blueprint paper.

Stannic oxide

In the industrial production of stannic oxide, tin metal is atomized and the finely divided metal burned. It is also prepared by calcining (heating to a high temperature) the hydrated oxide obtained by the hydrolysis (by boiling or by treatment with weak acid) of sodium stannate.

The soft white colour of fired porcelain enamel coatings opacified with stannic oxide is considered most desirable for the highest quality finish on bathtubs and other vitreous products. Stannic oxide is also used in making coloured ceramic glazes such as chrome–tin pink and vanadium–tin yellow. Substantial quantities of stannic oxide are consumed as a putty powder for polishing marble and granite.

Potassium stannate is prepared by dissolving tin in potassium hydroxide or by decomposing tin hydrate in potassium hydroxide. Potassium stannate is an ingredient in high-speed, high-conductivity tin-plating and tin-alloy-plating baths. It is the basis of a process for tin-plating aluminum pistons by immersion.

The commercial preparation of sodium stannate is accomplished by detinning scrap tinplate in a hot caustic soda solution containing oxidizers. The stannate is recovered by purifying the solution by evaporation and crystallization. Sodium stannate formed the basis of the first successful alkaline-tin electroplating bath. The plating speed and current-carrying capacity are less than those of the potassium stannate bath, but the process is cheaper and popular in job-plating shops.

Ceramic dielectric bodies are made from alkaline earth and heavy metal stannates, particularly stannates of barium, cadmium, copper, lead, and nickel. A soluble salt of the metal is reacted with sodium or potassium stannate, and the insoluble compounds are recovered.

Organotin compounds. Stabilizers, biocides and fungicides of varying toxicity, and catalysts all are industrially important organic tin (organotin) compounds.

Grignard and Wurtz processes

The commercial production of stabilizers and biocidal compounds depends on either the Grignard or the modified Wurtz processes. In the Grignard process, a tin tetrahalide reacts with an excess of alkylmagnesium halide to produce tetraalkyltins, the starting materials for many organotin compounds. Under the modified Wurtz method, dibutyltin dichloride, butyl chloride, and sodium react to yield tetrabutyltin. Tetrabutyltin can be converted to dibutyltin dichloride by treatment with stannic chloride and part of the product recycled.

Organotin compounds can also be prepared by reacting aluminum alkyls with stannic chloride or by direct synthesis in which tin metal reacts with alkyl halides in the presence of suitable catalysts at elevated temperatures to produce dialkyltin halides.

A number of different organotin compounds are useful as chemical stabilizers, to prevent various types of deterioration. One of these is dibutyltin dilaurate, prepared from the dichloride, which is hydrolyzed with caustic soda to form dibutyltin oxide. The oxide is then treated with lauric acid to yield dibutyltin dilaurate. This was one of the first compounds used as a stabilizer for polyvinyl chloride plastics. It also is added to poultry feed to control worms in chickens.

Biocidal and fungicidal compounds

The toxic properties of many tin compounds make them useful as insecticides and fungicides. Some of the more important of these compounds are as follows:

1. Bis-tributyltin oxide manufactured by the caustic hydrolysis of tributyltin chloride and industrially available in a number of formulations for use in antifouling paints, sanitary formulations, and for the control of microorganisms in industrial-process water.
2. Tributyltin chloride, as a rodent repellent, especially in coatings for exposed electrical cables.
3. Triphenyltin hydroxide, for controlling potato blight, leaf spot on sugar beets and celery, and pecan blight; it is used also on cotton and other crops. The acetate derivative, obtained by reacting the hydroxide with acetic acid, also is effective.
4. The hydroxide of tricyclohexyltin chloride, as a control of plant-feeding mites; its low order of mammalian toxicity allows its use on apple, pear, and citrus trees, as well as on vegetables and some field crops.

Catalysts

Stannous octoate, a stannous soap, acts as a catalyst in curing silicone oil emulsions. When mixed with certain amines, it catalyzes production of polyurethane foam rubber. Proprietary formulations for stannous soaps derived from the reaction of stannous oxide and oleic and stearic acid also are used as catalysts in foam rubber production. Stannous oxalate is one type of catalyst utilized in the hydrogenation of coal. (R.McK.MacI./Ed.)

TITANIUM

Titanium became an important structural metal with the advent of the space age. Its primary uses are as components in the compressor sections of jet engines, and as structural and nonstructural components of airframes and space-vehicle systems. These applications exploit the high strength-to-weight ratios of titanium alloys and their retention of strength in moderately high temperatures. In addition, titanium and its alloys are used in chemical-process equipment in which the excellent corrosion resistance can be used to advantage.

Although the element titanium had been known since its initial discovery in 1791 by William Gregor in England, and its identification in rutile in 1795 and in ilmenite in 1797, it was not until William Justin Kroll, a German refugee in the U.S., invented a magnesium-reduction process in 1936 that the production of metallic titanium became feasible. The process, which Kroll gave to the U.S. Bureau of Mines for development, remains in its original form or modifications thereof, the principal method used today for winning titanium metal from its ores.

Ores and mining. Titanium ores are plentiful, and numerous deposits of highly concentrated titanium minerals are readily accessible and easily mined. Rutile in the beach sands of Florida and Australia is merely scooped up and shipped to a mill for processing. Ilmenite, such as occurs in many near-surface deposits in North America, is similarly mined with large, mechanical shovels and transported to the processing plant.

Rutile and ilmenite are the two important ores of titanium. Rutile is almost pure (98–99 percent) titanium dioxide (TiO_2), containing small amounts of iron, silicon, and other elements as impurities. Ilmenite is a combined iron (II)-titanium dioxide usually expressed as iron(II) titanate ($FeO\text{-}TiO_2$) containing about 32 percent titanium and 37 percent iron. Rutile is the preferred ore for the production of titanium because of its high titanium content and the ease with which titanium tetrachloride, the starting material for the production of metallic titanium, can be produced from it.

Most of the world's rutile supply comes from the beach sands of New South Wales and Queensland along the east coast of Australia. Other sources include Sierra Leone, Sri Lanka, the Soviet Union, and Brazil. Ilmenite is much more abundant than rutile, and large deposits are known to exist in South Africa, India, the United States, Canada, Norway, and Australia. In addition to rutile and ilmenite, many other minerals contain titanium but not in sufficient concentration to be of commercial importance.

Refining and recovery. The production of titanium metal from its ores is accomplished in several steps, each involving chemical reactions and each designed to eliminate the gaseous elements oxygen and nitrogen from the chemical reactions. Titanium dioxide is produced from rutile or ilmenite by dissolving the ore in a sulfuric acid solution and precipitating the iron compounds. The solution is then hydrolyzed, producing a hydrous titanium oxide, which is washed and calcined. Titanium has a great affinity for oxygen and nitrogen, and unlike other metals, it actually dissolves oxygen and nitrogen in both the liquid and the solid states in addition to reacting to form oxides and nitrides. To eliminate oxygen and nitrogen from the

titanium production processes, the oxide prepared from the ore is first converted to titanium tetrachloride ($TiCl_4$) by reacting the titanium dioxide with chlorine gas in the presence of a reducing agent, usually carbon.

The resultant titanium tetrachloride—a liquid when at room temperature—contains many impurities, principally other metal chlorides, and it must be purified before being reduced to titanium metal. A sequence of distillations at temperatures above or below the boiling point is used to remove these undesired compounds. The high-purity titanium tetrachloride produced is reduced to metallic titanium by reaction with magnesium or sodium. The Kroll process, which uses magnesium, and modifications that use sodium, as the reducing agents, are batch processes conducted in large reactors. An argon or helium atmosphere is used within the reactor because all air and moisture must be excluded from the reaction to prevent oxygen, nitrogen, or hydrogen contamination of the final product. Both processes produce titanium in the form of a highly porous material, termed titanium sponge, with magnesium chloride or sodium chloride entrapped in the pores. This sponge is crushed to a suitable size for handling and then leached with water and acid or heated in a vacuum to remove the entrapped chlorides. Titanium sponge is the raw material from which ingots of titanium and titanium alloys are melted. The sponge's purity is dependent on the purity of the titanium tetrachloride from which it was made, the purity of the magnesium or sodium reducing agent, and the purity of the atmosphere during the reduction process. Common impurities in sponge titanium are oxygen, nitrogen, carbon, hydrogen, and iron.

The Kroll process

The metal and its alloys. Titanium in pure form is a soft, ductile metal. It has a density of 4.51 grams per cubic centimetre (0.163 pound per cubic inch), which is about midway between aluminum (2.70 grams per cubic centimetre [0.098 pound per cubic inch]) and iron (7.86 grams per cubic centimetre [0.284 pound per cubic inch]). It has a silvery gray colour and can be polished to a high lustre. It is not, however, as lustrous as chromium or stainless steel. Titanium melts at 1,675° C (3,045° F). The common impurities, carbon, oxygen, and nitrogen, raise the melting point; most metallic impurities, such as iron, manganese, chromium, and copper, lower the melting point. There are two crystal structures of titanium. At temperatures below 882° C (1,620° F), titanium has a hexagonal close-packed structure; this form is known as the alpha phase. At temperatures between 882° C and its melting point, titanium has a body-centred cubic structure, known as the beta phase (see below). Titanium is paramagnetic and has very low electrical and thermal conductivity. It occurs in three valence forms, 2, 3, and 4, the most stable of which is 4. The most common oxide is titanium dioxide, TiO_2.

Like stainless steel and aluminum, titanium, because of the formation of an oxide surface film that is relatively inert chemically, has excellent corrosion resistance, in many environments. Titanium will resist corrosive attack by oxidizing acids such as nitric acid and aqua regia, organic acids, moist chlorine gas, chloride solutions, dilute solutions of sulfuric acid and hydrochloric acid, and dilute solutions of alkalies. Although at room temperatures titanium is resistant to oxidizing atmospheres, at elevated temperatures it reacts with the oxygen in the air. At temperatures as low as 260° C (620° F), the surface of titanium becomes tarnished, forming an oxide film that ranges in colour from blue to gold. At higher temperatures, a thicker, yellow-brown oxide film, or scale, is formed. The rate of oxidation increases as the temperature is increased, but it is not until temperatures are above about 648° C (1,200° F) that oxidation becomes a problem if long-time exposures are involved. At temperatures between 648° and 1,215° C (2,200° F), short-time heating can be carried out in air. Forging and fabrication of titanium alloys are done at these temperatures with no detrimental effects as long as the oxide scales are removed after fabrication. In the liquid state, however, titanium is very reactive and reduces all known refractories.

Corrosion resistance

Titanium will ignite and burn in dry chlorine gas and is attacked by hydrofluoric and phosphoric acids and moderate concentrations of the alkalies. Titanium is severely attacked by fuming red nitric acid. The corrosion resistance of titanium is outstanding in marine atmospheres and seawater. No noticeable corrosion of titanium has been found after exposure to seawater for over three years.

Perhaps the most important property of titanium is its ability to be alloyed with most of the other metals and many of the nonmetals. Alloying additions increase the strength of titanium both at room temperatures and at elevated temperatures. Pure titanium has a tensile strength of about 2,800 kilograms per square centimetre (40,000 pounds per square inch). By alloying it, tensile strengths as high as 14,000 kilograms per square centimetre (200,000 pounds per square inch) can be obtained, thus making it useful as a structural metal.

The major alloying elements that are added to titanium are aluminum, vanadium, molybdenum, manganese, iron, and chromium. All of the titanium-base alloys contain one or more of these elements. The gaseous elements—oxygen, nitrogen, and hydrogen—and carbon are present in small quantities in almost all titanium alloys.

Basic types of titanium alloys

Titanium alloys can be classed in three basic types: alpha, alpha-beta, and beta alloys. Alpha alloys contain elements that dissolve in the hexagonal close-packed alpha phase. Alpha-beta alloys contain limited quantities of elements that dissolve in the body-centred cubic beta phase resulting in a two-phase alpha-beta alloy. Beta alloys contain large quantities of elements that dissolve in the beta phase resulting in an all-beta structure. Alpha alloys have medium strengths (8,400 to 10,500 kilograms per square centimetre [120,000 to 150,000 pounds per square inch]) and good elevated-temperature strengths, can be welded, and are used mostly as forgings. Alpha-beta alloys are the most versatile. In the annealed condition, they have medium strengths but can be heat-treated to very high strengths. These alloys have good forming characteristics and can be used in both sheet and bar-stock form. Generally, they are not weldable. Beta alloys have medium strengths and excellent forming characteristics but contain large quantities of high-density alloying additions. They also can be heat-treated to high strength levels. Compositions and properties of typical alloys are listed in Table 8.

Table 8: Typical Alloys

type of alloy	composition (balance titanium)	tensile strength (psi)	elongation (percent)
Alpha	5% Al–2.5% Sn	130,000	15
Alpha-beta	6% Al–4% V	140,000	15
Alpha-beta (heat treated)	6% Al–4% V	180,000	7
Beta	13% V–11% Cr–3% Al	150,000	15
Beta (heat treated)	13% V–11% Cr–3% Al	200,000	6

Titanium has long been used as a deoxidizer in steel, to which it is added as a ferrotitanium or a ferrocarbontitanium alloy. More recently, it has been used as an alloying addition to many steels to reduce grain size and control carbon content in stainless steel. In aluminum it is used to refine the grain size, while in copper it produces a precipitation-hardening alloy. In high-temperature nickel–cobalt–chromium alloys, titanium is added to produce a precipitation-hardening reaction providing high strengths at temperatures up to 815° C (1,500° F). Titanium is used in permanent magnet alloys of the iron–cobalt–nickel–titanium type.

Chemical compounds. Titanium dioxide (TiO_2) is the most important compound of titanium. It is used extensively in the pigment industry because of its excellent hiding power or opacity. A pure white compound having high reflectivity, it is ideal for use in white paints, enamels, and lacquers and can be used in conjunction with other compounds in coloured paints.

Use in the pigment industry

It is also used in pigments for rubber, paper, oilcloth, leather, textiles, inks, and cosmetics. The two main sources for titanium dioxide are rutile and ilmenite. Since less pure titanium dioxide can be used in pigments than can be used in producing titanium metal, most of the titanium dioxide produced from ilmenite is used by the pigment industry. The preparation of titanium dioxide for the

pigment industry follows the same procedures as those already described.

Titanium tetrachloride, the raw material for producing titanium metal, is the next most important compound. It is produced by chlorinating titanium dioxide, or titanium ores, in the presence of carbon. The minor amounts of iron, silicon, oxygen, and other impurities remaining after chlorination are removed by fractional distillation. Titanium tetrachloride is used, along with titanium dioxide, as a catalyst in many chemical reactions and as a smoke-producing compound for skywriting or smoke screens. Titanium carbide, a compound produced by reacting carbon with a titanium compound, is used in conjunction with a tungsten carbide in cutting tools and dies. Thin coatings of pure titanium carbide (3,800 on the Vickers diamond hardness scale as compared to about 1,300 for tungsten carbide) are applied by vapour deposition on tools, dies, punches, and other components. Many other compounds of titanium exist, but their uses are limited.

(H.R.O./Ed.)

TUNGSTEN

Tungsten, in its various forms, is a widely used industrial material. It is particularly important as an alloying element in the production of high-speed steel and, as cemented tungsten carbide, in cutting tools and other applications requiring hardness. The pure metal retains its strength at high temperatures, and additions of tungsten to other metals may improve their high-temperature strength.

The word tungsten was first used in 1755 by Baron Axel Fredrick Cronstedt, who applied it to the mineral later known as scheelite because of its high density. Other researchers identified the ore as a compound of lime and a previously unknown substance and discovered the same mysterious substance, which they named "wolfram," in a mineral that contained iron and manganese. Today the metal is known as tungsten in British and American usage and as wolfram in European. The mineral is known as either wolfram or wolframite.

Tungsten was a rare metal until 1847 when industrial development was started under a British patent for the manufacture of sodium tungstate and tungstic acid. A later patent (1857) described the manufacture of iron–tungsten alloys, which were to form the basis of modern high-speed steels. Industrial development followed an exhibition of tungsten tool-steels by Bethlehem Steel Corporation at Paris in 1900. The first application of unalloyed tungsten was as filaments in electric incandescent lamps, following the production of ductile filaments by a powder metallurgical process patented in 1909, a process that remains the basis of modern production methods.

Ores and mining. Tungsten minerals are widely distributed. The commercially important ones are wolframite and scheelite. Wolframite is iron and manganese tungstate (Fe, Mn) WO_4, a mixed crystal of huebnerite, $MnWO_4$, and ferberite, $FeWO_4$. Scheelite is calcium tungstate, $CaWO_4$.

Every tungsten deposit must be examined critically to determine the most suitable and economic method of concentration. The method chosen will depend upon the type of mineral, associated minerals, and the identity of the gangue. Gravity methods of concentrating the relatively heavy tungsten minerals are used frequently, though most large-scale mills use a combination of flotation and gravity methods.

If calcite and apatite are the principal contaminants, acid leaching can be used to upgrade the concentrate. Low-grade concentrate is also converted to synthetic scheelite by chemical treatment. Roasting, magnetic separation, or electrostatic separation may also be used to improve the quality.

Molybdenum, tin, and bismuth may be present in the ore in sufficient quantities to warrant their recovery, a circumstance that can complicate the process.

The aim of most mills is to produce a concentrate containing a minimum of 60 percent tungsten trioxide, WO_3. A range of 60 to 70 percent is normal.

Metal recovery. The first step in the recovery of tungsten from wolframite is to fuse it with sodium carbonate under oxidizing conditions or by digestion with a concentrated sodium hydroxide solution. Scheelite is usually decomposed by digestion with concentrated hydrochloric acid.

Wolframite reaction

The reaction between wolframite and sodium carbonate may be carried out in a gas-fired reverberatory furnace at 800° C (1,470° F) with constant rabbling (stirring) or in a gas-fired rotary furnace at a slightly higher temperature sufficient to cause the charge to flow down the kiln. In the case of digestion with a sodium hydroxide solution, the reaction is best carried out in a steam-heated autoclave.

The resulting product is leached with water to produce sodium tungstate liquor, which can be treated with calcium chloride solution to produce insoluble calcium tungstate (synthetic scheelite).

Hydrochloric acid is added, while stirring, to a slurry of the calcium tungstate to produce insoluble tungstic acid (H_2WO_4) and soluble calcium chloride. The tungstic acid is recovered by filtering, washing, and drying.

The digestion of scheelite with concentrated hydrochloric acid yields a crude tungstic acid. After the calcium chloride is washed out, the dried acid contains 80 to 98 percent WO_3. Purification is achieved by dissolving the tungstic-acid sludge in concentrated ammonia solution, filtering and evaporating to produce ammonium paratungstate crystals, $5(NH_4)_2O \cdot 12WO_3 \cdot xH_2O$. Further purification is obtained by decomposing the crystals with hydrochloric acid and redissolving the precipitated tungstic acid in ammonia solution and crystallizing again as ammonium paratungstate.

Most tungsten powder is produced by hydrogen reduction of tungstic acid, tungstic oxide, or ammonium paratungstate.

Particle size. The particle size of the powder is important in determining the success of the subsequent operations, and it is necessary to produce different types of powder for various applications. Tungsten powder for the production of massive metal (sheet, rod, and wire) has a wide range of particle size to produce maximum compaction, whereas the particle size of powder for conversion to carbide for hard metal production should be tightly controlled around the size specified, which size will vary from less than one micrometre to several micrometres, depending on the application. The reduction procedure must be carefully controlled to produce the required type of powder.

Reduction. The basic reduction process is to pass metal boats charged with tungstic acid or oxide through a heated tube with a counterflow of hydrogen. The acid or oxide is reduced to metal, and water vapour is removed in the flow of hydrogen. The temperature of reduction varies from 700° to 1,100° C (1,290° to 2,010° F); the higher the temperature, the coarser the powder. For the finest powder it is necessary to use a shallow charge of oxide with a high flow of dry hydrogen and carry out reduction at a low temperature, preferably in two stages. For coarser metal with a wide range of particle size, deep charges are reduced at high temperatures.

Some tungsten powder is produced by carbon reduction of the oxide, but this powder is less pure than hydrogen-reduced metal and is used mainly for alloy production.

High-purity tungsten powder is produced by the hydrogen reduction of tungsten hexafluoride in a fluidized bed at 550° C (1,020° F).

Powder metallurgy

The metal and its alloys. Tungsten powder is converted to massive metal by a powder metallurgy process. The powder is pressed into a bar shape that is presintered in hydrogen to give it enough strength to be handled. The bar is then clamped vertically between electrical contacts in a hydrogen-filled chamber, and a current of several thousand amperes is passed directly through it. The temperature of the bar is raised to about 3,000° C (5,400° F) and held at that temperature for about 30 minutes.

The sintered bar (93–95 percent of the density of solid tungsten) is too brittle to be worked at ordinary temperatures; it can, however, be forged and rolled at about 1,500° C (2,700° F).

Production of rod and wire is accomplished by swaging (passing the metal while hammering through a die). The first pass is made at 1,500° C and the temperature grad-

ually lowered to about 1,200° C (2,200° F). As the metal is worked it becomes more dense and more ductile.

Plate and sheet are produced from sintered bar by forging at 1,500° C, followed by rolling first at 1,400° C (2,500° F) and later down to 700° C (1,300° F) as the ductility increases.

Small additions of nickel (about 1 percent) result in lower sintering temperatures, but the resulting product cannot be worked.

The difficulties in melting tungsten (melting point about 3,400° C [6,100° F]) have been overcome with the development of melting by means of an electric arc or by electron-beam melting techniques, but the resulting product in both cases is an ingot with coarse columnar grains unsuitable for working.

Tungsten components have been made by other methods than powder metallurgy; *e.g.,* by centrifugally casting arc-melted tungsten, usually with a little added molybdenum; by building up shapes on formers by plasma-spraying tungsten powder; and by slip casting followed by sintering.

Depositing tungsten in vapour form from tungsten hexafluoride is used to produce coatings of tungsten.

Hot isostatic compaction is a recently developed means of consolidating tungsten powder. The tungsten powder is sealed in a metal container and heated in a high-pressure, inert gas. Full-density tungsten is produced at pressures of 700 to 1,400 kilograms per square centimetre (10,000 to 20,000 pounds per square inch) at temperatures between 1,500° and 1,600° C (2,900° F).

Applications of tungsten metal. The applications are based on tungsten's high melting point, excellent high-temperature strength, and good heat and electrical conductivity. It is used for lamp filaments, electrical contacts, X-ray cathodes, heating elements, and equipment (including thermocouples) for high-temperature furnaces.

Chief industrial uses

Aerospace applications have been investigated, but the good high-temperature mechanical properties of tungsten are marred by a lack of resistance to oxidation. Alloying has failed to overcome this fault. Protective coatings have achieved some improvement, but none is completely satisfactory in providing resistance to oxidation, abrasion, alternate heat and cold, and impact. Tungsten nozzles have been used in some types of rocket motors. The material is usually in the form of porous sintered tungsten infiltrated with silver.

Important alloys and their uses. Stellite is the name of an important series of alloys containing cobalt (30 to 56 percent), chromium (30 percent) and tungsten (10 to 20 percent), and small additions of carbon. These alloys combine excellent hot-hardness with good corrosion and oxidation resistance at high temperatures. The alloys can be used as solid tool bits or hard-facing rods for deposition as abrasion-resisting coatings on ferrous metals. Cast shapes are also used for surgical implants.

Additions of rhenium (35 atomic percent; that is, 35 percent of the atoms in the alloy are rhenium) greatly improve low-temperature ductility of tungsten. Aerospace and nuclear applications are considered possible, but the high cost of rhenium is a deterrent. Tungsten containing 26 percent rhenium is used in thermocouples for temperatures up to 3,000° C (5,400° F).

Heavy metal, an alloy of tungsten (93 percent), copper (2 percent), and nickel (5 percent), is used because of its high density for machine-balancing weights and for containers for radioactive substances.

Composites of "alloys" of tungsten and copper or silver, produced by infiltrating porous sintered tungsten with copper or silver, are used for electrical contacts for switch gear.

Small amounts of thoria (about 1 percent) are incorporated in tungsten to improve its high-temperature properties, particularly its resistance to sagging.

Cutting steels. An important use for tungsten is in making high-speed cutting steels that retain their hardness at high temperatures. These steels contain 6 to 20 percent tungsten with molybdenum, chromium, vanadium, or niobium. Tungsten can be added to steel in various forms, such as ferrotungsten (about 78 percent tungsten), melting base alloy (about 35 percent tungsten), scrap that is high in tungsten, and high-purity scheelite (natural or synthetic). Metal powder, pellets, and certain proprietary compositions are also used.

Most ferrotungsten in produced by carbon reduction of the concentrate in an electric furnace. A metallothermic process uses aluminum and silicon reducing agents.

Chemical compounds—preparation and uses. The most important compound is tungsten carbide (WC). It is used in the production of cemented carbides (hard metal) because of its great hardness, which it retains at high temperatures.

The carbide is prepared by heating an intimate mixture of tungsten powder with the required amount of carbon, in a nonoxidizing atmosphere, at about 1,400° to 1,500° C. The product is crushed and milled to a fine powder and intimately mixed with the requisite amount of fine cobalt powder by prolonged ball-milling.

The normal procedure is to press the mixed powder to the shape required. Usually a lubricant such as paraffin wax is incorporated in the powder to aid pressing and to reduce wear on the dies. The wax is removed by heat treatment before sintering at about 1,400° C. Originally, sintering was performed in a hydrogen atmosphere; but vacuum sintering, particularly when titanium and tantalum carbides are present, has largely replaced this technique. Large pieces of cemented carbide are produced by hot-pressing to shape in carbon dies.

The main use for cemented carbides is as cutting tools. The tool is brazed or mechanically fixed to a steel shank. Additions of titanium and tantalum carbides improve the cutting qualities. The application of a thin deposit of titanium carbide on the surface of the cemented carbide tools has, it is claimed, produced a very large increase in the cutting life.

The cobalt content of cemented carbides usually varies between 6 and 15 percent, the strength increasing and the hardness decreasing as the cobalt increases.

Other important applications for cemented carbides include drawing dies, rock drills, wear-resisting parts for machinery, automobile tire studs for use on icy roads, and deposits on other metals to reduce wear.

When tungsten is melted in a carbon crucible with an addition of carbon, the alloy $WC–W_2C$ is produced. It may be cast—*e.g.,* in the form of sleeves—and used as sand-blast nozzles, or crushed and incorporated in hard-facing welding rods.

(G.L.M.)

URANIUM

Uranium—a dense, hard, radioactive nickel-white metal with the heaviest atomic weight of any naturally occurring element—was known for 150 years before any serious application could be found for it. The discovery in 1938 that the uranium nucleus could be made to fission, or split, when bombarded with neutrons suddenly opened the dramatic possibility that the uranium atom could be made to give up its energy in a sustained chain reaction. About 0.5 kilogram (one pound) of uranium—in size, about 16 cubic centimetres (one cubic inch)—can, in such a reaction, yield as much energy as 1,360,000 kilograms (3,000,000 pounds) of coal. As a consequence, intense study of the element was launched together with extensive exploration for its ores. The first nonlaboratory application of the 1938 discovery was military, in the form of the original atomic bomb dropped on Hiroshima in 1945. After World War II, exploitation of nuclear energy as a power source and for other peaceful purposes was developed in the United States, Great Britain, the U.S.S.R., France, and other countries. By the 1970s uranium had become a metal of vast economic significance.

Nuclear fuel

The discovery of uranium is usually attributed to Martin Heinrich Klaproth, a Berlin apothecary and an outstanding analytical chemist, who described it in a lecture to the Royal Prussian Academy of Science in 1789. Mineralogists had tended to classify the mineral pitchblende, found in Joachimsthal (now Jáchymov, Czechoslovakia) and in the Erzgebirge (Germany), as an ore of zinc. Klaproth processed the mineral to produce specks of a black substance, which he termed uranit, a new metal, but which in

fact must still have been an oxide or mixture of oxides of uranium. A year later Klaproth changed the name from uranit to uranium, after the planet Uranus, which had been discovered in 1781.

First extraction

A French chemist, Eugène-Melchior Peligot, was the first to succeed in extracting the metal (1841). He passed chlorine over heated "uranium," studied the products of the reaction, and found that the "uranium" he used was an oxide. Peligot then tried to reduce uranium tetrachloride with potassium in a heated platinum crucible. The potassium chloride dissolved away, leaving a black powder that was truly the element uranium.

Ores and mining. Uranium is estimated to be present in the Earth's crust to the extent of about four parts per 1,000,000. It is thus considerably more abundant than gold and silver, ranking closer to tungsten and tantalum. Its presence in seawater amounts to 0.000001 to 0.000002 gram per litre. It is highly reactive and forms many compounds, some of which are water soluble.

Compounds of uranium are present in rocks of varying composition and origin. The actual concentration of uranium in most of these ores is low, however, and ore bodies suitable for economic mining are rare. Local concentrations of suitable ores may have occurred when uranium-rich compounds were exuded into rock faults after the molten uranium silicates in the Earth's crust had crystallized. Later the rock weathered, leaving deposits of the more chemically unreactive uranium minerals; other uranium minerals may have been taken into aqueous solution, transported, and subsequently precipitated as oxides after reacting with other materials.

Major ore locations. Significant deposits of uranium ore are located in Africa, in Zaire, and in the Witwatersrand region of South Africa; in France and Czechoslovakia; in Australia; in South America; in Canada, in the Great Bear Lake area and the province of Ontario; and in the western U.S. The Soviet Union and China are also believed to hold substantial deposits.

Primary, relatively unoxidized uranium minerals of particular importance are uraninite (a crystalline uranium form that is 45 to 85 percent uranium) and its black friable or massive variety, pitchblende; coffinite (hydrated uranium silicate; black, friable; 60 percent uranium); brannerite (uranium, thorium, calcium titanate; black; 30 to 40 percent uranium); davidite (rare earth, uranium, iron titanate; black; 7 to 10 percent uranium). Important secondary, oxidized uranium minerals are carnotite, autunite, and torbernite. Carnotite is a complex compound of uranium, potassium, vanadium, and oxygen; autunite is a calcium uranyl phosphate; and torbernite is a copper uranyl phosphate.

Mining techniques. Ore deposits may be deep or shallow, regular or irregular, and may be found in association with a variety of other minerals. Most of the customary metal mining techniques are used, many quite technically advanced because of the relatively recent commercial interest in uranium.

The methods of mining uranium at the surface are similar to those for mining other minerals. The Lucky Mc Mine in Wyoming has produced ore at the rate of about 400,000 dry tons per year and in the process has had to remove at least 10,321,000 cubic metres (13,500,000 cubic yards) of waste. Earth-removing vehicles with capacities of 1.1–1.9 cubic metres (1.5–2.5 cubic yards) follow the narrow, winding lodes, mining out 0.6-metre- (two-foot-) thick slices of ore, while scrapers remove the surrounding waste. Pit design is dictated largely by the stability of the rock; at the Lucky Mc, the slopes of the pit walls average 50 to 57 degrees with bench heights (the distance from one horizontal surface to another) of 21 to 27 metres (70 to 90 feet) and widths of six to 12 metres (20 to 40 feet). Keeping the wall angle as steep as possible reduces the total volume of waste material, or overburden, that must be removed.

To set up a surface-mining program, engineers must determine the pit limits and ore reserves, as well as uranium content of the ore. Radiometric surveys to determine the extent and amount of the ore's radioactivity both prior to and during working may provide the first two sets of data; rapid spectrographic analysis measures content. This information must then be coupled with other relevant data in scheduling production. Computer systems are widely used in this process.

Techniques of underground mining vary somewhat with the nature of the ore body. Deposits that lie at depths of 300 to 1,200 metres (1,000 to 4,000 feet), in thicknesses of up to 30 metres (100 feet), and in strata sloping at angles of up to 45 degrees are extracted by the usual underground mining methods. Basically, the ore is released from the main body of rock by blasting and then transported to collection points for transmission to the surface through vertical shafts.

Control of ore grade quality is important and, again, radiometric assessment methods are widely used underground. Ventilation is of particular importance in underground uranium mining because of the presence of radon gas, one of the products of the radioactive disintegration of uranium. Radon also is radioactive and must be kept at a strictly regulated low level, to protect miners from overexposure to radiation.

Problems of underground operations

In South Africa, where uranium is a by-product of gold-ore processing, mining practice is determined by gold-mining requirements.

Refining and recovery. The chemical inertness of uranium minerals makes recovery difficult, but because of their relatively high density and hardness, mechanical enrichment of the ore is possible (Figure 31). Such enrichment, or concentration, is necessary, since the bulk of uranium in most regions comes from ores that contain fewer than three kilograms (six pounds) of uranium per ton of ore; the uranium mineral is finely distributed throughout a quartz-containing matrix.

Mechanical enrichment

Mechanical separation. The mined ore is first coarse crushed, sometimes separated on the basis of its radioactivity, and then fine-ground and mixed with water to produce a slurry. The fine grinding exposes the uranium mineral for subsequent chemical attack. Mechanical concentration of the ore slurry is accomplished in a series

Drawing by D. Meighan

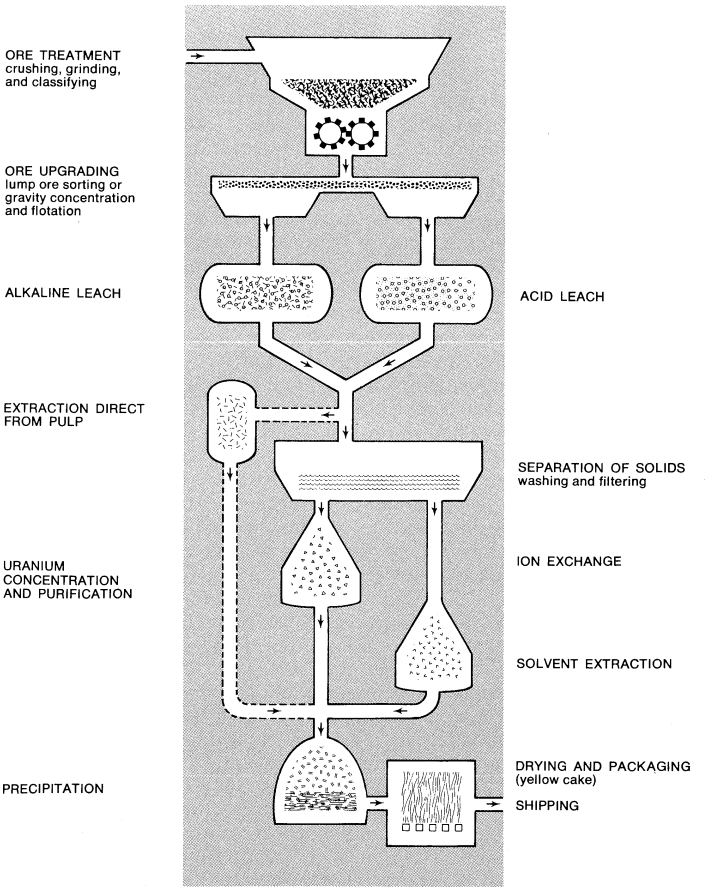


Figure 31: Steps involved in the recovery of uranium.

of steps. First, the water is drained off; then the use of thickeners or filters, or both, creates a product suitable for leaching (dissolving out the uranium). Either acid or alkali agents are used in such leaching processes.

Leaching and purifying. In the most common leaching process dilute sulfuric acid is added to the pulverized ore in a pachuca, a vessel in which the mixture is agitated with compressed air. Mechanical stirring may also be used, along with high pressures and temperatures. Such leaching produces a rather crude solution of uranyl sulfate. Some ores, particularly those rich in carbonates and alkaline earths that would be expensive to acid-leach, can be alkali-leached by sodium carbonate, yielding sodium uranyl carbonate. The alkali-leach solution is generally more concentrated than that for the acid process, but, because the chemical reaction is specific in nature, finer initial grinding is necessary to ensure efficient breakdown of the uranium mineral in the alkali method.

After leaching, the crude solution is separated from the ore pulp by filtration or settling, with the raw mineral solution gradually concentrated as it moves along the processing line; the solution is then purified by various chemical techniques. Precipitation from the purified solution, separation, and drying constitute the final stages of the extraction of uranium concentrates. The solution is neutralized with caustic soda, ammonia, or magnesia. By controlling the acidity, the iron is precipitated out first and then the uranium, as a uranate, usually sodium diuranate, the principal component of yellow cake. After thickening, the uranium precipitate passes through automatic filter presses or vacuum filter drums and is allowed to dry. This crude uranium compound concentrate is in the form of a bright, claylike material. It contains about 60 percent by weight of uranium together with a variety of impurities.

Preparation for nuclear uses. The nuclear uses of uranium demand high purity in general, particularly a high degree of freedom from certain rare earth metals, as well as from cadmium and boron, all of which have the ability to absorb neutrons. For the final purification stage uranyl nitrate solution is extracted in a process employing a concentrated solution of tributyl phosphate in kerosene or hexane. First the crude concentrate is dissolved in nitric acid, using either batch or continuous processes, yielding a solution of uranyl nitrate and silicic acid, which is separated out. The remaining uranyl nitrate solution containing about 200 grams of uranium per litre is purified by solvent extraction. The solvent is further purified to recover the uranium as an aqueous solution of uranyl nitrate of a purity suitable for nuclear purposes.

Final purification

Uranium and its compounds. Preparation of the oxides of uranium from the pure uranyl nitrate solution constitutes the next important processing stage. As these oxides are used in solid form to produce other uranium compounds, their physical state (particularly the size and distribution of particles) is of major importance and must be closely controlled.

Extraction processes. Extraction of the uranium oxide from the uranyl nitrate solution is accomplished by heating the nitrate in air. When heated below 300° C (570° F), the solution is converted to uranium trioxide; above 430° C (810° F), the oxide of uranium U_3O_8 is formed. Or uranium may be precipitated from the solution as bright yellow ammonium diuranate using ammonia or urea, or as hydrated uranium peroxide using hydrogen peroxide. These compounds are then thermally decomposed to give uranium trioxide, which is reduced to the dioxide by using hydrogen or cracked ammonia at 600° C (1,100° F) and above.

Uranium tetrafluoride, important in making uranium metal, is produced by reaction of uranium dioxide with anhydrous hydrofluoric acid at 450°–600° C (840°–1,100° F). The higher the temperature the more rapid the reaction rate, but a practical upper limit is imposed by the availability of structural materials and equipment able to withstand attack by hydrofluoric acid and the ability to control the reaction, which produces considerable heat. In all these refining methods the trend has been away from batch to continuous processes, which generally result in a more uniform product.

Preparation of uranium metal. Of the many processes that may be used for the preparation of uranium metal, the one in almost universal use employs uranium tetrafluoride with magnesium as the reducing agent. The compound is mixed with magnesium chips, compacted, loaded into a refractory container, and placed inside an electric resistance furnace in an argon atmosphere. The mixture is then heated; the reaction generates sufficient heat to melt both the resulting uranium and the magnesium fluoride. Because of the large difference between their specific gravities, the molten magnesium fluoride floats up and forms a slag on the uranium surface. Uranium ingots weighing about 1,360 kilograms (3,000 pounds) have been produced by this method.

Uranium oxides can also be reduced to a metal using calcium and magnesium in a similar manner, but of the reaction products only uranium is melted by the heat of reaction and is thus produced as a powder in a lime or magnesia matrix.

Uranium hexafluoride used in the gaseous diffusion enrichment process (see below) can be prepared by fluorinating the metal, an oxide, uranium tetrafluoride, or other compounds with fluorine, hydrogen fluoride, or halogen fluorides. The least amount of fluorine is needed for the fluorination of uranium tetrafluoride. This reaction, the basis of methods for industrial preparation of uranium hexafluoride, occurs at high temperatures (above 400° C [750° F]). Fluorination of the tetrafluoride with chlorine trifluoride, which requires lower temperatures (50°–150° C [120°–300° F]), is also used to produce uranium hexafluoride on an industrial scale. The hexafluoride may be handled as gas, solid, or liquid. Because this process, like much of the industrial chemistry of uranium compounds, entails major usage of highly corrosive fluorine and fluorine compounds, highly sophisticated chemical engineering techniques, standards, and equipment are required.

Applications of uranium and its compounds. Uranium in elemental or compound form fuels nuclear reactors, principally for the generation of electricity. In practice, when the uranium is used in reactors, it is alloyed with minor amounts of aluminum and iron, which serve to minimize dimensional changes resulting from fission. Uranium alloyed with molybdenum also is used as fuel in some reactors.

Reactor fuel for power production

Many different compounds of uranium have been explored as nuclear fuels. Uranium dioxide has so far proved the most economically attractive fuel for reactors in the production of electricity. The fuel consists of pellets of uranium dioxide of about 95 percent theoretical density packed into a can of stainless steel or zirconium alloy. The dioxide has a low thermal conductivity and tends to fragment in the can under the thermal stresses imposed in service. The can protects the fissionable material from contamination from coolants, and prevents contamination of the coolant and hence the remainder of the reactor by fission products from the uranium.

Uranium monocarbide, in contrast with uranium dioxide, is more metallic in nature, more chemically reactive, has a much higher thermal conductivity, and is better able to withstand thermal stresses. Its behaviour in a nuclear reactor is acceptable, but since its chemical reactivity is more akin to that of uranium metal, the choice of coolants for use with it is limited largely to molten sodium. Uranium carbide, therefore, is likely to find major application in sodium-cooled fast reactors.

Natural uranium contains only 0.7 percent of the fissionable isotope, uranium-235. In many reactor applications, this percentage must be increased, a process called enrichment. This is accomplished by a gaseous diffusion technique in which the lighter isotope uranium-235 diffuses through a membrane more rapidly than the heavier uranium-238. With a sufficient number of diffusion stages, enrichment can be carried out to any desired degree. Uranium hexafluoride, being gaseous at moderate temperatures and pressures, is the main feed material for this process. The by-product from such enrichment is uranium with most of its uranium-235 removed (primarily the isotope uranium-238), called depleted uranium.

(J.Wi./Ed.)

PETROLEUM AND NATURAL GAS

Petroleum is a naturally occurring hydrocarbon material of animal and vegetable origin. First formed in deep sedimentary beds, petroleum later migrated to its present position in porous underground rocks. Petroleum occurs as either a liquid, a gas, or a combination of both. Occasional deposits, elevated during the formation of mountain ranges, have been exposed by erosion to form tar (sometimes called pitch) deposits, and some of these have been known and used throughout recorded history. Other near-surface deposits of liquid petroleum seep slowly to the surface through natural fractures and fissures in the overlying rock, often aided by erosion or weathering. Accumulations from these oil seeps, called rock oil, were used commercially in the early 19th century. The vast majority of petroleum deposits, however, lie in natural rock at depths from 150 to 7,600 metres (500 to 25,000 feet) below the surface of the ground. The petroleum deposits, or reservoirs, are prevented from migrating upward by intervening layers of dense rock. As a general rule, the deeper deposits have higher internal pressures and contain greater quantities of the gaseous hydrocarbons.

In order to create a petroleum trap or reservoir, a porous rock (one capable of containing liquids), such as sandstone or limestone, must be surrounded on the top and sides by nonporous rock, such as shale. In such an event, either gaseous or liquid petroleum, which is less dense than the always-present water, may be trapped in the top portion of the porous formation.

Since petroleum reservoirs are the result of an accumulation of liquid or gas in rocks originally containing water, there will always be some water left in the reservoir. Water production often accompanies the production of oil and gas.

When it was discovered in the 19th century that rock oil could yield a distilled product (kerosene) suitable for lanterns, sources of this oil were eagerly sought. It is now generally agreed that the first well drilled specifically to find petroleum was that of one Col. Edwin Drake in Titusville, Pennsylvania, in 1859. The success of this well, drilled close to an oil seep, led to further drilling in the same vicinity and soon to exploration elsewhere. The growing demand for petroleum products led quickly to oil wells not only in the United States but also in other countries. By the last quarter of the 20th century, there were approximately 600,000 producing oil wells in more than 100 different countries. Oil is produced in every continent except Antarctica.

First drilled well

Because of the widespread distribution of petroleum reservoirs, it was possible for new discoveries to be made at a rapid rate through the 1950s. As one result, there was little incentive for the development of efficient production techniques. Because little use had been found for the natural gas produced along with the crude oil, the gas often was burned at the wells. Such wasteful practices were gradually halted, partly by the increasing difficulty of finding new petroleum reservoirs and partly by the increased uses found for natural gas. New engineering methods led to an economically optimum recovery from each reservoir. These methods often were based on the injection of some less expensive fluid, such as water, to force additional petroleum into the producing wells.

As the value of the petroleum produced reached impressive proportions, governmental agencies in every producing area sought to prevent wasteful exploitation.

Prospecting and exploration

Drake's original well was drilled close to a known surface seepage of crude oil, and for many years such seepages were the only reliable indicators of the presence of oil. As demand grew, however, new methods were devised for exploring the petroleum potential of underground rock formations. There are now three major exploration methods: (1) surface feature mapping, (2) seismograph observations, and (3) Earth gravity surveys. But there is still no accurate way to predict the actual presence of underground petroleum deposits. Exploration methods carried out on the surface can only indicate the presence or absence of underground formations that are favourable for petroleum accumulation.

Surface methods. A seepage of crude oil sometimes appears as a tarlike deposit in a low area such as the oil springs at Baku, on the Caspian Sea, described by Marco Polo. More often, it can be seen as a thin skim of oil on small creeks that pass through the seepage area. This latter phenomenon was responsible for the naming of Oil Creek (a place in Pennsylvania), where Drake's well was drilled. Such seepages of natural gas cannot usually be seen, though instruments can detect natural gas concentrations in air down to less than one part in 100,000. Exploration methods have been used in which air samples from areas immediately over or under the ground surface are tested. Such samples taken from a wide area can yield a map of gas concentration points. A similar method has been used to test for traces of petroleum gases dissolved in seawater in the hope that undersea seepages could be thus detected. These geochemical prospecting methods are not applicable to the large majority of underground reservoirs that do not have leakage to the surface.

Oil Creek

A second class of surface prospecting methods is based on surface indications of underground rock formations. In some cases, subsurface folds and faults in rock formations are repeated in the surface features. The presence of underground salt domes may be indicated by a low bulge in an otherwise flat ground surface. Since these features usually produce possible petroleum traps, drilling in the area may be productive.

Seismographic methods. Most folds and faults in subsurface rocks are not apparent at the ground surface. Seismic methods that have been useful for mapping such underground formations involve the transmission and echo-producing properties of rock in a manner similar to underwater sonar methods. Sound waves travel at differing velocities depending on the types of rock formations and are reflected at interfaces between different types of rock. The sound wave source is usually a small explosion in a shallow drilled hole. Microphones are placed at various distances and in various directions from the source to record the arrival times of the echoes. This procedure is repeated at intervals over a large area. An experienced seismologist can then interpret the records to map the underground formation contours. Possible locations for petroleum accumulations may then be located for drilling.

Gravity meter surveys. The gravity meter has become a primary exploration tool. Although gravity at the Earth's surface is very nearly a constant, it is slightly greater where large rock formations lie closest to the surface. The gravity field, therefore, increases over the tops of anticlinal (arch-shaped) folds and other similar subsurface anomalies. A salt dome results in lowered gravity because salt is less dense than other rocks. Gravity surveys are usually carried out from the air. A sensitive instrument, the gravity meter, is suspended on a cable below the aircraft, while the recording devices are operated from inside. As the aircraft is flown over an area on a precise grid pattern, a map showing the gravity strengths is prepared from the recordings. These maps can show the presence of such possible accumulation spots as anticlinal folds, fault blocks, and salt domes, even though there are no surface indications of such. The gravity meter survey is faster and less expensive than the seismic method. It can locate only large rock features that exist relatively close to the surface, however, and is generally used as a preliminary survey tool to locate areas in which seismic work may profitably be concentrated.

Drilling and completing wells

Early wells were drilled with impact-type tools in a method called cable-tool drilling, in which a chisel-like bit was

Cable-tool drilling

suspended from a cable to a lever at the surface, and an up-and-down motion of the lever caused the bit to pound the bottom of the hole and chip away the rock. Such drilling had to be interrupted periodically to allow loose rock chips to be removed with a basket-like device on the cable. Also, the well had to be free of liquids during the drilling process so that the bit could effectively remove rock. These necessarily dry holes allowed oil and gas to flow to the surface in large quantities as soon as a producing formation was penetrated. This created the image of a "gusher" as a successful well. Often, a large amount of oil was wasted before the well could be capped and brought under control.

ROTARY DRILLING

During the middle and late 20th century, rotary drilling has become the preferred method for penetrating underground formations (Figure 32). In this method, a special tool—the drill bit—rotates while bearing down on the bottom of the well, thus gouging and chipping its way downward. Probably the greatest advantage of rotary drilling over cable tooling is that the well bore is kept full of liquid during drilling. A weighted fluid (called drilling mud) in the bore hole serves two important purposes: by its hydrostatic pressure, it prevents the entry of formation fluids into the well, thus preventing "blowouts" and "gushers." In addition, the drilling mud carries the crushed rock to the surface, so that the drilling is continuous until the bit wears out.

By courtesy of the American Petroleum Institute

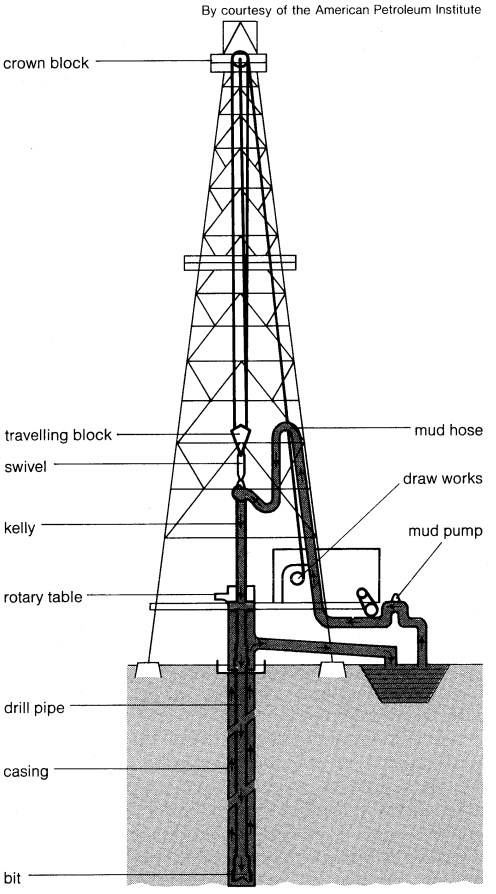


Figure 32: Rotary drilling rig.

The drill bit is connected to the surface equipment through a drill pipe, a heavy-walled tubing through which the drilling mud is fed to the bottom of the bore hole. In most cases, the drill pipe also transmits the rotary motion from a turntable at the surface to the bit at the bottom of the hole. The top piece of the drill pipe is a tube of square or octagonal cross section called the kelly, which passes through a square or octagonal hole in the turntable. At the bottom end of the drill pipe are extra-heavy sections called drill collars that serve to concentrate the weight on the rotating bit. The drilling mud leaves the drill pipe in such a way that it washes the loose rock from the bottom and carries it to the surface. Drilling mud is carefully formulated to assure correct weight and viscosity characteristics for its required tasks. After screening to remove the rock chips, the returning mud is held in open pits to be recirculated through the well. The mud is picked up by piston pumps and forced through a swivel joint into the top of the drill pipe.

The characteristic tall derrick contains the hoisting equipment that is used to raise and lower the drill pipe into the well. The drill bit wears quickly and requires frequent replacement, making it necessary to pull the entire drill string from the well and stack it at one side of the derrick. Joints of drill pipe are usually nine metres (30 feet) long. Sections of two or three such joints are separated and stacked vertically while the bit is being changed. Drilling mud is left in the bore hole during this time to prevent excessive flow of fluids into the well.

Such rotary drilling techniques as those described have made it possible for wells to be drilled to depths of more than 7,600 metres (25,000 feet). Wells have been drilled in both coastal waters and in the open sea. In relatively shallow waters (less than 60 metres [200 feet]), the wells are usually drilled from platforms that rest on the bottom; the drilling platform is constructed above the highest expected waves. In deeper waters and for exploratory work, drilling is done from a floating vessel that is kept in place by a combination of anchors and motors. Since the fixed platforms are quite expensive, many (20 or more) wells are often drilled from the same platform. Yet for most efficient coverage, the wells must enter the producing formation with more than 300 metres (1,000 feet) between them. Thus, each well except one has to be drilled in a precise slantwise direction. This directional drilling is achieved with a tool called a whipstock, which causes the drill bit to veer slightly in the desired direction. The whipstock is used at intervals until the total desired angular deviation is reached. Sharp bends in the bore hole must be avoided because the rather rigid drill pipe has to bend through the deviation while rotating.

Use of a floating vessel

One variation in rotary drilling has involved introducing a fluid-powered turbine at the bottom of the bore hole to produce the rotary motion of the bit. The turbodrill is about nine metres long and is made up of four basic parts: the upper bearing, the turbine, the lower bearing, and the drill bit. The upper bearing is attached to a drill pipe, which does not rotate or rotates at a much slower rate than the drill bit (six to eight revolutions per minute), but it is used to regulate the weight of the drill bit and to carry the drilling mud that powers the turbine. The drill bit rotates at a much faster rate (500 to 1,000 revolutions per minute) than conventional rotary drilling. The power source for turbodrilling is the mud pump, which pumps drilling mud through the drill pipe and upper bearing and into the turbine. The mud is diverted onto the rotors of the turbine, turning the lower bearing and the drill bit. As in conventional rotary drilling, the mud passes through a hollow portion of the lower bearing to cool the drill bit. Despite the widespread use of turbodrills in the Soviet Union, they are comparatively rare in the United States, where the rotary tool drilling method has remained predominant.

WELL LOGGING AND DRILL-STEM TESTING

After the bore hole has penetrated the possibly productive formations, these formations must be tested to determine if expensive completion procedures should be used. The first evaluation is usually made by well-logging methods, in which the logging tool is lowered past the formations while response signals are relayed to operators on the surface. Often these tools make use of the difference in electrical conductivities of rock, water, and petroleum to detect possible oil or gas accumulations. Other logging tools depend on difference in absorption of atomic particles. With well-logging tools, the exact depths of possibly productive formations are determined. Only a production test, however, can establish the potential productivity of a well.

Drill-stem test

The production test normally used is the so-called drill-stem test, in which the bore hole is sealed above and below the formation to be tested, with only the drill pipe open to the formation. The drill pipe is then emptied of the drilling mud so that the formation fluid may enter. When enough time has passed, the openings into the drill pipe are closed, and the pipe containing the formation fluid is brought to the surface for analysis. If there is gas present, the gas may flow from the top of the drill pipe during the test.

WELL COMPLETION

Casing. If the preliminary tests show that one or more of the formations in the bore hole will be commercially productive, the well must be prepared for the continuous production of the oil or gas. First, a large outside pipe, or casing, slightly smaller in diameter than the drill hole, is inserted to the full depth of the well. A cement slurry is forced between the outside of the casing and the inside surface of the drill hole. When set, this cement forms a seal so that fluids cannot pass from one portion of the well to the other through the bore hole. The casing is usually about 23 centimetres (nine inches) in diameter. It creates a permanent well through which the productive formations may be reached. After the casing is in place, a production string of smaller (about eight centimetres [three inches] in diameter) tubing is extended from the surface to the productive formation with a packing device to seal the productive interval from the rest of the well. If multiple productive formations are found, as many as four production strings of tubing may be hung in the same cased well. If a pump is needed to lift oil to the surface, it is placed on the bottom of the production string.

Since the casing is sealed against the productive formation, openings must be made to allow the oil or gas to enter the well. A down-hole perforator uses an explosive to shoot holes through the casing and cement and into the formation. The perforator tool is lowered through the tubing on a wire line. When it is in the correct position, the charges are fired electrically from the surface. Such perforating will be sufficient if the formation is quite productive. If not, an inert fluid may be injected into the formation at pressures high enough to fracture the rock around the well and thus open more flow passages for the petroleum. In early times, nitroglycerin was exploded in the well bore for the same purpose. An acid that dissolves rock may be used in place of the inert fluid.

Christmas tree valves. When the subsurface equipment is in place, a network of valves (called a Christmas tree) is installed on the surface, arranged so that flow from the well may be regulated and measured and so that tools to perform subsurface work may be introduced through the tubing. Christmas trees may be very simple, such as those found on a low-pressure well that must be pumped, or they may be very complex, as in the case of a high-pressure well with multiple producing strings.

Auxiliary apparatus. *Artificial lift.* Many oil wells do not have a formation pressure high enough to push the oil (which weighs about 0.7 kilogram per litre [six pounds per gallon]) to the surface. In these cases some artificial means for lifting the oil to the surface must be installed. The most common installation involves a motor and "walking beam" (like a seesaw) on the surface that operates the pump on the bottom of the production string. A chain of solid metal rods connects the beam and the pump. Another method, called gas lift, uses the buoyancy of a bubble of gas in the tubing to push the oil to the surface.

Gas lift

A third type of artificial lift forces some of the produced oil down the well at high pressure to operate a pump at the bottom of the well. Even though an oil field may have enough pressure to produce naturally at first, artificial lift will usually be required in the later stages of production. Gas wells that produce little or no liquid do not need artificial lift devices.

Surface equipment. Water often flows into the well along with oil and gas. Surface equipment separates the water from the petroleum and separates the oil and the gas for storage and transportation. The production from a number of wells also must be gathered at a central point for distribution. A settling tank is usually used to separate the water, although either heat or chemical treatment may be necessary to break down emulsions. Gas is separated from the oil by allowing the produced stream to pass through a series of pressure-reducing valves and removing the liberated gas to storage. The remaining crude oil is kept at normal pressures in cylindrical tanks after being brought through pipelines from the individual wells. Gas is usually piped directly to a central processing plant (called a gasoline plant), where remaining liquids are removed before it is fed to the consumer pipeline.

Recovery of oil and gas

As indicated above, both oil and gas reservoirs usually start with a formation pressure that forces fluid into the well and sometimes to the surface. With artificial lift reservoir fluids may be produced as long as enough reservoir pressure remains to cause the fluids to flow into the well. Since a decline in reservoir pressure invariably accompanies fluid production, a point is reached at which commercial quantities of oil or gas no longer flow into the well. In natural gas reservoirs this does not usually occur until most of the gas is recovered. In nearly all crude-oil fields, however, less than one-third of the oil originally present can be produced using only the reservoir pressure.

SECONDARY RECOVERY OF OIL

When a large part of the crude oil cannot be recovered simply by allowing the original reservoir pressure to furnish the driving energy, a method for supplying extra energy must be found. Usually, an injected fluid such as water or gas replaces the produced oil and thus maintains the reservoir pressure. When this procedure substantially increases the potential production of oil, it is called secondary recovery. When gas alone is used as the injected fluid, it is usually put into the top of the reservoir, where gas normally collects to form a gas cap. The effective use of this type of secondary recovery is restricted to reservoirs that have a high flow potential so that gravity allows the oil to drain to the lower part of the formation, where it may be produced. In such a situation gas injection can be a very effective recovery method. When gravity segregation of the gas and oil cannot occur, however, or when little vertical relief exists in the formation, only a small increase over the natural pressure decline production is realized.

Another even more widely used secondary recovery method is water flooding. Water, after treatment to remove certain impurities, is injected through some of the wells. In moving through the formation, it pushes oil toward the remaining production wells. The wells to be used for water injection are usually spaced evenly among the producing wells so that the oil has the shortest possible paths to follow. Water is an effective injection fluid and is not dependent upon gravity segregation. A water-flooding project often increases the oil recovery to at least twice that obtained by pressure decline methods. The major exception is oil reservoirs that are connected to an active subsurface water source and are thus subjected to water flooding without injection.

Water flooding

Although water flooding greatly increases the recovery of oil from a particular reservoir, it leaves one-fourth to one-third of the oil behind, and several other methods have been tried in efforts to achieve even better oil recovery. One such method uses a quantity of fluid, such as alcohol or liquefied carbon dioxide, ahead of the injected water to decrease the quantity of oil bypassed by the water. On occasion the injected water is made more viscous by the addition of jellying agents, which help to increase the reservoir volume invaded by water. Still another method requires that a small amount of liquid hydrocarbon, such as propane, be placed ahead of a gas injection to help push the oil ahead of the gas.

Several recent methods for secondary recovery involve the application of heat to the reservoir to make the oil more flowable. These methods are best suited for reservoirs that contain viscous crude oil. In one method (in situ combustion) a portion of the oil in the formation is burned by igniting the oil around the well and then forc-

ing air through the formation to push the burning region out into the reservoir. The heat thus generated makes the oil more flowable. In a successful project about 10 times as much oil is recovered as is burned. In another heat method, steam is injected into the reservoir; the heat liberated by the condensation decreases the viscosity of the crude oil. The heated oil is often produced from the same well into which the steam was injected. Several cycles of such injection and production may be used on a single well.

GAS CYCLING

Liquified petroleum gases

Natural gas reservoirs often contain appreciable quantities of liquid hydrocarbons held in the gaseous state. Since these liquids are liquefied petroleum gases (LPG) such as propane and butane, they are important sources of both fuel and chemical feedstocks. If the pressure in such a reservoir is allowed to decline during gas production, these liquids will condense in the reservoir and become nonrecoverable. To prevent this decline in pressure, a quantity of the produced gas from which the liquids have been removed is put back into the reservoir. This process, called gas cycling, is continued until the optimum quantity of liquids has been removed. The reservoir pressure is then allowed to decline during the production of the nearly dry gas. In effect, this process defers the use of the reinjected natural gas until after the liquids have been produced.

Overall, the production of oil and gas has become an exacting engineering job. Although new oil fields are being discovered, the huge consumption of petroleum requires that fields already in production be used more efficiently.

(B.H.C./Ed.)

Petroleum refining

GENERAL CONSIDERATIONS

In most cases, crude oil is unsuitable for direct use. It is converted into such useful products as gasoline, motor oil, and petrochemicals through the refining process.

Petroleum refining began in the United States and Russia in the second half of the 19th century, following the discovery in 1859 of rock oil in Pennsylvania. In the earliest refineries simple stills separated crude oil into impure gasoline, kerosene, lubricating oil, and fuel oil fractions. Kerosene was the principal marketable product. To improve its odour and appearance the kerosene was treated chemically with caustic soda or sulfuric acid.

The earliest automobile fuel was composed of those fractions of crude oil that were too light to be included in kerosene. Before the invention of the automobile this fraction had been virtually unsalable; as demand for it rose, refiners increased production and improved quality. Methods for the continuous distillation of crude oil were introduced.

After World War I a major improvement in refining came with development of the cracking process, consisting of heating surplus heavier oils under pressure and thereby cracking, or splitting, their large molecules into the smaller ones that form the lighter, more valuable fractions. After appropriate chemical treatment, gasoline manufactured by cracking performed better in automobile engines than gasoline derived from straight distillation.

During the 1930s and World War II, sophisticated refining processes involving the use of catalysts led to further improvements in the quality of fuels and increased their supply. These improved processes, including catalytic cracking, polymerization, alkylation, and isomerization (see below), permitted the petroleum industry to meet the high-performance demands of combat aircraft and, after the war, to supply the increasing quantities required by commercial aviation.

Catalytic reforming

The 1950s and 1960s brought a large-scale demand for jet fuel and for high-quality lubricating oils. Catalytic reforming was established as the leading process for upgrading automotive motor gasolines for use in the higher compression engines. Hydrocracking, accomplished by the addition of hydrogen during refining, also improved the crude-oil fractions.

Types of crude oils. The various hydrocarbon compounds that are mixed together in crude oil have different boiling points, but, apart from the lightest, the differences between the boiling points of neighbouring members in the rising scale of molecular weight are so small (only fractions of a degree) that they cannot be separated by ordinary distillation. Fortunately, separating is not usually necessary; most common petroleum products consist of mixtures of compounds whose boiling points fall within a specified range.

Typical boiling ranges are: motor gasoline 25°–150° C (75°–300° F), kerosene 150°–230° C (300°–450° F), diesel oils 230°–340° C (450°–650° F), and fuel oils above 340° C. In practice, however, there are many different grades of each product with different boiling-point specifications.

Crude oils vary widely in boiling range and composition. Figure 33 shows the proportions of the products obtained by distillation of six typical crude oils, ranging from West

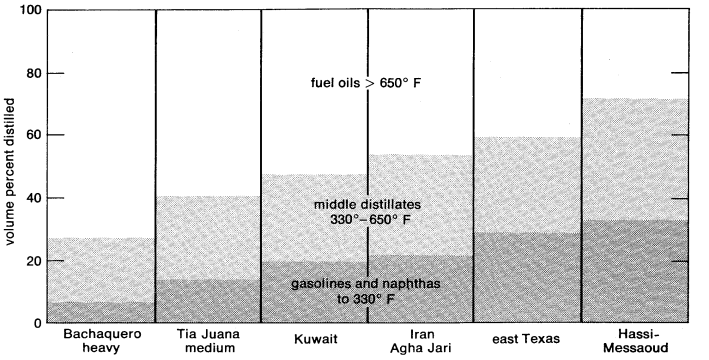


Figure 33: Proportions of the products obtained by distillation of six crude oils.

Venezuelan Bachaquero Heavy crude to the light Algerian crude Hassi-Messaoud. The percentage of lighter products (gasoline and naphtha) increases to the right of the chart. Middle distillates (which include kerosene and gas oil) also increase, while fuel-oil content diminishes.

Composition of crude oil. Four main types of hydrocarbons are present in crude oil: normal paraffins, isoparaffins, cycloparaffins (also called naphthenes), and aromatics. Some crude oils, such as Pennsylvanian, consist mainly of paraffins. Others, such as the heavy Mexican and Venezuelan crudes, are predominantly naphthenic and are rich in asphalt (a high boiling semisolid material). Wax is usually, but not always, associated with paraffin-base crudes. The U.S. Bureau of Mines classifies crude oils as paraffin base, naphthene base, or mixed base by determining the properties of key fractions distilled from the oil.

In addition to the hydrocarbons, compounds of sulfur, nitrogen, and oxygen are present in small amounts in crude oil. These are harmful unless removed. Also, there are usually traces of vanadium, nickel, chlorine, and arsenic.

SEPARATION INTO COMPONENTS

The primary refinery process is fractional distillation, which may be followed by other physical separation methods, such as solvent extraction, in which superior lubricating stocks are extracted by means of a solvent, usually furfural obtained from oat hulls.

Fractional distillation. Modern petroleum distillation units operate continuously over long periods of time and are large compared with those that carry out the same process in other industries. Units with 100,000-barrel-per day capacities are commonplace, and units of over 200,000 barrels per day are now in operation. The American barrel is the most widely used unit in the oil industry and comprises about 160 litres (42 U.S. gallons, or 35 imperial gallons).

Figure 34 shows the principles of operation of a fractional distillation unit. Crude oil is pumped at a constant rate through steel alloy tubes in a furnace, fired by gas or oil, and heated to a temperature between 315°–370° C (600° and 700° F), depending on the type of crude oil and the end products desired. A mixture of vapour and unvaporized oil passes from the furnace into the fraction-

ating column, a vertical cylindrical tower as much as 45 metres (150 feet) high, provided with 30 or 40 perforated fractionating trays spaced at regular intervals. The bubble-cap tray is the commonest type used, but the sieve tray (less expensive, but with a narrower operating range) is sometimes employed. It consists of a simple perforated plate with small holes about 0.48 to 0.64 centimetre (0.19 to 0.25 inch) in diameter.

The oil vapours rise up through the column and are condensed to a liquid in a water-cooled condenser at the top. A small amount of gas remains uncondensed; this is piped away to the refinery fuel-gas system. A pressure control valve on this line maintains fractionating-column pressure at the required figure, usually near atmospheric pressure.

Reflux

Part of the condensed liquid, called reflux, is pumped back into the top of the column and runs down from tray to tray, contacting the rising vapours as they pass through the slots in the bubble caps. The liquid progressively absorbs heavier constituents from the vapours and in turn loses its lighter components. Condensation and re-evaporation take place on each tray. Eventually, an equilibrium is reached in which there is a continuous gradation of temperature and of oil properties throughout the column, with the lightest constituent on the top tray and the heaviest at the bottom. The use of reflux and a column of this type distinguishes fractional distillation from simple distillation.

In the column shown in Figure 34, fractions called sidestreams are withdrawn at several points. These products have properties intermediate from the top and base of the column. Typical boiling ranges for various products are as follows: light gasoline (overhead) 25°–95° C (75°–200° F); naphtha (no. 1 sidestream) 95°–150° C (200°–300° F); kerosene (no. 2 sidestream) 150°–230° C (300°–450° F); and gas oil (no. 3 sidestream) 230°–340° C (450°–650° F).

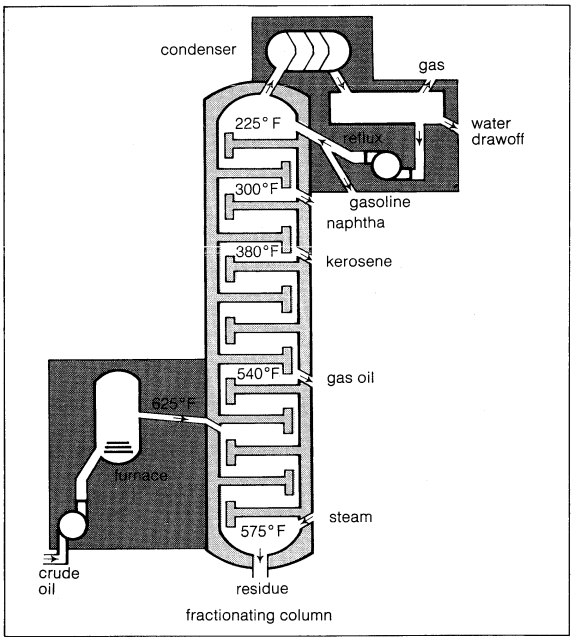


Figure 34: General operating principles of a fractional distillation unit (see text).

In practice, the boiling ranges of these products can be varied within wide limits according to requirements. This is accomplished either by selecting different draw-off points in the column or by varying the quantity of oil withdrawn and has the effect of changing the equilibrium concentration of the liquid on the tray concerned.

The degree of fractionation, or sharpness of separation between hydrocarbons, depends on the number of trays and their efficiency in achieving equilibrium between vapour and liquid. It also depends on the reflux ratio; *i.e.*, the volume of liquid pumped back divided by the volume of overhead product. Reflux ratios in crude oil distillation columns are usually between 1:1 and 3:1.

Treatment of unvaporized oil

Unvaporized oil entering the column (Figure 34) flows downward over another set of trays in the lower part of the column called stripping trays, which remove any light constituents remaining in the liquid. To assist in this, steam is injected at the bottom of the column. The residue that passes from the base of the fractionating column is suitable for blending into fuel oils. Alternatively, it may be distilled a second time under vacuum conditions and further quantities of distillate recovered for use as a starting material for manufacturing lubricating oil or as feedstock for catalytic cracking.

Vacuum distillation. The principles of vacuum distillation resemble those of fractional distillation, and, except that larger diameter columns are used to maintain comparable vapour velocities at the reduced pressures, the equipment is also similar. The vacuum is produced by steam ejectors in vacuum distillation. Components that are less volatile can be distilled without raising the temperature to the range at which cracking occurs, as it would at atmospheric pressure. Firing conditions in the furnace are adjusted so that the oil temperature does not exceed about 400° C (750° F). The residue after vacuum distillation is asphalt, or bitumen.

Superfractionation. Superfractionation is an extension of fractional distillation employing columns with a much larger number of trays (*e.g.*, 100) and reflux ratios exceeding 5:1. With such equipment it is possible to obtain fractions containing only a few hydrocarbons or even to separate pure compounds. By this method, isopentane of over 90 percent purity is produced for aviation gasoline; isohexane and isoheptane concentrates are also prepared for the same purpose. These isoparaffins have much higher octane numbers than the corresponding normal paraffins.

Other methods. *Absorption and stripping.* Absorption and stripping are processes used to obtain valuable light products such as propane/propylene and butane/butylene from the gasoline vapours that pass out of the top of the fractionating tower. In the absorption process, gasoline vapours are bubbled through an absorption oil such as kerosene or heavy naphtha in equipment resembling a fractionating column. The light products dissolve in the oil while dry gases such as hydrogen, methane, ethane, and ethylene pass through undissolved. Absorption is more effective under pressures of about seven to 11 kilograms per square centimetre (100 to 150 pounds per square inch) than it is at atmospheric pressure.

The light products are separated from the absorption oil in the stripping process. The solution of absorption oil and light products is boiled by steam and passes to a stripping column, where the light product vapours pass upward and are recovered by condensation by water-cooling under pressure. The unvaporized oil passes from the base of the column for reuse.

Solvent extraction. Solvent-extraction processes are used primarily for the removal of constituents that would have an adverse effect on the performance of the product in use. The quality of kerosene is improved by the extraction of aromatic compounds that burn with a smoky flame (these compounds can be used to advantage in tractor fuels).

Another important operation is the removal of heavy aromatic compounds from lubricating oils. Removal improves the viscosity–temperature relationship of the oil, extending the temperature range over which satisfactory lubrication is obtained. The usual solvents for extraction of lubricating oil are furfural and phenol. Other solvents are dichloroethyl ether, nitrobenzene, and a mixture of liquid propane and cresylic acid.

Adsorption. Certain highly porous, solid materials have the ability to select and adsorb specific types of molecules, thus separating them from other types. Silica gel is used in this way to separate aromatics from other hydrocarbons, and activated charcoal is used to remove liquid components from gases.

Adsorption is thus somewhat analogous to the process of absorption with an oil, although the principles are different. Layers of the absorbed material only a few molecules thick are formed on the extensive interior surface of the adsorbent; this interior surface may amount to several acres per pound of material.

Molecular sieves

Recent years have brought new developments in the use of adsorbents of a very selective nature called molecular sieves. Molecular sieves are produced by the dehydration of naturally occurring or synthetic zeolites (crystalline alkali-metal aluminosilicates). The dehydration leaves intercrystalline cavities that have pore openings of definite size, depending upon the alkali metal of the zeolite. Under adsorptive conditions normal paraffin molecules can enter the crystalline lattice and be selectively retained, whereas the other molecules are excluded. This principle is used commercially for the removal of normal paraffins from gasoline fuels, thus increasing their octane number. The use of molecular sieves has also been extended to the separation of hydrocarbons of higher molecular weight.

Thermal diffusion. When a mixture of hydrocarbons is passed through a narrow gap, of the order of 0.25 millimetre (0.01 inch), between hot and cold surfaces, some of the constituents concentrate near the hot surface and others near the cold. The phenomenon, known as thermal diffusion, is not clearly understood, but it is believed that separation occurs as a result of differences in the shapes of the molecules. Though this process has been applied in the laboratory as an analytical tool, it is unlikely to find much use in industry as the thermodynamic efficiency is low.

Crystallization. The crystallization of wax from lubricating oil fractions is essential to make the oils suitable for use. A solvent (*e.g.*, a mixture of benzene and methyl ethyl ketone) is first added to the oil and the solution chilled to about −20° C (−5° F). The function of the benzene is to keep the oil in solution and maintain its fluidity at low temperatures, whereas the methyl ethyl ketone acts as a wax precipitant. Rotary filters are used to filter off the wax crystals on a specially woven canvas cloth stretched over a perforated cylindrical drum. A vacuum, maintained within the drum, sucks the oil into it. The wax crystals are removed from the cloth by metal scrapers, after washing with solvent to remove traces of oil. The solvents are later distilled from the oil and reused.

ALTERATIONS OF MOLECULAR STRUCTURE

The separation processes described above are based on differences in physical properties of the components of crude oil. By chemically changing their molecular structure, it is possible to convert less valuable hydrocarbon compounds into those in demand.

The first of these conversion processes is cracking, or thermal decomposition of long-chain hydrocarbon molecules into shorter molecules with lower boiling points. For example, a paraffin molecule such as dodecane ($C_{12}H_{26}$) has such poor "antiknock" properties that it cannot be used in a modern automobile engine, but under intense heat it breaks down into shorter molecules such as paraffins (C_6H_{14}; C_7H_{16}) or olefins (C_6H_{12}; C_5H_{10}), which are suitable for motor fuels. The chemical reactions that take place in a cracking operation are complex.

The product derived from a cracking process is, in effect, a synthetic crude oil. Desired gasoline fractions must be separated by distillation.

Thermal cracking and reforming. The earliest cracking techniques are typified by processes in which kerosene or gas oil materials were converted by heating to temperatures between 450° and 540° C (850° and 1,000° F) at pressures of 18 to 35 kilograms per square centimetre (250 to 500 pounds per square inch). This process produced gasolines of about 70 octane number (low by modern standards).

Visbreaking, another thermal cracking process, reduces viscosity of heavy crude oil residues to make them more suitable for inclusion in fuel oils.

The steam cracking process, by which ethylene and other olefins are made from naphtha, differs from thermal cracking in that it is carried out at low pressures and higher temperatures.

Thermal reforming, a modification of the thermal cracking process, reforms or alters the properties of low-grade components such as naphthas by converting the molecules into those of higher octane number. Pressures used are somewhat higher than in cracking. At temperatures from 510° to 566° C (950° to 1,050° F), it is possible to obtain gasolines with octane numbers of 70–80 from components of less than 40.

Catalytic cracking. The development of catalytic cracking was one of the major achievements of the petroleum industry. Enormous catalytic cracking plants dominate the skylines of modern oil refineries and operate continuously for years on end. The catalytic cracking of gas oil is one of the most important processes for manufacture of gasoline.

Use of a catalyst (a material that assists a chemical reaction but does not take part in it) in the cracking reaction increases the yield of improved quality products under much less severe operating conditions than in thermal cracking. Typical temperatures used are from 454° to 510° C (850° to 950° F) at much lower pressures of 0.7 to 1.4 kilograms per square centimetre (10 to 20 pounds per square inch). At first natural or synthetic clays were used as catalysts. A typical synthetic clay contains 12.5 percent alumina and 87.5 percent silica. In the last few years zeolitic or molecular sieve-base catalysts have been introduced. These give greater gasoline yields while reducing the formation of gas and coke.

Cracking may be accomplished either with a "fixed bed" of catalyst, with a "moving bed" of granular catalyst, or with a "fluid" catalyst, which is a finely divided solid having properties analogous to a liquid when agitated by air or oil vapours.

Fluid-catalytic cracking process

The fluid-catalytic cracking process may be examined as an example. Figure 35 shows the principles of operation of a fluid-catalytic cracking unit.

In this arrangement a reactor and a regenerator vessel are

From *Hydrocarbon Processing* (September, 1968)

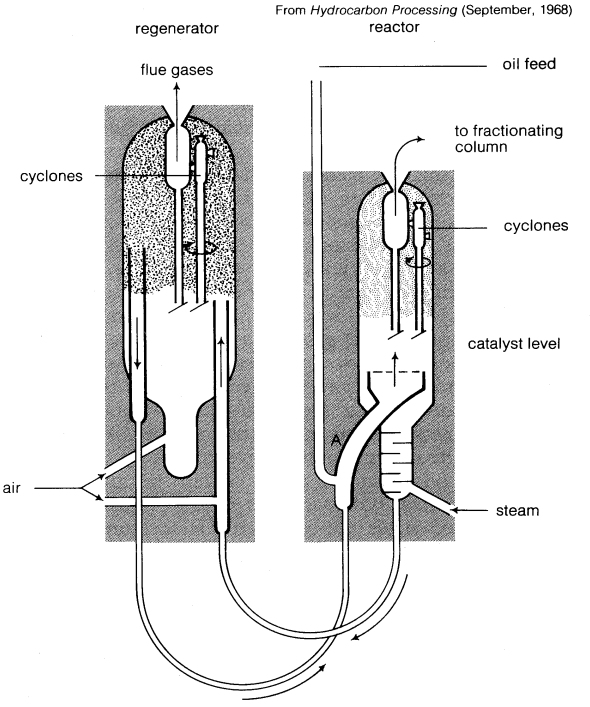


Figure 35: General operating principles of a fluid-catalytic cracking unit (see text).

located side by side. The oil is vaporized when it meets the hot catalyst at point A. The vapours force the catalyst up into the reactor, which (because its cross section is larger in area) allows the catalyst to settle into a bed whose depth can be varied to regulate the reaction time. The oil vapours maintain the bed of catalyst in a turbulent condition, thus effecting the close contact between vapours and catalyst necessary for cracking.

As the cracking reactions proceed, carbon is deposited on the catalyst particles. Since these deposits impair its efficiency, catalyst is continuously withdrawn from the base of the reactor and regenerated in the other vessel by burning off the carbon with a current of air. This also heats up the catalyst to the required reaction temperature.

Product vapours pass out of the top of the reactor through separators using centrifugal force, called cyclone separators, where the catalyst is removed and dropped back into

the reactor bed. Similar cyclones recover catalyst from the flue gases that emerge from the top of the regenerator. A small amount of fresh catalyst is added to the system from time to time to maintain the activity and a similar amount withdrawn.

The products are fractionated in a distillation column. The yield of gasoline is in the region of 40–60 percent of the gas oil feed.

Catalytic reforming. This process uses catalysts to mold molecular structure into desirable forms without the formation of unwanted heavy products and coke. As in thermal reforming, a naphtha-type material is used as the feedstock, but the reactions are carried out in the presence of hydrogen, which inhibits the formation of unstable unsaturates that polymerize into high-boiling compounds.

Use of platinum, chlorine, and fluorine

Most catalytic reforming processes use platinum as the active catalyst. It is distributed on the surface of aluminum oxide as a carrier. Small amounts of chlorine and fluorine act as catalyst promoters. In spite of the high cost of platinum, the process is feasible because of the long life of the catalyst and the high quality and yield of the products obtained. This process has almost entirely replaced thermal reforming.

Other important reactions include the breaking down of long-chain hydrocarbon molecules into smaller saturated chains and the formation of isoparaffins, made up of branched-chain molecules. Formation of ring compounds, or cyclization, of paraffins into naphthenes also takes place and naphthenes are in turn dehydrogenated into aromatic compounds. The desirable end products are isoparaffins and aromatics, both with high octane numbers.

Commercially available processes include platforming, powerforming, ultraforming, and catalytic reforming. In a typical platforming unit, the naphtha charge mixed with hydrogen (the mixture is about five molecules of hydrogen to one of hydrocarbon) is heated to a temperature normally in the range of 455°–535° C (851°–995° F). This gaseous mixture passes downward through the pellets of catalyst contained in three reactors in series. Operating pressures vary from 14 to 56 kilograms per square centimetre (200 to 800 pounds per square inch). Because heat is absorbed in the reforming reactions, the mixture is reheated in intermediate heaters between the reactors.

After leaving the third reactor, the liquid product is condensed, separated, and passed to the stabilizer column, where hydrocarbon gases produced in the reactions are removed. The catalytic reformate product is then available for blending into motor gasoline without further treatment. The hydrogen leaving the products separator is returned to the reactors by the compressor.

Operating conditions are set to obtain the required octane number, usually between 90 and 100. At the higher levels, reduced yields are obtained. Octane numbers can be raised by the addition of between 1.5 and 3 millilitres (0.05 and 0.1 ounce) of tetraethyl lead per 3.8 litres (one U.S. gallon). In some countries, legislation is proposed to limit the amount of tetraethyl addition to motor gasolines as an antipollution measure. Such restrictions will tend to increase the use of catalytic reforming, an important process for production of high-octane components.

Polymerization. Gaseous hydrocarbons produced in thermal and catalytic cracking processes may be used for refinery fuel, or they may be converted into more valuable products by various methods such as polymerization. In polymerization, the hydrocarbon molecules are induced to combine, or polymerize, into molecules of two or more times the molecular weight, forming a material that performs well as a motor fuel. Early commercial processes used 65 percent sulfuric acid as catalyst but were soon superseded by processes using catalysts of phosphoric acid on kieselguhr or the phosphates of copper and cadmium. High pressures, of the order of 28 to 84 kilograms per square centimetre (400 to 1,200 pounds per square inch) are required at temperatures from 175° to 230° C (350° to 450° F).

The polymer gasolines derived from petroleum fractions containing propylene and butylene have octane numbers above 90, and with the addition of tetraethyl lead, above 100.

Alkylation. The alkylation reaction also achieves a longer chain molecule by the combination of two smaller molecules, one being an olefin and the other an isoparaffin (usually isobutane). During World War II, alkylation became the main process for the manufacture of isooctane, which made possible the production of large quantities of aviation gasoline.

Sulfuric acid alkylation. Concentrated sulfuric acid (98 percent) is used as the catalyst, and the process is carried out in reaction vessels of mild steel, as corrosion of this material with concentrated acid is relatively low. The reaction must be carried out at low temperatures, 2° to 7° C (35° to 45° F). Refrigeration is necessary because heat is generated by the reaction.

The octane numbers of alkylates produced range from 89 to 95. For aviation fuel, butylene fractions are usually used as feedstocks. After the addition of tetraethyl lead, octane numbers of over 100 are attainable. The sulfuric acid alkylation process was announced in 1938 and is still extensively used for production of aviation gasoline and high-octane motor gasoline.

Hydrofluoric acid alkylation. Hydrofluoric acid is used as the catalyst in this variant of the alkylation process. The chemical reactions are similar, but it is possible to use higher temperatures (between 24° and 46° C [75° and 115° F]) thus avoiding the need for refrigeration. Recovery of the hydrofluoric acid is simple and it can be accomplished by distillation. Stringent safety precautions must be exercised when using this highly corrosive and toxic substance.

Isomerization. The demand for aviation gasoline became so great during World War II and after that the quantities of isobutane available for alkylation feedstock were insufficient. The deficiency was remedied by isomerization (changing molecular structure) of the more abundant normal butane into isobutane. The isomerization catalyst is aluminum chloride supported on alumina, promoted by hydrogen chloride gas.

Commercial processes have also been developed for isomerization of normal pentane and normal hexane to the corresponding isoparaffins. Here, the catalyst is usually of the platinum type. The reactions are carried out in the presence of hydrogen, as in catalytic reforming. Hydrogen is neither made nor consumed but is used to inhibit undesirable side reactions. The objective of these processes is to convert low-octane-number pentane and hexane feedstocks into high-octane-number gasoline-blending components or into highly pure isopentane or neohexane (2,2, dimethylbutane). In the latter case the isomerization product is purified by molecular sieve extraction followed by distillation. Application of this process is increasing and will be more valuable if limitations are placed on the use of tetraethyl lead in gasolines.

Hydrogenation and hydrocracking. One of the most far-reaching developments since World War II has been the use of hydrogen in refining, in part as the result of the availability of hydrogen as a by-product of catalytic reforming. But refining with hydrogen is now becoming so popular that hydrogen-manufacturing plants are being built.

One of the earliest applications was the conversion of octenes (eight-carbon-atom olefins) produced by polymerization into isooctane for manufacture of aviation gasoline, but the alkylation process has since rendered this obsolete. Today, two main types of hydrogenation exist; hydrotreatment and hydrocracking. Hydrotreatment is used to remove impurities such as sulfur, nitrogen, and oxygen (see below *Purification processes*).

Advantages of hydrocracking

Although capital costs are high, the flexibility of hydrocracking is such that it promises to become the most effective process available to the refiner for extending the use of the barrel of crude oil. This process can be used for producing gasoline or middle distillates from heavy gas oils, for converting residues into lighter oils, or for producing liquefied petroleum gases from straight-run naphthas. By cracking these in the presence of hydrogen and a catalyst, valuable products are obtained without simultaneous formation of coke and large quantities of gas. The activity of the catalyst can be maintained for long

periods. Continuous regeneration is not necessary, as it is in catalytic cracking.

In the United States, the main objective has been to increase gasoline production, particularly from heavy crudes. In the Eastern Hemisphere there has been a trend to install units designed for middle-distillate production, to augment the supplies of heating oils; but in Europe, gasoline-producing units are being erected, releasing naphtha (petrochemical feedstock) for catalytic reforming.

Hydrocracking is accomplished at lower temperatures than catalytic cracking—*e.g.*, at 260° to 427° C (500° to 800° F)—but at much higher pressures—70 to 140 kilograms per square centimetre (1,000 to 2,000 pounds per square inch). The design and manufacture of the large, thick-walled vessels for operation under these conditions has been a considerable engineering achievement.

The catalysts used vary widely. The cracking reactions are induced by materials of the silica–alumina type. Hydrogenation catalysts such as nickel, tungsten oxide, platinum, or palladium are incorporated. Molecular sieve-type catalysts with a hydrogenation constituent are also being used.

PURIFICATION PROCESSES

Before petroleum products are marketed, certain impurities must be removed or made less obnoxious. The most common impurities are sulfur compounds, which, apart from their evil odour, are technically undesirable. In motor and aviation gasolines these compounds reduce the effectiveness of the antiknock tetraethyl lead; in diesel oils, they cause corrosion in the engine. Finally, there is increasing need in many countries to limit sulfur content of fuel oils to reduce atmospheric pollution.

Common impurities

Most crude oils contain small amounts of hydrogen sulfide; this is increased by the breakdown of heavier sulfur compounds during distillation. Though the bulk of the hydrogen sulfide passes off in the overhead gases, some remains dissolved in the gasoline. It is easily removed by alkaline solutions. The light distillates and products from cracking processes also contain more complex sulfur compounds that are not removed so readily. The most common of these are the mercaptans, which have a particularly unpleasant smell. Other undesirable constituents include oxygenated compounds such as alkyl phenols and naphthenic acids, nitrogen bases, gums and gum-forming constituents, and unstable compounds that can lead to colour formation in the product.

Processes for the purification of petroleum products are described below. Product quality is assessed by standard test methods. In addition to chemical tests for impurities, they include many physical tests designed to measure performance characteristics against relevant specifications.

Sulfuric acid treatment. The use of concentrated sulfuric acid as a refining agent is as old as the petroleum industry. It has been applied to almost every type of product. Though still important, its use has declined in recent years, following the introduction of solvent extraction and catalytic processes using hydrogen. Sulfuric acid removes sulfur compounds by solution. It also polymerizes highly reactive hydrocarbons and neutralizes nitrogen bases. Asphaltic constituents and easily oxidizable compounds are also removed.

Sulfuric acid treatment improves the colour of cracked gasolines and inhibits the formation of gum in storage. Sulfuric acid is particularly valuable for refining special boiling-point solvents and white spirits when a very low sulfur content is required. In the past it was extensively used for treatment of lubricating oils. The introduction of more modern processes, such as solvent extraction, has displaced it except for special purposes.

A serious inconvenience associated with acid treatment is the need to dispose of the acid tars that are formed as waste. An acid recovery plant is required to regenerate the sulfuric acid for reuse, both to economize in acid consumption and to prevent pollution. Most modern refineries have adopted catalytic refining processes to avoid these problems. Another disadvantage is the slight loss of useful product in acid treatment, since there is some polymerization to high-boiling compounds. These compounds are eliminated by redistillation of the product.

Sweetening. Sweetening processes convert evil-smelling and corrosive mercaptans into relatively innocuous disulfides. Mercaptans are slightly acidic compounds. Their acidity decreases as the number of carbon atoms increases. Examples are methyl mercaptan, ethyl mercaptan, and propyl mercaptan. The lighter ones are partially removed by alkali treatment, but in sweetening reactions they are oxidized. Catalysts assist oxidation. In what is known as the doctor process of oxidation, the catalyst is a solution of sodium plumbite, made by dissolving lead oxide in caustic soda. This inexpensive process was very popular until recently, but the necessity of adding elemental sulfur to make the reactions proceed causes an increase in the total sulfur content of the oil. The doctor process is consequently being superseded by processes in which there are no undesirable side reactions and no increase in total sulfur content. One such is the copper chloride process, in which the catalyst is a suspension, or slurry, of copper chloride and fuller's earth. It is applicable to both gasolines and kerosenes. The oil is heated and then brought into contact with the slurry while being agitated by a stream of air or oxygen that oxidizes the mercaptans to disulfides. The slurry is then allowed to settle and is separated for reuse. The heater raises the temperature to a point that will keep the water formed in the reaction dissolved in the oil, so that the catalyst slurry remains at the right degree of hydration. After sweetening, the oil is water washed to remove any traces of catalyst and dried by a salt filter.

The copper chloride process

Mercaptan extraction. Simple sweetening is adequate for many purposes and is widely used, but other methods must be adopted if actual removal of sulfur compounds is required. One of these methods is mercaptan extraction. When certain chemicals called solutizers, such as potassium isobutyrate and sodium cresylate, are added to caustic soda solution, the solubility of the higher mercaptans is increased and they can be extracted from the oil. The oil is first pretreated with caustic soda to remove traces of hydrogen sulfide and alkyl phenols and then brought into contact with the solutizer solution, either in a packed column or other mixing device. The mixture is allowed to settle and the product pumped away to storage, via a water washer.

Regeneration of the solution containing sodium mercaptides can be accomplished by heating or by oxidation with air in the presence of a catalyst. In the latter case the regenerated solution is washed with an oil to remove the disulfides that are formed before being recycled to the extraction stage.

Clay treatment. Some natural clays, activated by roasting or by treatment with steam or acids, have been used for many years in the industry for removing traces of impurities. The phenomenon is similar to that described under the adsorption separation process; the clays retain the larger long-chain molecules within their highly porous structure.

Clay treatment is occasionally used to remove gum and gum-forming materials from thermally cracked and thermally reformed gasolines in the vapour phase. A more economic procedure, however, is to add small quantities of inhibitors, synthetic anti-oxidants, to the gasoline. These prevent, or greatly retard, gum formation.

Final purification of lubricating oils by clay treatment is in general use. Resins and highly coloured bodies remaining after solvent extraction are removed, and a pleasing appearance is given to the oil. The treatment may be by contact—that is, the clay is added to the oil, the mixture heated, and the clay filtered off—or by percolation, in which the heated oil is allowed to flow through a large bed of active adsorbent.

The spent clay is usually discarded, although there are processes in which the clay is regenerated by roasting. Disposal of spent clay presents a problem to refineries. Newer hydrofinishing processes are available in which mild hydrogenation achieves results similar to treatment by clay.

Hydrogen treatment. A major application of the hydrogen treatment is in the removal of sulfur compounds of all types. Hydrogen treatment can be applied to a wide range of products. The processes are known as hydrofining or hydrodesulfurization. In them, the oil is mixed

with hydrogen and vaporized before passing over a catalyst such as tungsten, nickel, or a mixture of cobalt and molybdenum oxides, supported on an alumina base. Mild operating temperatures are used—*e.g.*, between 260° and 427° C (500° and 800° F)—and pressures between 14 and 70 kilograms per square centimetre (200 and 1,000 pounds per square inch).

The sulfur in the oil is converted to hydrogen sulfide, which is eliminated from the circulating hydrogen by absorption in a solution such as diethanolamine, which can then be heated to remove the hydrogen sulfide and reused. The hydrogen sulfide recovered is useful for manufacturing elemental sulfur of high purity.

Hydrotreatment processes are used primarily for sulfur removal from gasolines, naphthas, kerosenes, and gas oils. There is no breakdown of hydrocarbons and so the oils are virtually unchanged, except for the purification.

Molecular sieves. Molecular sieves are also used to purify petroleum products since they have a strong affinity for polar compounds such as water, carbon dioxide, hydrogen sulfide, and mercaptans.

Usually sieves with a pore size of 13 angstrom units (13×10^{-8} centimetres) are used. These are prepared by the dehydration of a sodium aluminosilicate (zeolite). Fixed beds of adsorbent are used. The petroleum product is passed through for a predetermined period, depending on the amount of impurity to be removed. The adsorbed contaminants may then be eliminated from the sieves by purging with a gas stream at temperatures between 200° and 315° C (400° and 600° F). The cycle is then repeated.

PETROLEUM PRODUCTS AND THEIR USES

Refinery gas

Gases. Refinery gas has a heating value similar to natural gas and is used to enrich supplies. It is used as a starting material for petrochemical industries. It is also converted (by means of distillation and other treatment) into liquefied petroleum gas (LPG), a convenient, highly portable fuel for domestic and light industrial use.

Gasolines. *Motor gasoline.* A good motor gasoline must possess three characteristics. (1) The fuel must burn smoothly in the engine without detonating, called knocking. A common performance characteristic is the octane number, an expression of the engine performance of a gasoline when compared with two standard test fuels, normal heptane and isooctane. Normal heptane, a fuel of poor quality (*i.e.*, very prone to "knocking"), is assigned a value of 0 and isooctane, of very high quality, 100. Thus, when a motor gasoline is given an octane number of 90, it means that its test performance in a standard engine is equivalent to that of a mixture of 90 parts by volume of isooctane and 10 parts of normal heptane. There are two methods for carrying out the test: the Research Method, under mild conditions of temperature and running speed (49° C [120° F] and 600 revolutions per minute), and the Motor Method, under more severe conditions (149° C [300° F] and 900 revolutions per minute). In commercial practice, Research Method ratings are usually quoted. Motor gasolines are blended to required octane number (90 to 100 for most modern cars) from the components whose manufacture is described above. A small amount of tetraethyl lead (TEL) is usually added. (2) A good motor gasoline must start the car easily in cold weather. (3) The fuel system must not become vapour locked when temperatures are high. Volatility and vapour pressure define these latter two requirements quantitatively.

Octane number

Aviation gasoline. Piston-type aircraft engines are extremely fastidious in their demands on fuel quality. Knocking cannot be tolerated without risk of serious mechanical failure; this is true whether a weak mixture is being used, as under cruising conditions, or whether an enriched mixture is being used, as during takeoff or under maximum power.

Antiknock ratings of aviation gasolines are therefore determined under both conditions. The methods used, known respectively as the F3 and F4 methods, can rate fuels with octane numbers above 100. Aviation fuels are designated by quoting both figures. Thus a 100/130-grade fuel has an F3 rating of 100 and an F4 rating of 130. Because of the low temperatures in flight at high altitudes, a maximum freezing point of −60° C (−76° F) is specified to ensure that the fuel remains fluid.

Kerosene. Though its use as an illuminant has greatly diminished, kerosene still finds considerable use throughout the world in cooking and space heating. Newer uses are as fuel for tractors and as a turbine fuel for jet and turboprop aircraft.

Burning oil. Kerosene used as a burning oil must not give a smoky flame and there must be a minimum of charring of the wick or formation of "bloom" on lamp glasses. Standard laboratory procedures test these properties by burning the oil in special lamps. All kerosene fuels must satisfy minimum flash-point specifications, 49° C (120° F), to limit fire hazard in storage and handling.

Jet fuels. Jet fuels must burn cleanly in the aircraft engine combustion chambers. The first fuel used was a kerosene, and this type fuel still finds considerable use in civil aircraft. For military purposes, however, and to increase availability, it was necessary to extend the boiling range to include both gasolines and kerosenes.

As with aviation gasolines, the freezing point is also a significant factor, and the fuels are freed of suspended water that might cause blockage of the fuel system with ice particles.

Gas oils and diesel oils. The main use of gas oils is to power the diesel engines in trucks, buses, railway engines, and similar vehicles. In a diesel engine, combustion is induced by the heat of compression of the air in the cylinder caused by the moving piston. Detonation, which leads to knocking in a gasoline engine, is a necessity for diesel fuel. A good diesel fuel starts to burn at several places in the cylinder soon after the oil is injected as a mist. Once the flame is started, any further fuel entering the cylinder ignites at once.

Straight-chain hydrocarbons make the best diesel fuels. In order to have a standard reference scale, the oil is matched against blends of cetane (normal hexadecane) and alpha methylnaphthalene, the latter of which gives a very poor engine performance. High-quality diesel fuels have cetane numbers around 50, equivalent to a blend containing 50 percent cetane and 50 percent alpha methylnaphthalene.

For the slower diesel engines used in ships and in stationary engines on land, it is possible to cut deeper into the crude oil than with high-speed diesel fuels. The heavier, more viscous grades of oil are heated to permit easy pumping and to give the correct viscosity at the fuel injectors for good combustion.

Fuel oils. Fuel oil, or furnace oil, consists largely of residues from crude-oil distillation. These are blended with other suitable fractions to obtain the required viscosity, the basis on which fuel oils are sold. About 70 percent of these oils are used to generate steam in power stations, ships' boilers, and industrial boilers. Their calorific value ranges from about 19,000 British thermal units per pound (Btu/lb) for the lightest fuels to about 18,300 Btu/lb for the heaviest. (The British thermal unit is the amount of heat required to raise the temperature of one pound of water one degree Fahrenheit.) Calorific value depends on the carbon–hydrogen ratio of the fuel and also on its sulfur content, which may vary from about 1 to 4 percent, according to the source of the crude oil. The combustion properties of fuel oils of high sulfur content are satisfactory if the flue gases do not impinge on cool surfaces, which could lead to corrosion by the condensation of acidic sulfur trioxide.

With increasing attention given to air pollution, some highly industrialized countries propose to reduce the concentration of sulfur oxides in the atmosphere by limiting the sulfur content of fuels burned. Since the early 1970s the oil industry has been avoiding the cost of desulfurizing this product by processing mixtures of low- and high-sulfur crude oils.

Fuel oils derived from waxy crudes can solidify under storage conditions; it is usual, therefore, to specify a pour point for the oil. This point (the lowest temperature at which the oil still flows) varies from about 2° C (35° F) for the lighter fuel oils to about 21° C (70° F) for the heavier. The heavier oils are usually heated in storage and in use. A pumpability test, measuring the viscosity–tem-

Pumpability test

perature relationships of the oil, gives a good indication of its performance in actual use.

Lubricating oils. At one time the suitability of petroleum fractions for use as lubricants depended entirely on the sources of the crude oils from which they were prepared. Those from Pennsylvanian crudes were recognized as having very superior properties, being largely paraffinic in nature. But with the advent of solvent extraction the choice of suitable crude oils has been considerably extended.

Viscosity is the basic property by which lubricating oils are classified. Requirements vary from the thin oil needed for the high-speed spindles of textile machinery to the viscous, tacky compositions applied to open gears or wire ropes. Between these extremes is a range of products, many highly refined, with special characteristics.

Automobile oils represent a large proportion of those manufactured. In the U.S., the Society of Automotive Engineers established a system of SAE numbers to indicate the viscosity at a particular temperature, −18° C (0° F). Oils in common use have SAE numbers varying from 5 to 50.

When ordinary mineral oils having satisfactory fluidity at low temperatures are used over an extended temperature range, excessive thinning occurs at the high temperatures. To correct this, multigrade oils incorporating synthetic long-chain polymers have been developed. These have much more satisfactory viscosity–temperature relationships. Thus an oil designated SAE 10W/40 has the viscosity of an SAE 10W oil at −18° C (0° F), and of an SAE 40 oil at 99° C (210° F). Such an oil will help start an engine in winter (hence the suffix W) and will lubricate well under running conditions in summer. Other additives that improve the performance of lubricating oils are antioxidants and detergents, which maintain engine cleanliness by keeping fine carbon particles in suspension in the oil.

Gear oils. In gear lubrication the oil separates metal surfaces, reducing friction and wear. Extreme pressures develop in some gears, notably those in the rear axles of cars; special additives are incorporated to prevent freezing together (seizing) of the metal surfaces. The oils contain sulfur compounds that form a resistant film on the surfaces, preventing actual metal-to-metal contact. Such compounds are known as extreme pressure (EP) additives.

Greases. Greases are lubricating oils to which thickeners, or gelling agents, are added. Soaps of aluminum, calcium, lithium, and sodium are commonly used, while more recently nonsoap thickeners (*e.g.*, carbon, silica, and polyethylene) have been employed.

Other petroleum products. Highly purified distillate fuels are used in pottery kilns, for glass manufacture, and many other purposes, including the baking of bread. Metallurgical operations employing petroleum products include the smelting of ores, forging, and alloy melting.

White spirits are used as thinners for paints and varnishes and for dry cleaning. They are highly purified to avoid unpleasant odours. Similar materials are used in the rubber industry and for the extraction of vegetable oils.

Asphaltic bitumen

Waxes and petroleum jellies are familiar products. Asphaltic bitumen is the earliest known petroleum product, having been used by the pre-Babylonian inhabitants of the Euphrates valley before 2500 BC. Today it is widely used for the construction of roads and airfields. Specialized uses of bitumen include the manufacture of roofing felts, waterproof papers, pipeline coatings, and electrical insulation.

Carbon black was previously made by the incomplete combustion of natural gas. Today it is manufactured by decomposing liquid hydrocarbon fractions; it is compounded with rubber in tire manufacture and is a constituent of printing inks and lacquers.

Petroleum chemicals. By definition, petroleum chemicals are simply chemicals that happen to be made from a starting material obtained from petroleum. They are, in almost every case, virtually identical with the same product from other sources, such as coal and coke or fermentation processes.

Olefins. The thermal cracking process in refineries, as it developed from 1910 onwards, aimed at increasing the quantity and improving the quality of gasoline. As a by-product of this process, gases were produced that included a significant proportion of lower molecular weight olefins (unsaturated compounds of carbon and hydrogen that are chemically very reactive), particularly propylene and *n*-butylenes. Commercial production of chemicals from this source dates from 1919. Catalytic cracking is a valuable source of propylene and *n*-butylenes but not a very good one of ethylene, a most important petroleum chemical.

It is important to differentiate between the practices of the U.S. and most of the rest of the world in producing olefins. In the U.S., propylene and *n*-butylenes are produced from refinery catalytic cracking. Ethylene is largely produced in the U.S. by cracking ethane extracted from natural gas. In Europe, Japan, and other major centres of petroleum chemical production, catalytic cracking is a less important refinery operation, and appropriate supplies of natural gas are not readily available. As a result, both the Europeans and the Japanese generally crack a liquid hydrocarbon fraction in the gasoline range (light or heavy naphtha). In a single operation this yields a whole range of olefins, notably ethylene, propylene, the *n*-butylenes, and butadiene.

Cracking plants are usually situated adjacent to a refinery from which they derive the fractions used as feedstock and to which they return liquid cracking product. Plants that consume these various olefins are clustered around the cracker. Among typical plant complexes in Europe are ICI at Wilton, England; Shell at Carrington, England; BP Chemicals at Grangemouth, Scotland; Shell and Naphtachimie near Marseille, France; Shell, Dow, Gulf, ICI, and other companies near Rotterdam, The Netherlands; Shell/BASF at Wesseling, West Germany; and BP/Bayer at Dormagen, West Germany. Similar major industrial sites are characteristic of the Japanese petroleum chemical industry (for example at Yokkaichi and Tokuyama).

Petroleum chemical plant complexes

After World War II, petroleum became a major source of organic chemicals. Since the early 1970s, however, partly because of the increasing cost of crude oil, other sources of organic chemicals have also been exploited.

Inorganic chemicals. Various inorganic chemicals are also made from petroleum, some, particularly ammonia and sulfur, in large quantities. Ammonia production requires hydrogen from a hydrocarbon source. Traditionally, this hydrogen was produced from a coke and steam reaction, but today most ammonia is synthesized from liquid petroleum fractions, natural gas, or refinery gases. It has been pointed out that sulfur is an unwanted impurity in many petroleum products. In the process of eliminating it from oil products, it is ultimately recoverable either as elemental sulfur or as sulfuric acid. Though various inorganic chemicals manufactured from petroleum represent a substantial tonnage, the total is still only a modest proportion of the world production of such chemicals.

Aromatic compounds. The aromatic compounds, prominent in the refinery process of catalytic reforming, are major sources of petroleum chemical products as well as the lower olefins mentioned above. In the traditional chemical industry, aromatics such as benzene, toluene, and the xylenes were made from coal during the course of coal carbonization in making coke or town gas. Today a much larger volume of these compounds is made in petroleum refineries by catalytic reforming. A further source of supply now increasingly used in Europe is the aromatic-rich liquid fraction produced in the cracking of naphtha to make ethylene and other olefins.

Polymers. A highly significant proportion of products made from these basic chemicals are the plastics, synthetic rubbers, and synthetic fibres known as high polymers because their molecules are the high-molecular-weight compounds made up of the repeated structural units that have combined chemically, or polymerized. In plastics the four major products are polyethylene, polyvinyl chloride, and polystyrene, all made from the simple structural unit, or monomer, ethylene and polypropylene, made from the monomer propylene. Major raw materials for synthetic rubbers include butadiene, ethylene, benzene, and propylene.

Among synthetic fibres the polyesters, which are a combination of ethylene glycol (made from ethylene) and terephthalic acid (made from one of the xylenes), are the most widely used; they account for about one-half of all

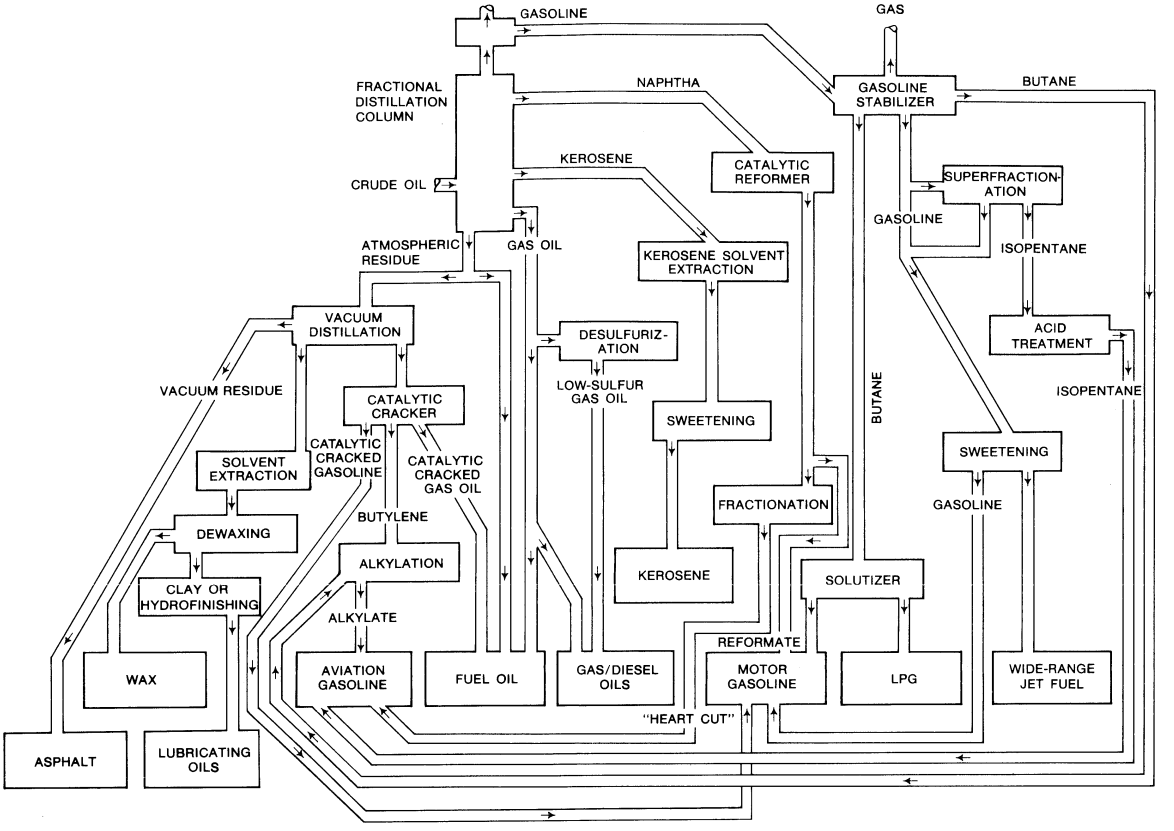


Figure 36: Generalized flow chart of the refinery process.

synthetic fibres. The second major synthetic fibre is nylon, its most important raw material being benzene. The acrylics, in which the major raw material is the propylene derivative acrylonitrile, make up most of the remainder of synthetic fibres.

By-products. Petroleum chemicals have advanced from the original concept of by-product usage and represent a large investment of money and technology. An additional group of products, however, that represent a direct by-product of oil-refining operations, the naphthenic acids, can be abstracted for use in paint dryers and other derivatives. Naphthasulfonates recovered in sulfuric acid refining of oil products serve as emulsifying agents. Many cresylic acid products can be recovered as by-products.

REFINERY PLANT AND FACILITIES

Refining facilities. Integration of the separate processing plants in a modern refinery has resulted in economies of services, intermediate tankage, and manpower. The simplest refinery contains a distillation unit, catalytic reformer, and various treatment units. Small refineries of this type supply local markets in many parts of the world. With such units, it is possible to produce motor gasolines of different octane numbers, lamp kerosene, jet fuel, gas and diesel oil, and residual fuels. There may also be a plant for recovery of propane and butane.

Effect of marketing research on processing

In deciding which process to use, refinery managers analyze the market for residual fuel oil, which may amount to 40 or 50 percent of the crude oil processed. In some countries of the Eastern Hemisphere this fuel oil is easily marketed. But in the U.S. and Canada, where large quantities of gasoline are required, less fuel oil is used because natural gas is available within reach of consuming areas. Cracking plants, usually catalytic, together with vacuum-distillation units to produce fractions for use as feedstock, solve this problem. The production of gasolines can then be further augmented by polymerization or alkylation, superfractionation, and possibly by hydrocracking.

With further additions of plant it is possible to build a fully integrated, comprehensive refinery by stages, including solvent extraction (of kerosene), hydrodesulfurization, lubricating-oil plant (propane de-asphalting, solvent extraction, dewaxing, and clay treatment or hydrofining), and plant for the production of asphalt, wax, sulfur, and special products. The chart in Figure 36 shows how these plants are interrelated.

Tankage. The manufacturing units form only part of a refinery, though the most important one. A large proportion of the cost, and the area covered, is taken up by tankage, handling facilities, etc.

The storage capacity allocated to crude oil may be minimal if the refinery is connected to the oil fields by pipeline. When crude oil has to be imported, however, sufficient tankage is provided to enable continuous refinery operation while still allowing for the irregular arrival of tankers. Scheduling of tanker movements is important in integrated oil operations, particularly with very large tankers (up to 250,000 tons and more) in service.

Finished products require much more tankage in terms of storage capacity than crude oil. Provision is made for short-term variations in the sales of refinery products to avoid frequent alterations in the refinery process units. Nonvolatile products such as diesel oils and fuel oils are stored in large-diameter cylindrical tanks with low-pitched conical roofs. Tanks with floating roofs are used to reduce evaporation losses in gasolines and other volatile products, including crude oils. The roof, which resembles a pontoon, moves up and down inside the tank, thus eliminating the air space that could contain a considerable proportion of petroleum vapour. Fire hazards are also reduced by the use of floating-roof tanks.

For liquefied petroleum gases, pressure vessels (usually spherical) are used. Alternatively, these materials may be stored in suitable geological formations underground, such as cavities in salt domes. Large refrigerated tanks are a recent development. In these, the liquefied gas is stored at normal pressure but at reduced temperatures.

Bulk transportation. The supertankers have sharply cut the cost of transporting crude oil, making it practical to locate refineries near areas of consumption rather than adjacent to oil fields. Deepwater ports have been built to receive these large tankers, particularly in Europe, which obtains most of its supplies from the Middle East and Africa. Examples are Rotterdam, Gothenburg, Le Havre,

Foss near Marseilles, and Milford Haven in the United Kingdom. Groups of refineries have been constructed near these ports or connected to them by pipeline.

If pipelines are not practical or economic, water transport is the first choice for petroleum products. Full-size tankers or smaller coastal vessels, with capacities of 15,000 barrels or less, distribute the different grades of oil to depots situated on the coastline.

Countries having networks of canals and navigable rivers afford many opportunities for using barges, a very cheap method of transportation. On the Mississippi, barges of over 5,000-tons (37,000-barrels) capacity are in use; the Rhine and the Seine are also suitable for large barges. Many canals, however, can only accommodate 200-ton barges (1,500 barrels). Each barge is divided into separate calibrated compartments so that more than one type of product can be carried on a voyage; pumps are fitted for the discharge of the cargo.

Transport by rail tank car is still widely used. The cars normally have capacities of up to 50 tons (390 barrels), but there is a trend to larger sizes.

The final stage of delivery to the majority of customers is by the familiar road tank trucks, whose carrying capacity is about 150 to 200 barrels.

Pipelines and pumping stations. *Pipelines.* Development of crude-oil trunk lines began in the United States in the early years of the 20th century when refineries in the populous eastern area had insufficient supplies from convenient oil fields. Pipelines were used for the transportation of crude oil in the United States in the late 19th century, supplying refineries in the general areas of the known oil fields. After 1900 there was a rapid development of crude-oil trunk lines following the discoveries of oil in Texas, Kansas, and Oklahoma. These fields were used to supplement the declining production of the eastern oil fields. By 1940 more than 85 percent of the crude oil supplied to eastern refineries came from west of the Mississippi River. By the last quarter of the 20th century, there were about 400,000 kilometres (170,000 miles) of oil pipeline in operation in the United States; about one-third of the total mileage consisted of crude-oil trunk lines and a similar amount smaller feeder lines in the oil field areas. A pipeline to bring crude oil from the fields on the North Slope of Alaska to the southern port of Valdez was completed in 1977. Europe also has several crude-oil pipelines supplying inland refineries. In the Middle East large pipelines have been built for transporting crude oil from the oil fields of Iraq, Saudi Arabia, and other oil-exporting countries to deepwater terminals in the eastern Mediterranean.

Crude oil and product pipelines

More and more large pipelines are coming into operation to pump batches of different products successively down the line, since it is possible to do this without intermingling them. The "Big Inch" line, completed in 1943 to transport crude oil from Texas and Louisiana to refineries in the eastern United States, was converted for carrying gas from Texas to New Jersey, a distance of 2,375 kilometres (1,475 miles). In 1964 it was replaced by the world's largest products line, which can transport 1,000,000 barrels of products per day from Houston, Texas, to New Jersey. In Europe the first product line was the TRAPIL line from the lower Seine region to Paris, completed in 1953. This and other such lines on the Continent and in the United Kingdom enable inland refineries to dispose of products economically when the local market pattern does not match the proportions manufactured.

Centrifugal pumps are almost invariably used for main-line pumping, and with long lines the capacity can be increased by installing booster stations at intermediate points along the line. In recent years, many of these have been designed for unattended operation, the starting of pumps and the opening or closing of valves being carried out by automatic sequence control initiated by signals sent over telephone lines or by radio or by instruments responding to line pressure at the unattended pump station.

Service plant and facilities. Present-day refineries contain much equipment for servicing refinery operations. This includes the plant for steam generation and in some cases electric power, water treatment, effluent systems, and antipollution facilities. This equipment is essential for the operation of the industry within the community. One of the advantages of the self-contained refinery is its flexibility regarding surplus products; usually gas and heavy residue can be used for the generation of steam and power and for combustion in the oil-plant furnaces.

(J.E.C./A.L.W.)

Processing, transportation, and storage of natural gas

TREATMENT AND PURIFICATION

Natural gas is not directly usable when it comes from the well. Often it is extremely rich in methane and contains heavy hydrocarbons; in addition, it contains varying proportions of nitrogen and carbon dioxide, hydrogen sulfide, mercaptans, and other sulfur compounds. The aim of gas processing is to make the gas conform to standards of use and to recover the liquid or solid components, which are often of greater commercial value than the gas itself.

After its arrival at the central plant, the gas must undergo the following series of treatments: (1) expansion at processing pressure, which causes the formation of condensates; these are removed, and the gas is then dehydrated by contact with glycol, which is regenerated; (2) sweetening, which removes the hydrogen sulfide and carbon dioxide by diethanolamine treatment, and fixes them in an absorber; (3) elimination of the mercaptans by soda fixation; (4) stripping (either by cooling or by absorption in an oil), which extracts the hydrocarbons with longer chains than methane and produces a gas conforming to commercial specifications. The rich absorption oil and the purified condensate are fractionated by distillation to produce propane, butane, stabilized gasoline, and residues. These successive operations are performed continuously and require an extremely sophisticated control system.

TRANSPORTATION AND DISTRIBUTION

Transportation of gas. Gas fields are often located far from major consumption centres; consequently, the gas has to be transported, although chemical industry transformation plants are frequently installed in the vicinity of production fields. Transportation of natural gas depends upon its form—in a gaseous form it is transported by pipeline under high pressure, and in a liquid form it is transported by boat.

Overland transport by pipeline

Large gas pipelines enable gas to be transported over thousands of kilometres. Examples are the North American pipelines, which extend from Texas and Louisiana to the northeast coast, and from the Alberta fields to the Atlantic Seaboard.

Transportation pressure is generally 70 bars (70 kilograms per square centimetre [1,000 pounds per square inch]) because economic studies, in keeping with technological advances, have shown that transportation costs are at a minimum for pressures of this magnitude. Pipeline diameters for such long-distance transportation have tended to increase from an average of about 60 to 70 centimetres (24 to 29 inches) in 1960 to a size of 1.20 metres (about four feet). Some Soviet projects involve diameters of more than two metres (6.5 feet) and pressures of about 40 bars (41 kilograms per square centimetre [580 pounds per square inch]).

Because of pressure losses, the pressure has to be boosted at regular intervals to keep a constant flow rate in the pipeline, and economic surveys indicate that a compression station should be located every 80 or 100 kilometres (50 or 60 miles) in order to obtain maximum flow under the most economically advantageous conditions. At greater intervals, the pipeline is doubled in size if the gas flow is to be increased.

Despite the high pressure and circulation rates of 10 to 15 metres (33 to 49 feet) per second for gas, the energy carried by a gas pipeline is still one-third or one-half as much as that carried by an oil pipeline because one cubic metre of oil contains as much energy as 1,000 cubic metres (35,000 cubic feet) of natural gas.

Marine transport

Petroleum prospecting has revealed the presence of large gas fields in Africa, the Middle East, Alaska, and else-

where. Transportation from such areas is effected by boat. The gas is liquefied to −160° C (−256° F) and transported as oil in tankers designed especially for this purpose. One cubic metre of liquid natural gas is equivalent to 600 cubic metres (21,200 cubic feet) of gas at atmospheric pressure, with the specific gravity of the liquid being relatively low (about 0.55).

This technique has developed very quickly since the first industrial operations involving it began in 1965 from Africa to Great Britain and France. Two other lines opened up in 1970, the first from Libya to Spain and Italy, and the second from Alaska to Japan. Since then other large-scale projects, including lines from African, South American, and Indonesian fields to the United States, Europe, and Japan, have been undertaken.

After 1965, when methane tankers with a capacity of 20,000 cubic metres (706,300 cubic feet) began to be used, the size of these ships rose to 90,000 and even 120,000 cubic metres (3,178,000 and 4,237,000 cubic feet), an increase that enabled them to carry from 50,000,000 to 70,000,000 cubic metres (1,800,000,000 to 2,500,000,000 cubic feet) of gas per trip.

Distribution and consumption. Production and marketing arrangements vary in different parts of the world. In some nations, such as Mexico and the Soviet Union, the industry is completely nationalized. In Britain, Japan, Italy, and France exploration and development are carried out by either private firms or the government, but transportation and distribution are a government monopoly. In the United States, exploration, production, and transportation are handled by private firms, regulated by federal, state, and local commissions.

Consumption patterns for natural gas also vary. In western Europe, natural gas was not available in most countries until after the mid-20th century; therefore, it supplied only a relatively small percentage of gas demand, but production and consumption of natural gas increased in the late 20th century to account for about one-seventh of the total energy requirements of western Europe. Natural-gas demand in North America, on the other hand, remained fairly constant during the same period, contributing about one-quarter of the total energy requirements. Residential and commercial uses have consumed the largest proportion of natural gas in North America and western Europe, while industry consumes the next largest amount and electric-power generation is a distant third in natural-gas consumption. In western Europe, natural gas provides the same advantages to industrial users as it does in homes and thus is expected to continue making significant inroads into this market at the expense of coal and, to a lesser extent, oil.

By far the major use of natural gas is as fuel, though increasing amounts are being taken by the chemical industry for raw material. Among industries consuming large volumes are food, paper, chemicals, petroleum refining, and primary metals. In the United States, a large amount fuels household heating plants; in the Soviet Union, a considerable volume goes for electric-power generation.

Ideal feedstock for chemical industry

In the chemical industry, natural gas supplies an ideal feedstock (raw-material source) for the manufacture of such widely used industrial gases as methanol, acetylene, helium, and hydrogen.

Although helium, for example, can be separated from the air, all commercial production is from natural gas, containing helium in relatively high percentages. Helium is extracted by cooling natural gas to the point at which it condenses to a liquid, leaving behind an uncondensed mixture of helium and nitrogen. Known as crude helium, this mixture is separated from the liquid natural gas and then further cooled until the nitrogen condenses. The liquid nitrogen is removed, and the helium gas remaining is purified to a concentration of 99.995 percent by passing it through a bed of activated charcoal at liquid-nitrogen temperatures.

With the widespread introduction of natural gas into commercial markets, hydrogen manufacturers shifted from coke to natural gas, from which hydrogen can be made by two different methods. In the first, the steam-reforming process, natural gas is reacted with steam over a catalyst at high temperatures to produce a mixture of hydrogen and carbon oxides. In the second, the partial-oxidation process, the natural gas is reacted with oxygen under pressure, forming a mixture of hydrogen and carbon monoxide along with small quantities of other impurities.

GAS STORAGE

In the course of marine transport, necessary land storage requires double-walled tanks with special insulation. Such tanks may have unit volumes of as much as 50,000 cubic metres (1,766,000 cubic feet). The low temperatures attained require the use of special steels with a high nickel content or aluminum-base alloys. Because the problem of heat loss is less important at ports of arrival, where the liquid gas has to be vaporized, frozen-earth tanks, consisting of enormous holes dug in the ground and made leakproof by freezing the groundwater, are used.

Storage facilities on a more extensive scale have been created by using depleted oil or gas reservoirs near consumption centres (northeastern United States, northern Italy), or by the creation of artificial gas fields in aquifer layers. The latter technique developed rapidly, and the number of storage facilities of this type in the United States has increased tremendously since 1950. There are a number of such underground storage areas in France, Germany, and the Soviet Union. (H.Pe./Ed.)

COAL

Mining

Coal mining is a worldwide basic industry. Though the industry has operated on a huge scale for decades, the existence of vast coal reserves in many parts of the world guarantees its future, at least until coal has been superseded as a source of energy and of many chemical compounds.

HISTORY OF COAL MINING

Ancient use of coal

There is archaeological evidence that coal was burned in funeral pyres during the Bronze Age, 3,000 to 4,000 years ago, in Wales. Aristotle mentions coal ("combustible bodies") in his *Meteorologica*, and his pupil Theophrastus also records its use. The Romans in Britain burned coal before AD 400; cinders have been found among the ruins of Roman villas and towns and along the Roman wall, especially in Northumberland, near the outcrop of coal seams. The Hopi Indians of what is now the southwestern United States mined coal by picking and scraping and used it for heating, cooking, and in ceremonial chambers as early as the 12th century AD; in the 14th century they used it industrially in pottery making. Marco Polo reports its use as widespread in 13th-century China. The Domesday Book (1086), which recorded everything of economic value in England, does not mention coal. London's first coal arrived by sea in 1228, from the areas of Fife and Northumberland, where lumps broken from submarine outcroppings and washed ashore by wave action were gathered by women and children. Thereafter, the name sea coal was applied to all bituminous coal in England. Later in the century, monks began to mine outcroppings in the north of England.

Early mining technology. Except for the Chinese, who may have mined coal underground, all the early coal seams were worked from the surface, in fully exposed outcroppings. In the later Middle Ages, however, exhaustion of outcrop coal in many places forced a change from surface to underground, or shaft, mining. Early shaft mines were little more than wells widened as much as miners dared in the face of danger of collapse. Shafts were sunk on high ground, with adits—near-horizontal tunnels—for drainage driven into the side of the hill. In England, some shallow

mine shafts were exhausted as early as the 14th century, making it necessary to go deeper and expand mining at the shaft bottoms. These remained small operations; a record of 1684 shows 70 mines near Bristol, employing 123 workers. Greater depth created many problems. First, water could no longer simply be drained away. Crude methods were devised to lift it to the surface. A bucket and chain device was first powered by men and later by horse; a continuous belt of circular plates was drawn up through a pipe. Windmills were used for pumps. But shafts had to be restricted to depths of 90 to 105 metres (300 to 350 feet) and a mining radius of 180 metres (600 feet). It was not until 1710 that the water problem was eased by Thomas Newcomen's steam atmospheric engine, which supplied a cheap and reliable power source for a vertical reciprocating lift pump.

Raising the coal itself was another problem. Manpower, operating a windlass, was replaced by horsepower; and as the shafts went deeper, more horses were added. At Whitehaven in 1801, coal was hoisted 180 metres (600 feet) by four horses at the rate of 42–44 metric tons (46–48 tons) in nine hours. The introduction of the steam engine to hoist coal was a major turning point for the industry. Small steam-powered windlasses were successfully tried out about 1770. About 1840, the first cage was used to hoist the loaded car; and from 1840 onward, advances in coal-mining techniques were rapid.

Early European miners wedged coal out of the seam or broke it loose with a pick. After explosives were introduced, it was still necessary to undercut the coal seam with hand tools. The advent of steam, compressed air, and electricity brought relief from this hard, dangerous work. In 1868, after almost 100 years of trial and error, a commercially successful revolving-wheel cutter for undercutting the coal seam was introduced in England. This first powered cutting tool was soon improved by introduction of compressed air as a power source in place of steam. Later, electricity was used. The longwall cutter was introduced in 1891. Originally driven by compressed air, and later electrified, it could begin at one end of a long face (the vertical, exposed cross section of a seam of coal) and cut continuously to the other.

Early mining methods

Early transportation methods. In the first shaft mines, coal was loaded into baskets that were carried on the backs of men or women, or loaded on wooden sledges or trams that were then pushed or hauled through the main haulage roadway to the shaft bottom to be hung on hoisting ropes or chains. In drift and slope mines, the coal was brought directly to the surface by these and similar methods. Sledges were pulled first by men and later by animals, including mules, horses, oxen, and even dogs and goats.

Steam locomotives designed by Richard Trevithick were used in the fields of South Wales and Tyne, and later in Pennsylvania and West Virginia, but they created too much smoke. Compressed-air locomotives, which appeared in the 1880s, proved expensive to operate. Electric locomotives, introduced in 1887, rapidly became popular; but mules and horses were still working in some mines as late as the 1940s.

UNDERGROUND COAL MINING

Methods of gaining access to coal. Several methods, drift, slope, and shaft (or various combinations thereof), are used to gain access to coal located too deep underground for surface mining (see Figure 4).

A drift is a horizontal, or nearly horizontal, tunnel excavated, or dug, from a place at which a level or only slightly inclined seam crops out at the surface. The drift usually follows the seam and provides the cheapest method of access, permitting continuous transportation from the coal face to the surface by any conventional means. A drift is sometimes called an adit or crosscut, but this latter type of horizontal, or near-horizontal, opening generally intersects a seam or vein of ore while a drift follows or parallels a seam. An adit or crosscut is sometimes tunnelled into a mountainside or hillside in order to intersect an inclined coal vein. Crosscuts along different levels of a mountainside are often interconnected by vertical tunnels, called raises if they are driven upward and winzes if they are driven downward. Crosscuts are advantageous for several reasons: water is easily drained, extracted ore is transported on a nearly level horizon, and prospecting for parallel seams of ore is made easier.

Drift

A slope is an inclined opening driven to work a seam that is not horizontal, to access a seam from the surface, or to reach between seams. Men and supplies frequently are moved on a track in the same slope opening in which a belt conveyor for the coal is located. A slope often will be dug parallel to a narrow inclined seam of ore, and crosscuts will connect the slope and seam of ore at different levels beneath the surface.

A shaft is a vertical opening driven from the surface to provide access to a seam or seams. Such an opening driven from one seam down to another is called a blind shaft. Shafts are used to mine both nearly horizontal and inclined seams of ore, and crosscuts, interconnected by raises and winzes, are commonly dug from the shaft to intersect inclined veins of ore.

Early shafts were small, multipurpose openings, lined at first with wood and later with brick. Modern shafts from the surface are normally lined with reinforced concrete, brick, or steel. Usually circular or oblong in cross section, they measure up to 7 metres (22 feet) across. If shafts are the only means of access, more than one is required at a mine to transport coal, workers, and mining equipment as well as provide ventilation. Inside the mine, the coal is dumped into a bin at the side of the shaft, below the level of the seam, and fed into a skip (a large, rectangular, cage-like container), which then is hoisted from the mine. At the Prosper Colliery in the Ruhr district of West Germany, a fully automatic system hoists more than 1,000 metric tons per hour from a depth of 1,000 metres (3,300 feet).

The shaft is used almost exclusively today to reach a deep seam. In the past, belt conveyors were used on slopes, but now, with automatic, high-speed hoisting machinery available, shafts are replacing the long slopes.

Mining techniques. *Room-and-pillar, longwall, and shortwall mining.* The room-and-pillar method, the oldest of the basic underground mining techniques, grew naturally out of the need to recover more coal as mining operations became deeper and more expensive. In this system, galleries, or tunnels, are driven into the coal seam with wider openings driven from them. Thick pillars of coal are left to support the overhead strata of earth and rock (overburden).

During the 20th century two methods of pillar recovery were developed as the demand for coal increased. The "pocket-and-fender," or " split-and-fender," method divides the pillar into several sections, and these sections are extracted in a sequence so that the interior of the pillar is mined first and the exterior mined last. The "open-end" method mines the pillar in sequential sections from the gob side or the mined-out side of the pillar. Both the pocket-and-fender and the open-end methods employ roof-support systems during the pillar recovery process.

Pillar recovery

In the longwall method, two widely spaced tunnels, 75 to 210 metres (250 to 700 feet) apart, are driven into the seam from the main haulage tunnel and are linked by a fourth tunnel. The fourth tunnel may be near the main haulage tunnel (the advancing longwall method) or may be at the ends of the widely spaced tunnels (the retreating longwall method). Coal is removed by widening the fourth tunnel. The need for support posts limits the amount of open working space next to the coal face to such an extent that cars cannot be used; conveyors must be employed, and for the first half of the 20th century longwall mining was less productive than room-and-pillar mining. The productivity of longwall mining was significantly increased by the British in the 1950s, however, when they introduced powered roof supports that can be quickly advanced as mining progresses (Figure 37).

The productivity record of the longwall system since 1950 has been very good, but longwall equipment malfunctions can cause a work stoppage in a large area of the mine. Furthermore, ground irregularities along portions of the longwall can give rise to problems with the conveyor-powered equipment. These shortcomings of the longwall method, combined with advancements in con-

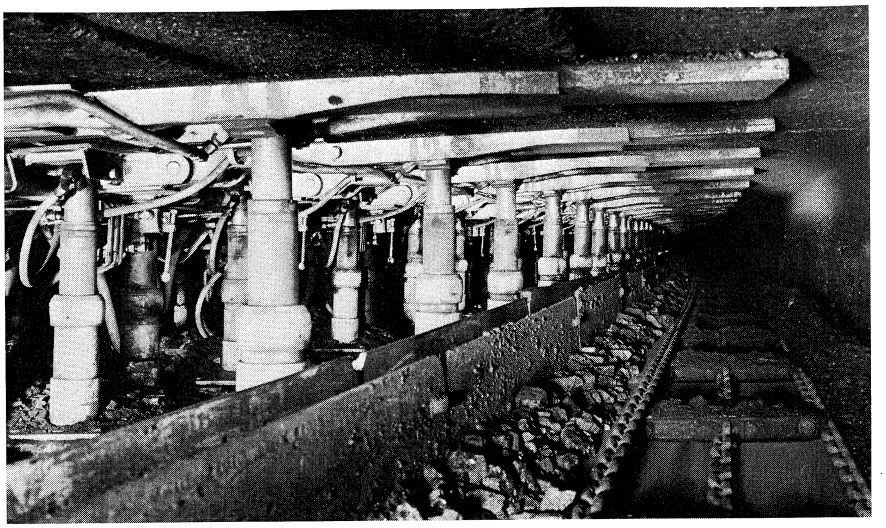


Figure 37: Longwall face with powered roof supports and segmented conveyor in a coal mine in Nottinghamshire, England.
By courtesy of Dowty Mining Equipment Limited, Gloucestershire, England

tinuous mining technology (see below *Mechanization and automation: Continuous mining*), led to the development of a derivative method, the shortwall system, though it might be pointed out that earlier shortwall systems had evolved from the room-and-pillar method, which used conventional or cyclical mining techniques.

The configuration of the shortwall system resembles that of the longwall except for the distance between the entry tunnels, 30 to 60 metres (100 to 200 feet); and the length of the entry tunnels from the main haulage tunnel to the face of the coal seam, a 120- to 910-metre (400- to 3,000-foot) stretch compared to the length of 300 to 1,500 metres (1,000 to 5,000 feet) for the longwall. Unlike the longwall method, the shortwall system employs the retreating face method exclusively; the intersecting fourth tunnel is therefore located at the end of the entry tunnels.

Coal is extracted by a self-propelled continuous miner from the mining face and transported by rubber-tired electric trucks called shuttle carts from the point of extraction to one of the entry tunnels, where the coal is then loaded onto a panel belt conveyor for transport to the main haulage tunnel. The roof-support system of the shortwall method is similar to that of the longwall method except for the extensive roof canopy, which provides an area free of support posts and wide enough for the continuous miner to operate. In both the longwall and shortwall methods, coal is extracted by widening the fourth tunnel, but the means used to widen the tunnel differs in each system. Longwall-mining machinery extracts coal from the mining face in parallel slices ranging from five to 75 centimetres (two to 30 inches) wide by conveyor-powered shearing or plowing devices pulled along the wall the entire distance between the entry tunnels. In the shortwall method, on the other hand, the continuous miner extracts coal by making cuts two to three metres (six to 10 feet) deep perpendicular to the mining surface.

Pressure considerations

In choosing between these methods, a very important consideration is pressure. Because internal stresses increase with depth and density of rock, underground workings become progressively more hazardous as they go deeper. The zones of depth may be grouped into three areas of rock pressure.

In the first zone (the exact depth of which varies because of different types of overlying strata), rock pressure presents no problem if mining dimensions are kept within established limits. Most of the coal fields in the U.S., New South Wales, South Africa, and the Lorraine region of France lie within this zone. Room-and-pillar mining is practicable, and many strata permit use of the longwall and shortwall methods; the choice is a matter of economics.

In the second zone, pressures are greater; regardless of precautions, collapses occur, though damage can be limited by providing adequate support. Most of the coalfields of the United Kingdom and the Soviet Union now being worked (almost entirely by the longwall method) lie in this zone.

In the third zone of depth are coalfields in the Ruhr Valley, The Netherlands, and Belgium. At these depths, the longwall method is employed predominantly.

Mining plans. The idea of planning a mining operation developed quite early. The panel system, in which the mine was divided to isolate problems, was in use in England by 1795. Multiple entries replaced the single entry for improved ventilation and safety. Planned backfilling became a practice in the anthracite mines toward the end of the 19th century. At the beginning of the 20th century, the replacement of man and animal power by mechanization, followed by a trend toward concentration, further modified mining plans, particularly in the United States, where conditions are more adaptable to variation.

Submarine mining. Both room-and-pillar and longwall methods are used under the ocean, with certain restrictions based on the enormous weight of the water over the mine. In Canada, for example, for total (longwall) undersea extraction, a minimum of 100 feet of cover (distance between the seafloor and the top of the seam) per foot of thickness of seam is required by law. In practice, however, longwall extraction is not conducted under less than 300 metres (1,000 feet) of cover. This is considered safe, even if parallel seams have already been worked.

Cover requirements

In the United Kingdom, 60 metres (197 feet) of cover is required for undersea room-and-pillar extraction, but pillar size and roadway width are restricted according to depth and thickness of seam.

Mechanization and automation. The two principal methods of underground mechanized mining are conventional, also called "cyclic," mining and continuous mining. Hydraulic mining has been adopted, however, on a limited basis, particularly in Europe.

Conventional (cyclic) mining. Conventional mechanized mining involves the following sequence of operations: supporting the roof, cutting, drilling, blasting, and loading. After the roof above the seam has been made secure by timbering or rock bolting, normally one horizontal—and sometimes a second, vertical—slot, 15 centimetres (six inches) wide and up to 4.5 metres (15 feet) deep, is cut into the coal seam. The slots, which make room for expansion of the coal upon blasting, are cut with a mobile cutting machine, which is essentially a large chain saw. The holes for the explosive charges are bored by power drills. The coal is broken down by special explosives that produce less heat and fewer unpleasant fumes than ordinary high explosives do.

After breakage, the coal is loaded into shuttle cars or onto chain or belt conveyors. Several types of hand-loaded conveyors developed in the first three decades of the 20th

century were made obsolete by the development of the more productive mobile loader and of surface mining.

Mobile loaders are most adaptable to the room-and-pillar method of mining. The Stanley Header, the first coal-loading machine used in the U.S., was developed in England and tested in Colorado in 1888. Others were developed, but few progressed beyond the prototype stage until the Joy machine was introduced (1914), employing the gathering-arm principle and providing the pattern for future successful mobile loaders. Upon the introduction of rubber-tired shuttle cars in 1938 and, in the late 1940s, rock bolting for roof support, which provided an unobstructed working area, mobile loading rapidly supplanted track loading.

Conventional mining is associated commonly with the room-and-pillar method in which the different cyclic operations (*i.e.*, roof-support, cutting, drilling, blasting, and loading) are conducted simultaneously in different rooms. Once an operation has been completed in one room the next operation in the cycle is undertaken, and the workers and equipment used in the completed task, such as drilling, are transported to the next room that requires this operation. This maximizes the productivity of workers and conventional equipment and causes minimal lags in room-and-pillar coal production.

Continuous mining. Introduced during the late 1940s, continuous mining began to replace the cyclic operations used in the room-and-pillar method. Continuous (noncyclic) mining implied that no break or lag time occurred in the mining process; drilling and blasting were eliminated; and a single machine, the continuous miner, broke off the coal from the seam and transferred it back to the haulage system. The Joy Ripper (1948) was the first continuous miner applicable in the room-and-pillar method. In 1950 the Jeffrey Colmol, a boring machine adaptable to thin seams, was placed on the market. Many such machines were installed underground, some producing more than 10 tons per minute. Their use reduced the number of men required in a mining crew and eliminated the expense of explosives and machines for cutting and drilling the coal.

Loebbe's plow system

One continuous longwall mining system involving the "plow" was developed near the end of World War II by Wilhelm Loebbe of Germany. Pulled across the face of the coal and guided by a pipe on the face side of a segmented conveyor, it carved a gash five to 15 centimetres (two to six inches) wide along the bottom of the seam. The hardness of the coal determined the depth of the cut. The plow was forced against the coal face by the conveyor, which in turn was under uniform pressure from equally spaced compressed-air rams. The conveyor snaked against the face behind the advancing plow to catch the coal that chipped off from above the gash. Substantially reducing the labour required at the coal face except that needed to install roof support, the Loebbe system quickly became popular in Germany, France, and the Low Countries. U.S. installations met with varying degrees of success. Competition from other, more productive continuous mining devices and from strip mining temporarily made the plow outmoded in the United States toward the end of the 1950s, but development of powered roof supports reinstated the system by drastically reducing the need for roof-control crews. Improvements in the design of the plow increased the range of applications of the system. In 1968 a new drive design increased the plow and conveyor speed, resulting in record production capacity.

Though the Loebbe plow had limited application in British mines, the power-advanced segmented conveyor became a fundamental part of equipment used at the coal face in continuous mining. In 1952 a simple continuous

Shearer

loader, called a shearer, was introduced (Figure 38). The shearer is a longwall cutting machine that is pulled along the face astride the conveyor, bearing a series of disks with picks on their perimeters, on a shaft perpendicular to the face. These revolving disks cut a slice from the coal face as the machine is pulled along. A plow behind the machine cleans up any coal dropped between the face and the conveyor.

Shearer installations increased rapidly in the United Kingdom, in step with the adoption of powered roof supports. More than 300 were in use in 1958, when a novel combination boring-and-cutting machine called a trepanner was introduced. Though the trepanner was slower and harder to control, it produced a greater proportion of lump rather than fine coal, desirable in the U.K. The basic flexibility of the shearer design permitted modification, however, and in 1961 installation of shearers resumed. Later, hybrid machines called trepan shearers were introduced.

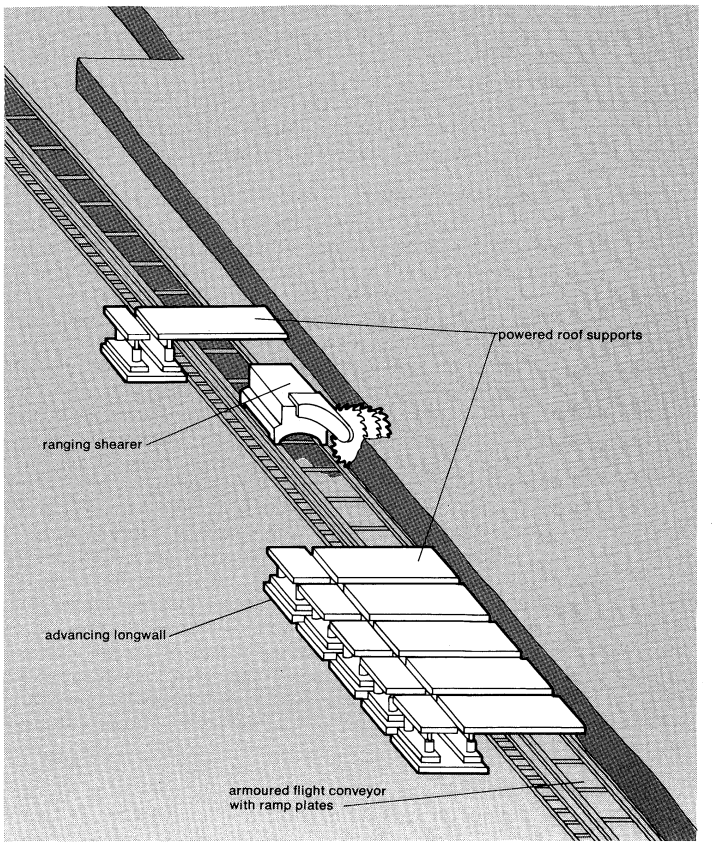


Figure 38: Section of longwall mining as practiced in the U.K., showing the ranging shearer cutting a horizontal seam and loading broken coal on a conveyor.

Hydraulic mining. The hydraulic mining of coal involves the erosion of ore by high-pressure water. The broken coal is crushed and mixed with water and then transported from the seam face to the surface by water, utilizing gravity whenever possible. Four basic types of hydraulic coal mining are used; two types combine hydraulic forces with conventional mining methods; and two types use primarily hydraulic forces for all mining procedures.

Coal sometimes is broken from the face of the seam by conventional methods and then flushed from the face and transported by moving water to the surface. Water jets sometimes are sometimes used to pre-loosen coal prior to conventional extraction. This type of hydraulic mining is undertaken in hard and impermeable coal seams, such as anthracite.

Conventional methods also are used in conjunction with hydraulic pressure to extract coal from hard seams. In this situation either the mine is designed to hydraulically flush and transport the ore to the surface, or shuttle cars and conveyors can be used.

The most widely used form of hydraulic mining involves the cutting and breaking of coal from the seam face with water jets, followed by the flushing of coal into flumes by high-pressure water, and hydraulic transport of the coal ore to the surface.

The fourth type of hydraulic coal mining not only uses high-pressure water to extract and transport coal, but it also utilizes hydraulic power to operate other mining equipment such as fans, lights, and ventilation machinery. Although this type is less widespread, it may be considered pure hydraulic mining.

An integral part of a hydraulic-mining facility is the surface plant, in which several operations take place, such as the dewatering of the extracted coal and the treatment and recycling of water. The technology of hydraulic coal mining in an underground environment is less developed than the conventional and continuous methods, and the majority of the hydraulic mining operations are located in Europe.

Underground hauling. In underground mechanized mining, shuttle cars or conveyors transfer the coal from the production area to the main haulage system. A conveyor belt was successfully used in an anthracite mine in central Pennsylvania in 1924 to carry coal from a group of room conveyors to a string of cars at the mine entry. In the early 1930s, such a relay system, with a continuous car feeder, provided uninterrupted coal transportation. The use of these moving belts gained acceptance gradually, and the first mine completely equipped with conveyors began operating in the anthracite field in 1933. During the 1950s and 1960s, relay belts almost completely replaced rail cars for intermediate transportation in the mines.

Use of the conveyor belt

Battery-powered, rubber-tired cars, designed to carry coal from the loading machine to the elevator, were first used successfully in northern West Virginia in 1938. Later, shuttle cars were designed to permit direct discharge of coal onto a belt conveyor or into a quick-discharge belt feeder. With shuttle cars came complete off-track face equipment, supplied with electric power by trailing cables. Shuttle cars also were gradually changed over to electric cables in place of batteries, reducing manoeuvrability but increasing payloads. In the late 1960s, convertible battery-cable cars appeared.

The trend toward larger capacity had to be balanced against awkwardness. In thin seams, the limitation on car size led to direct loading onto the relay belt conveyor by way of a bridge conveyor or conveyors, with excellent results. Beginning in the late 1960s, this method was tried in thicker seams, with substantial gains in production. In Australia, where conditions are similar but the coal seams are thicker, a large proportion of the coal is hauled from the face by shuttle cars.

Moving men and supplies

The transportation of men and supplies must be independent of coal transportation, to prevent interruption of movement of the coal. Cars on tracks are normally used in most countries, though in the U.S. there is a trend toward trackless transportation by battery-powered tractors and rubber-tired cars. In some instances, the transfer of supplies from rail-borne cars to trackless equipment is eliminated by the use of rubber-tired cars that can operate on as well as off the tracks. In deep mines the floor of the mined-out area (called the gateway) is subject to heaving, and tracks usually are laid to the face; if the gateways are trackless, supplies may be carried by monorail, while men walk. In the horizontal, thicker seams of South African mines, men move about the mine on bicycles.

Completely trackless mining has expanded steadily, mostly in mines with slope access. Coal is transported by belt from the seam face to the surface, and men and supplies are hauled from the slope bottom by electric tractors and rubber-tired cars. In some instances, to avoid transfer delays at the slope bottom, the tractor-pulled cars are brought directly to and from the surface by hoists. Many small drift mines in the U.S. have no track underground. Battery tractors haul trains of rubber-tired cars to and from the working face. Underground coal in the United Kingdom is transported almost entirely by belt conveyor, and conveyor transport is increasing elsewhere. In the U.S., where conditions permit high-capacity track haulage, the proportion of coal moved by belt conveyor is smaller.

Special considerations. *Powered roof supports.* The necessity of laborious and slow operations of shoring up the roof of the mine prevented longwall mining from realizing its potential until the 1950s. The powered, self-advancing roof supports introduced by the British provide a prop-free area—between the coal face and the first row of jacks—for equipment and a canopied pathway between the first and second rows of jacks. Individually or in groups, these supports, attached to the conveyor, are hydraulically lowered, advanced, and reset against the roof.

The introduction of powered supports was followed by rapid adaptation of their design to the many mining conditions in the United Kingdom and continental Europe. The large number of designs made it possible for the United Kingdom, in the short period of 15 years, to equip 760 faces with powered supports.

Rock bolting. The origin of rock bolting (or roof bolting) to support and reinforce the roof and ribs, or walls, of coal-mine openings is obscure, but the technique was in limited use in Europe and the U.S. before World War I. In this technique, steel rods are installed in holes drilled into exposed strata and anchored under tension (Figure 39).

Roof bolting is intended (1) to tie weak strata to overlying strong strata; (2) to reinforce horizontal strata; or (3) to combine suspension and reinforcement. To make the system effective, bolts must be installed as soon as possible after the roof and ribs have been exposed.

Ventilation. The presence of noxious and flammable gases caused miners to recognize the critical importance

By courtesy of Inland Steel Company, Sesser, Ill.; photograph, Louis C. Pavledes

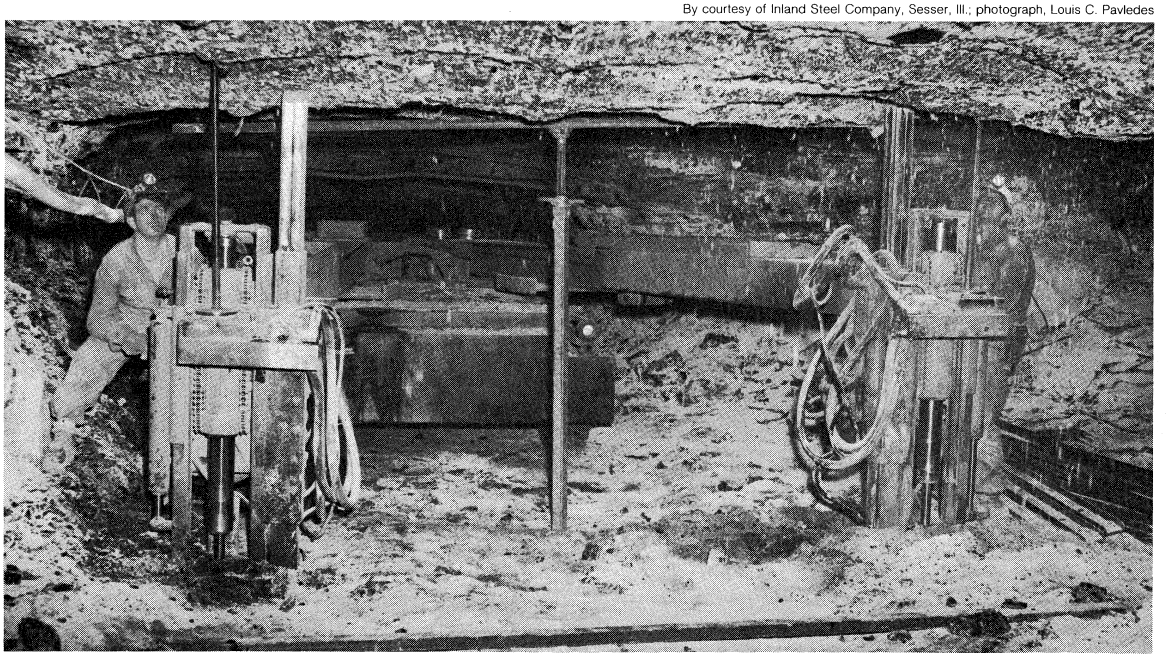


Figure 39: Rotary roof bolter installing bolts that support the roof.

Natural ventilation

of ventilation in coal mines from the earliest days (Figure 40). Natural ventilation was afforded by level drainage tunnels driven from the sloping surface to connect with the shaft. Surface stacks above the shaft increased the efficiency of ventilation; their use continued in small mines until the early 20th century. The most reliable method, before the introduction of fans, was the use of a furnace at the shaft bottom or on the surface. Despite the hazard of fire and explosion there were still a large number of furnaces operating, at least in nongassy mines, in the early 20th century.

From *Coal Age Magazine* (October, 1936); McGraw-Hill

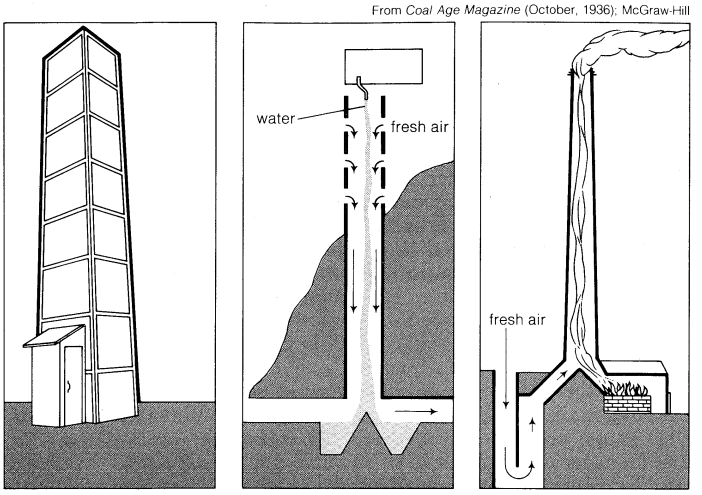


Figure 40: *Methods of ventilation.*
(A) Ventilation stack. (B) Water trompe; air is drawn into the mine shaft by falling water. (C) Furnace ventilation.

Open-flame illumination, however, was a much commoner cause of explosions until the introduction of the Davy safety lamp (about 1815), in which the flame is enclosed in a double layer of wire gauze that prevents ignition of flammable gases in the air of the mine. Presence of strong air currents, however, made even the Davy lamp unsafe.

Rotary ventilating fans were introduced in mines in the 18th century. Originally of wood and powered by steam, they were improved throughout the 19th and 20th centuries by the introduction of steel blades, electric power, and aerodynamically efficient shapes for the blades. The practice of making highly accurate airflow measurements was initiated to evaluate the effectiveness of ventilating systems. The quality of the atmospheric environment in coal mines is now ensured by supplementing ventilation with automatic controls of the temperature and the humidity.

Introduction of fans

In small mines, in the past, methane that accumulated near the roof was often cleared, before the working shift, by burning with lighted candles mounted on poles. Portable electrical instruments for detecting gas were developed before the middle of the 20th century. More recently, methane detectors, which give a flashing signal or sound an alarm when the methane content of the mine atmosphere reaches a certain value (0.25–3 percent), have been introduced. Monitoring systems now can keep a constant check on fan operation and the mine's atmosphere.

The rapid advances in longwall mining made since the mid-1950s have drawn attention to more effective use of "bleeder systems," which drain methane from large mined and caved-in areas, and other advanced ventilation techniques. In the United Kingdom, gas removed from the mines is often collected and sold.

Research. Continuous mining, especially by the longwall method, has created many problems that have led to intensive research by government, the coal industry, and machinery manufacturers. Subjects of research include instrumentation, controls, environmental conditions, suitable fluids and metals, communications, improvement of machinery, coal-sensing and steering of machines, ventilation control, and development of supports.

Mining research by use of models has been carried on extensively in West Germany and the Soviet Union since the late 1950s. Australia has had model equipment similar to that used in West Germany since 1966. The United Kingdom, Canada, and the United States are participating in similar investigations.

In the late 20th century, computers began to play a vital role in underground mining and research. Computers were initially employed by the coal-mining industry to handle accounting and payroll. Computer technology, however, was later extended to many areas of underground mining operations. Computer hardware and software were designed to simulate coal-mining procedures, to prepare production schedules, to estimate coal reserves, and to monitor mining operations. The development of computer graphics and mapping has enabled mine workers to produce detailed pictorial surveys of underground facilities. In addition, advances in telecommunications have made it possible to link several mines into a network for production monitoring by large mainframe computers.

SURFACE COAL MINING

The size of the property, the thickness and slope of the coal seam, and the type of overburden greatly influence the size and type of surface operation. The basic determinant of economic feasibility is the ratio of depth of overburden to thickness of coal. This varies from 1:8 in thick brown coal deposits of Australia to 30:1 in some U.S. strip mines.

Open-pit mining. Open-pit (also called opencut, or opencast) mining is used to remove coal from thick veins covered by little overburden. When open-pit operations begin, the valuable topsoil is carefully scraped away; after mining and backfilling, it is redistributed over the surface for resumption of agriculture. Two specific operations under entirely different conditions illustrate the open-pit method.

The Yallourn Mine, Australia. One of the largest single deposits of brown coal in the world, containing nearly 13,000,000,000 tons of coal recoverable by present methods, lies in the Latrobe Valley, east of Melbourne, Australia. The coal seams are approximately horizontal and from 30 to more than 90 metres (100 to 300 feet) thick. In some places several superimposed seams, up to 264 metres (880 feet) in thickness, are broken only by thin layers of waste. A clay and gravel overburden varies from 12 to 27 metres (40 to 90 feet) in depth. Overburden and coal are excavated by large, mobile, continuous bucket-wheel excavators that operate on benches within the coal seam at levels depending upon the capacity of the machine. These remove the material from the working faces with large, toothed buckets mounted on a wheel. The bucket-wheel excavator cuts 21 metres (70 feet) above and 8 metres (25 feet) below the crawler tracks that transport the machine. Similar machines with buckets mounted on an endless chain are called bucket-chain excavators. They reach 24 metres (80 feet) above and 26 metres (87 feet) below the crawlers. To prevent fires, water is sprayed on the coal faces during summer months.

Bucket-wheel excavators

The Garsdorf Mine, West Germany. The Garsdorf property, located in the Rhineland deposit in the state of Nordrhein-Westfalen, contains a 1,000,000,000-ton reserve of lignite of Miocene age (about 7,000,000 to 26,000,000 years old). Broken by faulting into offset blocks, the 15- to 60-metre- (50- to 200-foot-) thick seam dips northward to a cover depth of 240 metres (800 feet). The overburden consists of water-bearing strata of gravels, sands, and clays. The open-pit technique was chosen after efforts, made before 1950, to deep-mine the deposit were defeated by water and other problems. The Garsdorf pit is expected to reach a depth of 300 metres (1,000 feet), with ratio of five overburden to one coal. Though this ratio is not exceptional, most of the overburden must be transported to a distant point, which makes disposal a serious problem. Large-capacity equipment is employed to keep down costs. Because the seam is faulted and because it is not horizontal, lignite and waste must be handled alternately as the machinery travels along a level path.

In open-pit work, the slope of the walls must be carefully considered; the slopes at the bottom of a deep pit cannot be as steep as those near the surface, because the

forces causing collapse increase with depth. Mined-out areas must be promptly backfilled; water is drained from the next lower level to assure slope stability. At Garsdorf, draining is accomplished with submersible pumps.

Area strip mining. Area strip mining is applicable to relatively flat terrain and a nearly horizontal seam. A trench, or box cut, is made from one end of the area to the other, and the coal is removed. The overburden from each succeeding parallel cut is placed in the previous cut. In this manner, the final cut before grading is the only excavated area; the remainder of the property is shaped into a series of rounded ridges.

Box cut

Area strip mining is practiced widely in the United States (Figure 41). The first U.S. strip mining was carried out near Danville, Illinois, in 1866. Overburden, removed by horse-drawn plows and scrapers, was hauled away in carts and wheelbarrows. Steam-powered shovels first proved successful in the Mission field of Illinois in 1911. Capacity of strip-mining equipment has continuously increased since then.

First U.S. strip mining

Shovels or draglines are normally used for strip mining in the United States because the earth strata are usually consolidated, requiring drilling and blasting. It would be prohibitively costly to blast consolidated strata fine enough for handling by a wheel excavator. When unconsolidated glacial deposits are found above the normal strata, however, as in Indiana, Illinois, and northern Missouri, the efficient wheel excavator is used in combination with the shovel and dragline. The determining factor in the purchase of such large equipment is productivity. An all-purpose dragline is more versatile than a shovel. On the other hand, a shovel for the same boom and carrying unit works faster.

A large operation in eastern Ohio exemplifies the equipment used in strip mining. The property had two overburden conditions, an average 24-metre (80-foot) and 31-metre (104-foot) cover of consolidated strata and topsoil. Basically, the machines chosen are the same; for the first overburden condition, a 51-metre (170-foot) boom with a 99-cubic-metre (130-cubic-yard) capacity; for the second, a 60-metre (200-foot) boom with a 76-cubic-metre (100-cubic-yard) capacity.

The width of coal uncovered is determined by the angle of repose, or natural slope, of the loose material; for the 24-metre cover area, a width of 39 metres (130 feet) is used. The large shovel is followed by a 7.2-cubic-metre (9.5-cubic-yard) coal-loading shovel. The operating cycle consists of: (1) drilling and blasting, (2) excavation, (3) coal loading, and (4) restoration. Allied equipment for each uncovering unit includes a bulldozer on the highwall (the ledge above the mined area) to cut a road for the drilling rig; a blast-hole drill; a bulldozer in the pit to push loose debris into heaps for the shovel; a road grader or small bulldozer (and, if necessary to meet ash requirements of the mined coal, a powered-type street sweeper) to remove loose soil and rock from the exposed surface of the coal; a 7.2-cubic-metre shovel to load coal; power-cable handling equipment with a hydraulically operated reel for a 7,200-volt cable for the large shovel and 4,100-volt cable for the small shovel; and 50- and 120-ton trucks.

Operating cycle

The first requirement of preparation is thorough fragmentation of rock overburden by blasting to allow steady operation of the shovel with the bucket filled to capacity. The trend is toward larger drills and blasting holes of larger diameter. The drilling pattern is staggered, and hole spacing depends upon the size of the hole, as uniform fragmentation requires the same blasting ratio (the volume of explosive needed to fragment a given volume of rock).

An indication of the trend in equipment is illustrated by the change since 1940, when a 4.6-cubic-metre (six-cubic-yard) shovel having a capacity of 228 cubic metres (300 cubic yards) of fragmented material per hour was used in starting the Ohio operation: the more recent 99-cubic-metre shovel has a capacity of 5,320 cubic metres (7,000 cubic yards) per hour. By the late 1970s, bucket-wheel excavators capable of moving 8,300 cubic metres (10,900 cubic yards) per hour were available.

In response to serious abuses in the past by some mining firms, most U.S. state laws require bonding to ensure complete compliance with regulations concerned with restoring a strip-mined area. After backfilling, limestone is deposited next to the top with sufficient select topsoil on the surface to promote prompt growth of vegetation. Piles of overburden are graded to a gentle, rolling contour and planted with grasses and trees to prevent erosion.

Contour strip mining. Contour strip mining is practiced where deposits occur in high, rolling, or mountainous terrain and the overburden ratio is high. About one-half of the U.S. surface mining has been accomplished by this method. Removal of overburden starts at the outcrop,

By courtesy of the National Coal Association, Washington, D.C.

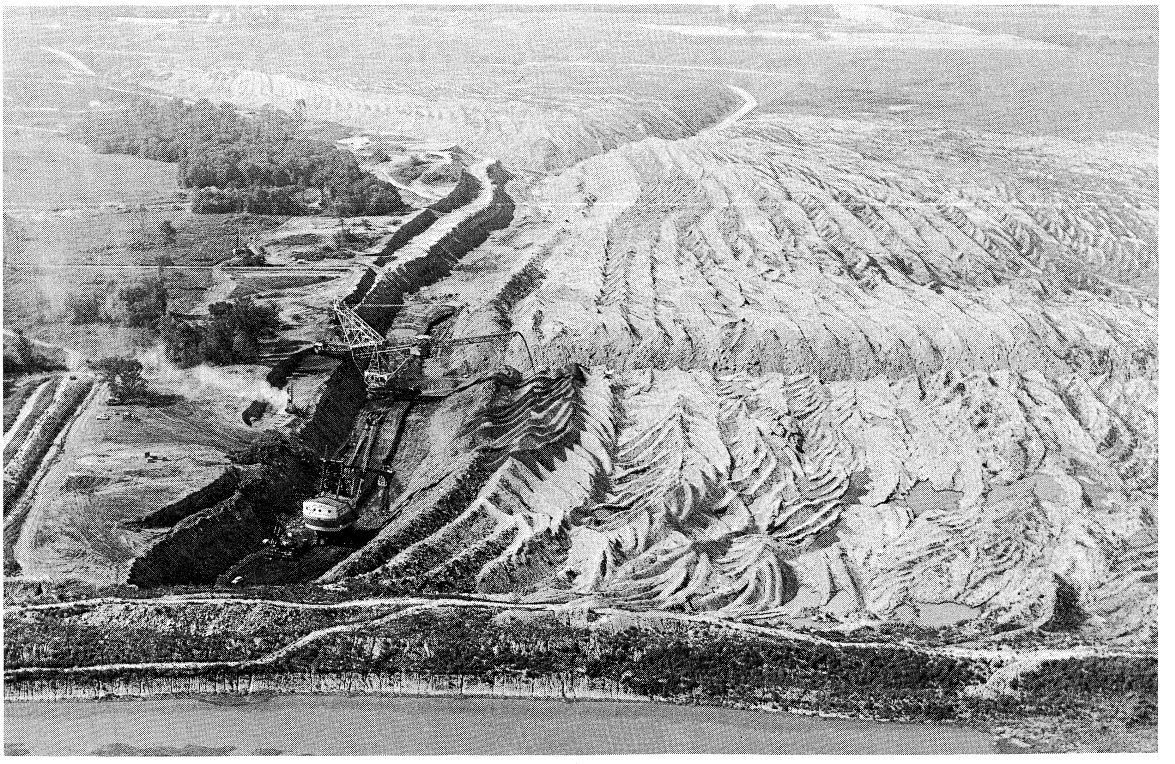


Figure 41: Strip mining near Freeburg, Illinois.

following along the hillside, creating a shelf with a slope on the side below at the angle of repose of the material. The uphill side of the cut is an almost vertical highwall. As succeeding cuts are made, the overburden is piled on the previous cut area until the economical overburden height is reached.

Generally, the operating procedure is similar to that of area strip mining, except that equipment is normally smaller because of the need for adaptability to varying conditions and because it must frequently be moved from one site to another as the deposits are exhausted. In the Appalachian area, where conditions normally require contour strip mining, several coal seams that occur close together may be mined simultaneously.

Restoration of mined areas

Restoration is more difficult and more expensive in rolling terrain. No single set of procedures is applicable to all conditions. The ideal is total backfill, which is practical in some of the rolling hills of Pennsylvania. When restoration is completed, no signs of mining are evident. Restoration in the steep hills of West Virginia has created slightly sloping benches now useful for grazing and agriculture in places where previously only difficult-to-recover timber could be grown.

Regardless of the restoration plan used, it is necessary to cover the remaining coal and other toxic material, establish noneroding gradients on the benches, control the movement of water to prevent saturation of the soil and erosion of the spoil banks, and provide a good vegetation cover. These steps can be taken most effectively when they are planned as part of the whole mining operation.

Auger mining. Recovering coal by boring into a seam located at the base of a highwall is normally one of the cheapest mining systems, but it is limited to horizontal or lightly pitching seams that have been exposed by geologic erosion, such as the coalfields of the Appalachian area. Recovery is approximately 50 percent.

Augering is usually associated with contour strip mining, recovering coal for a limited depth beyond the point at which the overburden ratio has reached its economical limit for stripping. It is also used to recover coal between the outcrop and the depth needed for safe underground mining.

The only preparation required for auger mining is establishment of a well-drained pit or bench of sufficient width for the operation and a highwall clear of loose material. Where the operation closely follows stripping, and care is taken on the final highwall blasting, little or no additional work is necessary. Only a bulldozer is needed in addition to the auger equipment and trucks. Where auger mining is conducted independently of stripping, a small shovel may be needed to prepare the bench and highwall.

Multiple augers

Power-driven auger heads range from 40 to 245 centimetres (16 to 96 inches) in diameter and are used singly or in groups of two, three, four, or five. Multiple units can be mounted side by side or, where two are used, one above the other for selective mining above and below a band of impurity. Multiple augers may overlap to leave no coal between them. With these, production is increased, and holes can be held straighter. Production varies with so many factors that generalizing has little meaning. For example, one triple-head auger in West Virginia averages 225 tons per shift, while the same type of device in Ohio averages 324 tons.

TRANSPORTATION OF COAL

The cost of transporting coal from the mine to the consumer is an important part of the total cost; when exported, it is often a major portion. Coal-conveying facilities are competitive; there is a continuous effort to stabilize delivered costs. Current trends are toward larger unit capacities, more automation, and new techniques.

By vehicle. *Railroads.* Coal and railways, among other things, were the base for industrialization of Europe and the U.S. Furthermore, coal has been one of the railroads' most important cargo since the middle of the 19th century. Thus, the coal-mining industry and the railways have found themselves sharing common problems.

Rail transportation techniques improved with growing tonnage, but by the 1950s, the coal industry was looking to alternative means of transportation for further cost savings.

Unit train operations

In the 1950s a plan originated in France for assembling whole trainloads of coal destined for a single major customer. Unit train operations, with rapid, continuous loading and unloading, are now common in Europe and North America. Trains with large loads can be hauled by several locomotives scattered through the train. On a western Canadian line, 13 locomotives, controlled by radio, pulled 105 cars holding 11,000 tons. A U.S. line assembled a 500-car train, more than six kilometres (four miles) long, to carry almost 25,000 tons of coal over a distance of 251 kilometres (157 miles). In Europe, where distances are shorter and interruptions for passenger train traffic more frequent, shorter, faster trains have been favoured.

A number of trains operate continuously, without ever uncoupling, moving even during loading and unloading. In the United Kingdom, this practice is called "merry-go-round." One U.S. short haul of 24 kilometres (15 miles) is completely automatic, handled from a control centre. In Europe, many unit trains operate on the "block system," in which coal is carried from several mines for a single customer or from a single mine for several consumers in a single region.

Water carriers. New approaches are also evident in transportation of coal by means of inland waterways. For example, conveyors are employed to carry coal from mines to the rivers or canals. Transportation over integrated rail-and-water routes is increasing; *e.g.,* coal is moved by rail to docks on the Great Lakes and in the Tidewater region of Virginia. On the Great Lakes, coal is conveyed by belts to the ships, at rates of about 6,000 tons per hour. At piers accommodating vessels, the loading rate is even higher.

Integrated rail-and-water routes

Motor vehicles. Shipment of coal by truck, while substantial in volume, is generally more expensive than shipment by rail or barge; restrictions on weight also contribute to making highway movement of coal less significant than movement by rail or water in most places.

Trucks are most useful in off-the-highway haulage and in delivery to small consumers. Coal for power stations, however, may be trucked substantial distances if a return load of fly ash is available.

By wire. Extra-high voltage (EHV) transmission of electric power (sometimes in excess of 500,000 volts) from generating stations located at or near the mine mouth is rapidly increasing as a means of transporting energy to distant cities. "Coal by wire" is low in cost, is unimpeded by difficult terrain, allows location of power plants away from populated areas, and can be interconnected with other power lines. Disadvantages of the system include power-line losses and the need for purchasing costly rights-of-way for transmission lines in populated areas. Transport of coal by wire over short distances, however, has been accepted practice in some countries for years. The Yallourn Mine in Australia began power transmission in 1925. Frimmersdorf, West Germany, has a very large mine-mouth generating station and steam power plant.

Mine-mouth generating stations

Mine-mouth generating stations have developed into huge complexes in the U.S., both in established coal-producing areas in the Appalachians and in newer fields in western states. Plants in the Chestnut Ridge generating area of western Pennsylvania, completed in 1971, use about 16,000,000 tons of bituminous coal per year. In the western U.S., extensive reserves, isolated because of terrain and distance from markets, are being exploited by utilizing the coal-by-wire method to transmit energy from coalfields in New Mexico, Arizona, and Wyoming to serve the expanding population of the West, particularly Southern California.

It is possible also to burn, or heat, the coal in the ground, collect the resulting gases, burn the gas to generate electricity, and transmit the electrical energy by wire.

By pipeline. The long-distance coal-slurry pipeline was the outcome of a research evaluation of alternative transportation methods in an effort to reduce the cost of taking coal to the consumer. After five years of research, including pilot-plant operation, a pipeline 25 centimetres (10 inches) in diameter, buried 1.05 metres (3.5 feet) underground, was placed in service in 1957 in Ohio, carrying pulverized

Table 9: Comparative Data of Coal Mining in Selected Countries

item	Australia	Canada	U.S.	West Germany	U.K.	South Africa	Soviet Union	Japan	India
Geologic age*									
Anthracite	P	Cr	Pa–Te	none	C	C–Tr	C	none	Pa–C
Bituminous	Tr or P	Cr–C–Te	C–Te	C–Cr–Te	C	C–Tr	C–Ju	Tr–Ju–Te	P–C–Te
Lignite or brown	Te	Cr–Te	Cr–Te	Te–Pl	none	none	Te–Me	Te	Te
Tectonic movement									
Folding (pitch)	90% < 10°, 1% > 20°	heavy in areas	limited areas	22% > 20°	heavy in areas	<5°	heavy in areas	heavy	heavy†
Faulting	limited	in the west	limited areas	extensive	extensive	none	prevalent	prevalent	prevalent
Igneous intrusion	extensive	...	very limited areas	limited	limited	heavy in areas	...	heavy in areas	heavy in areas
Depth of overburden									
Average (ft)	520	1,700	<500‡	2,400	1,468	<300	1,130	1,510	490
Maximum (ft)	1,200	2,500	...	10% > 3,300	3,765	1,430	3,600	3,025	1,970
Seam thickness									
Average (in.)	102	96	63.6§	62.6	51.2	...	60	71.6	118
Minimum (in.)	36	60	24§	19.7	18	24	20	...	31
Maximum (in.)	216	...	7.1% > 86§	197	150	240	1,180	...	827
Underground methods									
Longwall	3.3%	32%	1.34%	99%	96.5%	2 faces	85%	87.5%	1.3%
Longwall, with powered roof supports	100%	...	100%	27.5%	760 faces	100%	549 faces	14.7%	—
Room-and-pillar (%)	96.7	68	98.65	0	3.5	>98‡	1.7, 8 incl. slicing, 3.2 shield, 2.1 other	7.1	98.7
Hand loading (%)	0.5	0	4.55	8.7	8.0‖	60	...	...	97
Mechanical loading (%)	99.5	100	95.45	90.3	92	40	2 hydraulic	...	3
Conventional (%)	23.3	0	51	0	9	96	48	...	99.63
Continuous	75.7	100	49	100	91	4	50	67	0.37
Productivity¶									
Underground	10.06	6.14	15.07♀	3.89	3.42	2.83	2.19	2.44	—
Surface	52.4	43.4	36.1♀	...	...	none	13.33	—	—
Total average	30.9	27.4	19.17♀ 18.8♂	3.44	...	2.83	2.78	2.44	0.65
Largest operation (000,000 tons)	12.0 opencut	...	5.8 strip	20.0 opencut	...	4.27 underground	5.5 opencut	...	1.19 opencut

*P–Permian; Tr–Triassic; Te–Tertiary; Cr–Cretaceous; Pa–Pennsylvanian; C–Carboniferous; Ju–Jurassic; Pl–Pleistocene; Me–Mesozoic. †Mining mostly 6–12 degrees pitch seams—access 59% shaft. ‡Estimated. §1965. ‖1969. ¶Based on tons per day for all employees to point of shipment. ♀Bituminous and lignite, 1967. ♂Bituminous, lignite, and anthracite, 1967.
Sources: U.S. Department of the Interior, Bureau of Mines, IC 8345 (August 1967); *Coal—Bituminous, Lignite and Anthracite*, Bureau of Mines Yearbook (1968 preprints); National Coal Association, *World Coal Trade* (1969); *Mining Congress Journal* (December 1969).

coal, mixed with water, a distance of 189 kilometres (108 miles) from Cadiz to a power plant on Lake Erie. During approximately six years of operation, it transported more than 7,000,000 tons of coal. The project was a complete success, technically and commercially, even though it was closed down when unit-train freight rates between the two points were established, reducing the cost per ton from $3.47 to $1.88.

In the same year a 115-kilometre, 15-centimetre (72-mile, six-inch) line for transporting gilsonite (a black, lustrous asphalt) entered service in Utah. For this project, a boiler for the use of coal in liquid form was designed and demonstrated.

These two successes proved that the technology required for the construction of cross-country coal pipelines was available. Included were the discovery of a "stabilized" coal slurry that solved the problem of solids settling out during a shutdown; the proving in of an automatic remote-control system; and developments that made obsolete the original terminals—that is the slurry preparation, the dewatering, and the drying.

Black Mesa pipeline

Early in 1971, the Black Mesa pipeline for coal slurry was put into operation. This is a 45-centimetre, 440-kilometre (18-inch, 275-mile) system designed to carry 5,000,000 tons of coal per year from the Kayenta Mine in Arizona to the Mohave steam-generating plant near Davis Dam, Nevada. Opposition from other modes of coal transport, particularly railroads, has stymied coal pipeline development in the United States. In the late 20th century, pipeline legislation, granting the right of eminent domain to pipeline projects, was opposed by the railroads and subsequently defeated in the U.S. Congress.

Coal may also be transported by pipeline in the form of gas, after underground gasification at the mine.

PRODUCTIVITY

Stimulated by competition, the coal industry throughout the world has made continuous improvement in productivity (output per man-hour) through ingenuity and technological innovation. In the U.S., productivity increased dramatically from the middle to the late 20th century, actually reducing the average cost of a ton of coal at the mine during the 20 years from 1948 to 1968. The average hourly earnings of the miners and the costs of supplies and equipment also increased rapidly.

Other countries have also made tremendous strides in productivity of underground mining, though their gains do not equal those of the U.S. Difficult geologic conditions impose restrictions in some areas, increasing the expenditure of work in moving the coal from the working face to the surface. Some geologic phenomena that have a marked effect on productivity are (1) great depth of overburden, which, through excessive pressure, increases roof and floor maintenance problems at the face and in the roadways out from the face; (2) faults that displace the seam, restricting the direction and extent of mining and sometimes requiring excavation through the rock strata to reach the continuation of an interrupted seam; (3) igneous intrusions (molten materials forced from within the Earth's crust through fissures or weak planes in the strata) that isolate coal areas or limit continuity of operation by making the seam nonuniform; (4) structural deformations of the Earth's crust that when light can be responsible for clay veins and when heavy can be responsible for steep, pitching seams that may require a series of tunnels through the rock to reach different levels.

Geologic effects on productivity

The influence of various factors on productivity is indicated in Table 9. The table reflects conditions in selected countries that for the most part have excellent mining research and development programs but that vary widely in mining conditions. The listing is in order of overall productivity, reflecting more difficult mining conditions in the same order. (M.A.E./Ed.)

Processing

Coal processing involves the conversion of coal into useful solid, gaseous, and liquid products. Most of them are

used as fuels, but others are important in the chemical, building, and road-making industries.

The general objectives of coal processing may be classified as follows: (1) to render coal suitable for various processes or purposes (*e.g.,* to make metallurgical coke or smokeless domestic fuel); (2) to convert coal into more valuable fuels (*e.g.,* briquetting of small coals, complete gasification to make fuel gas, or conversion into gasoline and fuel oil); and (3) to enlarge markets by using coal as a chemical raw material.

Although such carbonization by-products as benzol and tar have been very important in the development of the chemical industry, carbonization has rarely been carried out for the sake of these by-products alone. The manufacture of the main products, coke, smokeless fuel, and town gas, has been the main end.

DEVELOPMENT OF COAL-PROCESSING TECHNOLOGY

Coal processing probably originated in the 16th century, when it became necessary to find a source of carbon to supplement wood charcoal for making iron in blast furnaces. Consequently, processes of carbonization were devised, leading to the manufacture of coke. By 1709 coke was regularly used by Abraham Darby in his blast furnaces in Coalbrookdale, Shropshire. The original primitive carbonization processes were progressively replaced by more complicated methods in larger plants as technology advanced. Although alternative methods for making iron have been developed, the main method throughout the world still requires use of the blast furnace, and the manufacture of metallurgical coke seems likely to remain as the major example of coal processing.

Town gas

The carbonization of coal with the prime object of making town gas, with coke as a by-product, began commercially about 1800, after William Murdock had first used gas lighting in England in 1792. The practice spread rapidly in many countries until practically every town had its own gasworks; a visit to such a gasworks, often the only "chemical" plant in the neighbourhood, played a considerable part in arousing the interest of schoolchildren in chemistry. Other methods of making town gas from petroleum products and the use of natural gas, however, eliminated the need for the classic gasworks. The change was completed in most parts of the U.S. by 1950; in the U.K. the change began around 1960. Very little coal was carbonized in gasworks after 1970.

The by-products of coal carbonization are tar, benzol, and ammonia. Tar was first recovered from gasworks soon after 1800, but by-product coke ovens were not successfully operated until after 1850. The recovery of tar, and later of benzol, grew rapidly because of the requirements of the chemical industry, particularly for dyestuffs and explosives. Ammonia formed the basis of the only manufactured nitrogenous fertilizers until it was made synthetically after 1913. Later, the importance of these by-products decreased as the chemical industry more frequently turned to petroleum sources for its raw materials. Nevertheless, since the late 19th century, coal tar and benzol have provided the livelihood for three generations of chemists and have formed the basis for the growth of organic chemistry on a commercial scale.

In addition to high-quality town gas, the need for a lower grade industrial fuel gas led to "complete gasification" processes, producing no coke as a by-product. Processes for making "producer gas" from coal or coke were in use from the mid-19th century, reaching their peak in the U.K. in the period from 1930 to 1950, when several million tons per annum of coal were used in this way. Producer gas from coal was particularly valued in steelworks because of its high flame luminosity and good heating qualities. The high price of coal relative to oil and high labour costs, however, had closed practically all such plants by 1970. From about 1880 to 1920, the economics of the Mond producer-gas process, which used a large excess of steam, was helped materially by the exceptionally high yield of ammonia obtained from it.

In the 1930s, the Lurgi Company in Germany produced a complete gasification process that was particularly successful with noncoking coals. Plants were built in several countries, but they too suffered from natural gas and petroleum product competition and many had already closed by 1970. The Lurgi plant in South Africa, based on very cheap coal, makes "synthesis gas" (see below).

Much work continued in the U.S. to make town gas of high caloric value from coal, against the day when natural gas and oil become scarce and the relatively large reserves of coal have to be employed. Technically, the problem is solved, and the time is almost ripe for an economic application. In the U.K. coal is too expensive for any chance of economic success.

"Synthesis gas"

From 1920 to 1960, much effort was devoted throughout the world to devise processes for making "synthesis gas" from coal. Such gas can be used to make ammonia, methanol, and gasoline. Germany has been prominent with several processes, particularly with the more reactive brown coal and lignite. Another approach to "oil-from-coal," also pioneered in Germany in the 1930s, was hydrogenation of coal under pressure. The process, using bituminous coals, was technically very successful. It seems likely indeed that Germany could not have entered World War II without the support of such processes. Germany depended on them to supply not only ammonia (and hence explosives and fertilizers) but also most of her motor and aviation fuel. At the peak (May 1944), oil was being produced in Germany's oil-from-coal plants at the rate of about 5,500,000 tons per annum. The only oil-from-coal plants to survive after 1945, however, were in East Germany, particularly those using brown coal. The only other plant operating on bituminous coal by 1970 was in South Africa. But the U.S. is also developing oil-from-coal processes against the day when oil becomes scarce and expensive.

Work on coal processing has had profound effects on the chemical industry, both in technological developments and on the structure of the industry. The Winkler process, introduced in Germany in 1926 for the gasification of brown coal, was the first large-scale use of the technique of fluidization and later became widely used in the petroleum industry. Experience gained from the new processes and technologies has greatly influenced the direction of company growth and the spread of ideas.

Solid smokeless fuels

The development of solid smokeless fuels from coal has been an objective in many countries for centuries. Methods of cooking and house heating evolved from the use of wood and charcoal. When wood grew scarce, people searched for a substitute that they could use in their existing appliances. As coke was developed for metallurgical processes, the smaller sizes naturally became available for domestic use. With the manufacture of town gas by the carbonization of coal, large quantities of gas coke also were applied to domestic use. In the period 1920–40, much work was carried out on low-temperature carbonization processes, especially in the U.K., to make solid smokeless fuel together with oils suitable for use as gasoline and diesel oil. Not until the passing of the Clean Air Act in 1956, however, did the demand for smokeless fuels in the U.K. become great enough to warrant much expansion in production. In 1969, the production of manufactured smokeless fuels by the National Coal Board and private industry used more than 5,000,000 tons of coal; this was in addition to anthracite and gas coke. No other country has produced manufactured smokeless fuel on this scale.

The conversion of small coal and dust into larger lumps by briquetting has been of considerable importance since the early 19th century. While still important, later emphasis was on adding a further stage to make the briquettes smokeless.

Employing coal directly as a chemical raw material is of fairly recent origin, since the chemistry, like the coal itself, is complicated. But throughout the 20th century, workers in many countries have attacked coal with various chemicals, trying to convert it to a material that could then be used for a variety of purposes. They have also tried to convert it to active carbon. Although this research has uncovered some commercial applications, the quantities of coal consumed in this way have been small. Nevertheless, the work goes on, and outlets for modest quantities of coal are likely to be found in these directions.

Summing up, the processing of coal represents an important outlet for coal, in many countries exceeding 20 to 30 percent of the total market. The manufacture of metallurgical coke represents the biggest use in most countries and is likely to continue. In the U.K. and to a lesser extent in other countries, the manufacture of solid smokeless fuels is also important and likely to remain so. The amount of coal converted to town gas and industrial gases has declined rapidly in most countries, but coal may be used extensively again for making gas as natural gas and oil products become scarce and expensive. There are other minor uses for processed coal, and some new ones are being developed in the chemical, building, and road-making fields, but the quantities of coal involved are likely to remain fairly modest.

CLASSIFICATION AND EVALUATION OF COALS

The types and ranks of coal are described in the article FUELS: *Coals.* Any of these coals may be processed in one or many ways, but not all coals can be used in a particular process. Economic considerations are as important as technical considerations in selecting the coal to be used.

Clearly, coals with a high ash content are undesirable for making solid products of low ash content unless the coal can be cleaned first. The behaviour of the coal on heating (in particular, whether it becomes plastic and can make a good coke) determines its suitability for carbonization and the type of oven or retort that can be used. In general, coals with a volatile content of 15 to 35 percent, often as blends, are used to make metallurgical coke, while coals with 30 to 45 percent volatile content are used to make town gas. Coals of medium rank are used for low-temperature carbonization because they yield reactive cokes. Coals of very low or very high volatile-matter content will not form strong coke merely by heating. When such coal is not in lump form, briquetting at some stage is necessary if satisfactory metallurgical fuels and smokeless fuels are to be made from, for example, small anthracite, low-rank bituminous coal fines, or brown coal.

Any coal may be totally gasified by the appropriate process; the ash content is then a minor consideration. Any tarry material evolved during the gasification process has to be separated or gasified as well.

Large and graded coal is usually best suited for processing, because it usually is more easily cleaned and dried. Crushing is economical, but briquetting is expensive. Nevertheless, a large proportion of coal is mined as small pieces. Most commercial processes have been developed to use small coal because it is available at lower prices and overall is cheaper, even with the disadvantages noted.

CLEANING AND OTHER PREPARATORY STEPS

Mined coal contains a varying proportion of noncoal substances: for example, shales and rocks from roof and floor and shales embedded in the coal seams. Coal and other material may vary in size from 300 millimetres (12 inches) or more to dust. Modern mechanical mining methods yield a high proportion of small coal; *i.e.,* below 50 millimetres (two inches). The coal and other material with it is frequently wet. Since more than one seam is often worked at the same time, a varying mixture reaches the surface.

The function of coal preparation is to separate this mixture and prepare from the cleaned coal fractions of consistent size and quality for various markets. For example, clean large coal may be separated for domestic use, clean graded coal (25 to 75 millimetres [one to three inches]) for domestic and industrial use, clean small coal for carbonization, or a partially cleaned coal for steam raising in central power stations. Waste is discarded by being stowed underground or tipped at the surface. Separation into wanted and unwanted fractions must be conducted with maximum efficiency not only to obtain clean products but to avoid undue loss of coal to the waste. Inevitably, in practice, compromises have to be made to achieve optimum economic working.

The largest material brought up from the mine is often first crushed, sometimes after large rock is removed by hand. Subsequently, coal is separated from other materials by taking advantage of the difference in density between coal (sp gr—specific gravity—1.3 to 1.6) and other material (sp gr 1.8 to 2.4). The actual separation takes place in special washers. The upward velocity of the water can lift the coal but not the heavier material, as in the Baum jig, where a pulsating motion of the water assists the separation. Alternatively, the coal may be floated in a "dense medium"; *i.e.,* water containing fine solids of high specific gravity such as magnetite, which can be easily recovered magnetically. The smallest sizes may be treated by froth flotation. Air is blown through a suspension of the fines in water, and the coal (but not the shale particles) attach themselves to the air bubbles. A reagent may be added to help the separation, and the clean coal froth that accumulates is floated off. Water is removed from the washed products by drainage, filtration, centrifuging, or by heat drying.

Density separation

It is not always economic to clean coal for central power stations. It is also desirable to avoid adding water to the coal by trying to wash it. If the ash content is sufficiently low, the run-of-mine material is merely screened and the oversize is washed, crushed, and mixed back with the unwashed undersize.

When seams of low inherent ash content were mined by hand methods, the coal did not need washing. Such coal is still being worked, and the large sizes at least are not washed. Some of the brown coal and lignite in Germany and Australia are also clean enough, after briquetting and carbonization, for domestic and industrial use. The dirtier brown coals and lignites are generally not washed; they are used for power generation or gasification.

Rocks and shales, the waste products of coal preparation plants, sometimes find low-grade uses. Some of them go into road making, while certain shales may be treated in kilns to make lightweight aggregates for building construction, a process that takes advantage of the small amount of combustible material (coal and carbonaceous shales) present.

Rocks and shales

PYROLYSIS

When coals are heated in the absence of air (pyrolysis), changes occur as illustrated in Table 10.

Table 10: General Effects of Pyrolysis

temperature (°C)	changes
Above 100°	free water evaporates
Above 200°	combined water and carbon dioxide evolved
Above 350°	bituminous coals soften and melt to some extent: decomposition begins; tar and gas evolved
400° to 500°	most of the tar is evolved
450° to 550°	decomposition continues and residue turns solid
Above 550°	solid becomes coke; only gas now evolved
By 900°	no more gas evolved; coke only remains
Above 900°	small physical changes occur

Coals differ greatly in their behaviour on heating. Anthracite does not soften and evolves only a little gas. Low-rank coals yield much gas, carbon dioxide, and water, and give a coke of low mechanical strength. Medium-rank coals are intermediate in behaviour and yield strong cokes.

Low-temperature carbonization. Low-temperature carbonization usually describes the pyrolysis of coal to a final temperature of up to 700° C (1,300° F). The final solid product is a rather weak coke. Compared with high-temperature carbonization, this process yields more oil and tar but less gas. Between 1920 and 1940, low-temperature carbonization was applied extensively. The objective then was partly to obtain the high yield of tar and oils. After 1945, low-temperature carbonization was carried out commercially only to obtain the solid residue for use as a domestic, smokeless fuel. Nevertheless, the tar and oils formed simultaneously are processed further and sometimes add significantly to the economics of the process.

Processes. More than 100 processes have been developed but only a few have operated successfuly. Some examples follow: (1) noncaking lump coal or briquettes are

heated in internally heated retorts (*e.g.*, the Lurgi-Spülgas and Rexco processes); (2) small caking coal is heated in externally heated tubes, from which the agglomerate of coke is eventually discharged (*e.g.*, the Coalite process); (3) small caking coal is heated in a rotary kiln, the plastic char so formed balling up to form lumps of coke (*e.g.*, the Disco process); (4) small coal is heated in a fluidized bed to form a smokeless char that is compacted, while still hot and plastic, in a press with or without the addition of a binder (*e.g.*, the Char Briquetting process of the National Coal Board [U.K.]); (5) low-volatile coal is briquetted with some substance such as pitch and then carbonized (*e.g.*, the Phurnacite and Syntraciet processes).

Comparison of processes

The Rexco process uses a simple plant of low capital cost, but suitable lump coal is more expensive than others and not always available. The Lurgi-Spülgas process is more complicated but upgrades cheap brown coal. The Char Briquetting process leaves over 20 percent volatile matter in the smokeless fuel, providing a free-burning fuel in high yield. In the Coalite process, the by-products are particularly valuable; typical yields of this process appear in Table 11.

Table 11: Yields of By-products from the Coalite Process

	kg per 1,000 kg coal (9% moisture)
Coalite (large)	700
Coalite (other sizes)	50
Gas	83 (700×10^6 cals)
Tar	75
Light oil	12

In processes for carbonizing briquettes, the starting material is usually lignite or small coal. Briquetting and carbonizing are expensive, but the final product is dense and uniform, commanding a premium price. One process for rendering pitch-bound briquettes smokeless treats them at 350° to 400° C (660° to 750° F) in an atmosphere containing oxygen (*e.g.*, the Anthracine process in France, the Inichar process in Belgium, and the Multiheat process in the U.K.).

The Disco process, formerly operated in Pittsburgh, has been abandoned. All the other processes mentioned operate in Europe, often in competition with one another. No one process predominates because it is necessary to use the variety of coals that are naturally available and because some of the final fuels suit some applications and not others. The prices of the final products delivered to the householder also tend to differentiate the market. Processes based on lignite are commercially important in Europe and elsewhere, but in the U.S. are still under development.

Products. Low-temperature cokes and other smokeless fuels burn readily; they contain 5 to 20 percent volatile matter but nevertheless burn without smoke. The properties of low-temperature tars and light oils are described below.

High-temperature carbonization. In high-temperature carbonization, coal is heated to temperatures of 900° to 1,200° C (1,600° to 2,200° F). The lower temperatures are used for the manufacture of town gas, with coke as a by-product, and the higher temperatures for the manufacture of metallurgical coke, with gas as a by-product. High-temperature coke may also be made especially for domestic use, with gas as a by-product.

For producing town gas, the type of coal and the carbonizing conditions are selected to give the maximum yield of gas. The gas coke is weak and small but satisfactory as a domestic fuel. For making metallurgical coke, a blend of coals and the carbonizing conditions are selected to give hard strong coke.

Processes. For a high-temperature carbonizing process to be commercially satisfactory, it is necessary to (1) pass large quantities of heat into a mass of coal at temperatures up to 900° C or more; (2) extract readily the completed coke from the vessel in which it is carbonized; (3) avoid atmospheric pollution and arduous working conditions for the operators; and (4) carry out the whole operation on a large scale and at a low cost.

Metallurgical coke and coke ovens. Carbonization on a tiny scale was probably first employed by the smiths, who gave coal a preliminary heating on the hearth to leave a residue of coke. Then followed processes akin to charcoal burning. Beehive ovens, developed earlier for the carbonization of wood, were used with coal in the 18th century in Durham, England. These were roughly hemispherical furnaces, containing a coal layer about 750 millimetres (30 inches) thick. The gas and tar evolved were burned in the upper part of the furnace, and the heat so developed was reflected back to the coal. The operation could easily take three days. These ovens were small, required much labour, and led to atmospheric pollution. The last beehive ovens in the U.K. closed in 1954, but an improved form persisted in the U.S. in the 1970s. Because of their simplicity and low capital cost, they are useful in meeting peak demands for coke.

Beehive ovens

Successful ovens that recovered by-products followed the use of narrow side-heated or slot ovens, operating in Belgium in 1850–60. Constructed with a good grade of firebrick, each oven held about four metric tons of coal. Increasingly larger ovens for higher temperatures and faster carbonization followed. Fuel gas (usually coke oven gas but sometimes blast furnace or producer gas) was burned in flues outside the oven walls. Silica walls, first used in the U.S. before 1900, introduced the refractory most widely used with flue temperatures up to 1,450° C (2,600° F). Modern coke ovens can be as large as 6.5 metres (21 feet) high, 15.5 metres (50 feet) long, and 0.46 metres (1.5 feet) wide, each oven holding up to about 36 tons of coal. The coking time, *i.e.*, between charging and discharging, is about 15 hours. Such ovens are arranged in batteries, containing up to 100 ovens each. A coking plant may consist of several such batteries. Large coking plants in the U.S. and in the U.S.S.R. carbonize annually approximately 5,500,000 to 11,200,000 tons of coal each, but older coking plants are still operating throughout the world with quite small ovens and annual throughputs of only 112,000 to 336,000 tons. Modern coke ovens are highly mechanized, thus minimizing atmospheric pollution and lessening the labour needed. Massive machines load coal into each oven, push coke sideways away from the oven, and transfer red-hot coke to a quenching station, where it is cooled with water. In some plants the red-hot coke is cooled in circulating inert gases, the heat abstracted being used to generate steam. This is called dry quenching.

Preparation of charge

The preparation of the coal charge for coke ovens becomes increasingly important as suitable coals become less available. Formerly, single coals were used on their own to yield good strong coke, but today there is rarely enough such coal to supply large plants. Consequently, less good coking coals have to be used. By judicious selection and crushing, however, followed by intimate blending, equally good cokes can be made from a variety of coals. Broadly speaking, suitable blends can be obtained by mixing high-volatile with low-volatile coals, and often small additions of ground, small coke and anthracite are helpful. Drying the coal and even preheating it to 200° C (390° F) may also be helpful and economic. Thus in modern plants the facilities for preparing the blend may be quite elaborate.

During the last hours in the ovens, the coke shrinks and fissures. When it is discharged, it is partly in discrete pieces up to 200 millimetres (eight inches) long or more. After the coke leaves the quencher, it is screened into various sizes and the larger pieces may have to be cut.

Another important and expensive part of the plant is the by-product plant. Hot tarry gases leaving the ovens are collected, drawn away, and cooled. Crude tar separates and is removed for refining. The crude coke oven gas is scrubbed free of ammonia and then usually crude benzol is removed from it. Some of the gas (up to 45 percent) is used to heat the coke ovens, while the rest is available for use as town gas or as fuel.

Typical yields of products from high-temperature carbonization in ovens are shown in Table 12. The gas yield includes gas subsequently used for firing the ovens.

Some of the liquors from the by-product plant contain noxious substances that have to be removed before the effluents can safely be discharged to inland waters. Chemical

Table 12: Typical Yields of Products from High-temperature Carbonization in Ovens

percentage of volatile matter (dry basis)	23	30	35
	Yields kg per 1,000 kg (9% moisture)		
Coke (5% moisture)	750	680	620
Coke breeze (less than 12 mm)	40	60	80
Gas	135	148	160
(10^6 cal)	1,300	1,430	1,540
Tar	27	38	42
Benzol	9	12	15
Ammonia	2	3	4

means may be employed, but, increasingly, systems that involve biological treatment, similar to sewage treatment, are being used.

Town gas and gas coke. Although coke ovens are used to a limited extent to make town gas, other kinds of ovens and retorts are normally employed. This circumstance follows from the need to use high-volatile coals and steaming to maximize gas yields. Flexibility of output and ease of repair are also important considerations.

Horizontal and vertical retorts

An early type, with externally heated horizontal retorts, was originally made of cast iron and later of firebrick. Since the coke was raked out by hand, the task was arduous and led to atmospheric pollution. Early in the 20th century vertical retorts were developed, made of refractory and externally heated with producer gas made from small coke. Coal is fed into the top and coke is extracted from the bottom, with the operation continuous or intermittent. Steam is usually introduced into the bottom of the vertical retorts, serving both to cool the coke and to convert some of it into gas. Several such retorts were normally grouped together in a retort house, sharing certain services and facilities. Gasworks of this kind were often quite small. Even in large towns, there rarely was more than 1,100,000 tons per annum of coal carbonized in one gasworks. In many countries, gasworks that make gas from coal are being rapidly replaced by natural gas and town gas made from petroleum products.

The by-products plant at a gasworks is similar to that described for coking plants. The gas is cooled, and tar, ammonia, benzol, and hydrogen sulfide are removed. Pressure boosters then send the gas through distribution mains.

Coke discharged from the retort is cooled with water and prepared for market by cutting and screening. Noxious effluents from the by-product plant are usually treated in domestic sewage plants after preliminary treatment.

Typical yields of products from vertical retorts with steaming are shown in Table 13.

Table 13: Typical Yields of Products from Vertical Retorts with Steaming

	per 1,000 kg coal (3% moisture)
Town gas	458 cubic metres (1,930 × 10^6 cal)
Coke and breeze*	635 kg
Tar	63 kg
Benzol (recovered)	10 kg
Ammonia	3 kg

*Of which 140 kg is used to fire the retorts.

Products and their uses. *Coke.* The bulk of oven coke (sized about 40 to 100 millimetres [two to four inches]) is used throughout the world in blast furnaces to make iron. Much smaller quantities are employed in other metallurgical processes, such as the manufacture of ferroalloys, lead, and zinc, and in kilns to make lime and magnesia.

Foundry coke

Exceptionally large strong coke, known as foundry coke, is used in foundry cupolas to melt iron. The smaller sizes (15 to 50 millimetres [one to two inches]) of both oven and gas coke are suitable for heating houses, hospitals, schools, and commercial buildings. Coke in 10- to 25-millimetre (0.5-to 1-inch) sizes is much used in the manufacture of phosphorus and calcium carbide; from the latter acetylene, mainly for chemical purposes, is made. Large quantities of the smallest sizes (less than 12 millimetres [0.5 inch]), known as coke breeze, are suited for sintering small iron ore prior to use in blast furnaces. Any surplus breeze serves as an industrial boiler fuel. There are a number of small uses for coke where a cheap source of impure carbon is required.

Until 1940 coke was much used in water-gas generators, in which it reacted with steam and air to make a mixture of nitrogen and hydrogen for the manufacture of practically all-synthetic ammonia. But by 1969 hardly any ammonia was made from coke, since cheaper sources of hydrogen could be obtained from natural gas and petroleum.

Coal tar. Crude tar is refined; *i.e.*, separated into its various constituents by distillation and chemical processing. Typical analyses of tar are shown in Table 14, although

Table 14: Typical Analyses of Tars
(percent by weight)

	coke oven tar	gasworks tar	low-temperature tar
Pitch	59.0	44.0	26.0
Creosote	31.0	42.0	55.0
Light oils	2.5	5.4	6.7
Heavy oils	5.4	6.5	9.4
In the oils and creosote occur:			
Tar acids (*e.g.*, phenol, cresols)	3.1	16.0	29.3
Tar bases (*e.g.*, pyridine)	1.8	2.1	2.1
Naphthalene	8.8	3.0	0.7
Anthracene	1.0	0.26	0.06

the recovery of the pure chemicals as such is not complete in practice.

Either pot stills or more usually pipe stills with a fractionating column are employed. Pitch is run off the bottom of the column, while all other fractions, boiling (up to 400° C [750° F]), are distilled overhead, being separated into several fractions according to boiling range. Each fraction is then further worked up for the recovery of any desired chemical component. Tar acids are recovered from aqueous liquors by solvents and from certain oil and creosote fractions with caustic soda, while tar bases are recovered by extraction with sulfuric acid. Naphthalene is recovered by distillation or crystallization, and anthracene by crystallization of appropriate fractions. The crude chemicals so recovered from several tar refineries are processed into pure chemicals in secondary plants.

Pitch products

Pitch is a valuable binder for briquetting and in making electrodes. The aluminum industry of the world depends heavily on electrodes made from petroleum coke, pitch coke, and pitch. Pitch and mixtures of pitch and creosote in various proportions constitute road tar and fuels. Creosote is valuable for timber preservation. Because of its valuable waterproofing properties, pitch is employed in a number of formulations for treating walls, roofs, and floors. Digesting coal in pitch and heavy oils produces pipe-coating enamels that protect gas and oil mains buried under land and water. The heavy oils are also used for making carbon black.

Of the simpler chemical substances that appear in some of the fractions, the light oils are processed further into benzene, toluene, and xylene for chemical use or as solvents. Naphthalene is processed into mothballs but more importantly is used in dyestuffs and to make phthalic anhydride, in turn used to make plasticizers, resins, and paints; while anthracene is used mainly by the dyestuffs industry. Tar acids and bases enter into a wide range of chemicals, such as plastics, explosives, dyestuffs, agrochemicals, and pharmaceuticals.

Historically, coal tar is important because it has supported the chemical industry with such chemicals as benzene, toluene, phenol, cresol, naphthalene, and anthracene. In the 1930s, however, such substances began to evolve from petrochemical sources, which took over the major supply of the requirements of the chemical industry. Only anthracene and some of the higher tar acids and bases are still obtained more from coal tar than from petrochemicals.

Light oil. After tar has been removed, coal gas still contains some hydrocarbons that can be removed and recovered as a light, oily liquid often referred to as benzol,

Table 15: Typical Analyses of Light Oil
(percent by weight)

	coke ovens (benzol)	gasworks (benzol)	low-temperature (light oil)
aromatics			
Benzene	65	45	
Toluene	15	8	15–20
Xylene	5	4	
Naphtha	12	20	30–35
paraffins	2	20	40–45

which can be refined and separated into fractions. Typical analyses of light oil are given in Table 15.

The light oils from tar distillation are somewhat similar to, and sometimes are refined together with, the benzol. Benzene and toluene are used as chemicals, along with larger quantities of these substances made from petrochemical sources. During both world wars, large quantities of toluene were employed to make the explosive trinitrotoluene, but the uses of toluene later became more modest and various. The light naphtha contains coumarone and indene, which are extracted and converted to resins. All fractions are to some extent used as solvents over a wide field. The benzol from gasworks and low-temperature carbonization is usually too paraffinic, containing too few aromatic hydrocarbons to justify recovering benzene from it. After fractionation and treatment, such benzol is often used as gasoline and diesel fuel, though these are now low-value outlets and often such benzols are left in the gas or otherwise used as fuel.

Gaseous products. Coke oven gas is employed as fuel and town gas. Unpurified or partly purified gas is used in steelworks. Purification (removal of sulfur compounds) is effected by passing the gas through beds of iron oxide mixtures or by scrubbing it with chemical solutions. In Germany, coke oven gas is sent through a pressure grid over long distances, a practice that justifies operating the by-product plant, after cooling and tar removal, at elevated pressure. Coke oven gas from large plants formerly was used as a source of hydrogen to make synthetic ammonia, but this practice is fast disappearing, except in the U.S.S.R., since it is much cheaper to make hydrogen from natural gas or petroleum naphtha.

The coal gas from a gasworks is employed as town gas. Sulfur removal by the method described is then required by law. Most gasworks carbonize coal at atmospheric pressure, thus permitting limited quantities of gas to be stored economically in gasholders, a prominent feature of such plants. Some of these hold as much as 350,000 cubic metres (12,360,000 cubic feet) of gas. Newer gasworks, operating at pressure on petroleum feedstocks, cannot economically employ this method of storage.

Typical analyses are shown in Table 16.

Table 16: Typical Analyses of Coal Gas
(percent by volume)

	H_2	CH_4	CO	CO_2	N_2	O_2	other hydrocarbons
Coke oven gas	54	28	7	2	6	0.4	2.6
Town gas*	49	20	18	4	6	0.4	2.0

*From continuous vertical retorts with steaming.

OTHER PROCESSING TECHNIQUES AND PRODUCTS

Briquetting. In mining it is economically necessary to utilize small coals to the best purpose. If lump fuel is required for any purpose, it is necessary to convert the small coal into strong coherent lumps, which can be done by briquetting: mixing the coal with a binder and then pressing the mixture together to form shapes of the required size.

With brown coal, no binder is needed since the coal, after drying, can be extruded or otherwise pressed into shapes. With bituminous coal it is necessary to use pitch, bitumen, or other materials. The production of briquettes from cheap brown coal is still popular in Germany, where they find ready use as domestic fuel. Since they are not smokeless, however, such briquettes are increasingly carbonized before sale. Bituminous coal briquettes also are not smokeless and their use has declined rapidly. By contrast, the use of briquettes rendered smokeless by carbonization or other means has increased steadily in many countries, notably in the U.K. and northwestern Europe.

Smokeless binders such as starch, synthetic resin, and sulfite lye have been used to make smokeless briquettes, but these are costly or have other drawbacks. Nevertheless, it is likely that processes utilizing such binders will grow.

Coal, if heated to about 400° C, can be briquetted hot without a binder. The property is utilized in the Char Briquetting process of the National Coal Board (U.K.).

Metallurgical fuels

Metallurgical fuels, sometimes known as formed coke, can also be made by carbonizing briquettes produced from coal, char, or coke. These can be made from non-coking coal (incapable of being formed into coke in ovens). They have uniform size and other properties, from piece to piece. Successful processes of this kind have been developed in Poland (to make foundry fuel), in the U.S. (the National Fuels Corporation process), and in Germany (the Steinkohlenbergbauverein and Lurgi processes).

In the German processes, part of the coal is converted into char in a fluidized bed at about 700° C (1,300° F). The remainder of the coal is added prior to briquetting hot at about 400° C, and the briquettes are subsequently carbonized. It seems likely that formed-coke plants for making metallurgical fuels will be built in those countries that have noncoking coals but insufficient coking coal.

Gasification. Here the objective is to convert most of the coal substance into gas, leaving behind only ash and a little carbon as a solid residue.

Carbon reacts with steam chemically to form hydrogen and carbon oxides; thus

$$C + H_2O \rightarrow H_2 + CO$$
$$C + 2H_2O \rightarrow 2H_2 + CO_2$$

This process is endothermic (absorbs heat), and heat therefore must be added to the vessel in which coal and steam are reacting. The various ways of doing this and the varying nature of the fuel used and the gas obtained have led to the large number of gasification systems, dozens of which have reached commercial importance.

The simplest process is to add air with the steam to form producer gas; some of the coal is burnt to provide heat for the gasification of the remainder. This producer gas contains up to 55 percent nitrogen. Since it has a low caloric value, it can be used only as fuel gas. In this form, however, it can be used where coal itself cannot. The presence of nitrogen may be avoided by using oxygen instead of air or a mixture of oxygen and air if some nitrogen can be tolerated. Such gas was once widely used as synthesis gas for making ammonia. By operating under pressure, the gas produced contains some methane, which much improves the calorific value. One such process, developed by Lurgi in Germany, has been employed both to make town gas in Germany, the U.K., and Australia, and synthesis gas in the Sasol synthetic oil plant in South Africa.

Continuous and cyclical gasification

When using air or oxygen with steam, coal may be gasified in the following ways: in lump form in fixed beds (as in gas producers or in the Lurgi process); small coal in fluidized beds (the Winkler process); or in total suspension (the Koppers-Totzek and Texaco processes). The coal is sometimes carbonized first and the resultant coke used advantageously in the gasifier. It is necessary to avoid clinker forming from the ash: either the gasifier must be run below clinkering temperatures (about 1,200° C [2,200° F]) using plenty of steam or above 1,300° C (2,400° F) with the minimum of steam so that the ash is melted to form a slag. The Texaco process is a slagging process in which coal and oxygen and a little steam react at elevated pressure and at temperatures exceeding 1,300° C. The Leuna fixed-bed process can be worked under either nonslagging or slagging conditions according to the amount of steam used.

The preceding are continuous processes. Gasification may also be carried out cyclically. Some fuel is burned with air and the heat stored in a regenerator. Steam is then preheated in the regenerator before passing through a fuel bed with some air or oxygen. In a coke water-gas generator,

air and steam are passed alternately; heat evolved when air is passed is stored in the coke bed and made available when steam is passed. Oil can also be injected into the hot gases to be cracked and to provide additional caloric value. Carburetted water gas made in this way was once much used to meet peak demands for town gas.

In theory, a reactor could be designed to preheat steam sufficiently to satisfy the heat requirements of the steam–coal reaction. Existing materials of construction so far have prevented an economic solution. Some work is commencing in the use of nuclear heat for this purpose. It would perhaps be easier to do this with lignite, which is reactive, than with bituminous coal, which requires higher temperatures.

Finally, gasifiers have been designed in which coal is held in metal or ceramic retorts and heated externally while steam is passed through the coal. This process is used in vertical retorts to increase the yield of town gas, but it is not normally economic to gasify coal completely in this way.

Typical analyses of some manufactured gases are contained in Table 17. In addition, the gases contain various amounts of sulfur compounds.

Table 17: Typical Analyses of Some Manufactured Gases
(percent by volume [crude gas])

Fuel	Gas	H_2	CH_4	CO	CO_2	N_2	other hydrocarbons
Coal	Producer	12	3	29	4	51.6	0.4
Coke	Producer	11	0.5	29	5	54.5	0
Coke	Water gas	49	0.5	41	5	4.5	0
Coal	Lurgi	39.8	11	16	32	1	0.2
Coal	Koppers-Totzek	33.9	0.1	51	13	2	0

Hydrogenation. Coal can be converted substantially into oil and gas by reaction with hydrogen at pressures of 250 atmospheres or more. Such oils may be fractioned and purified to make excellent gasoline. In the Bergius process, developed in Germany in the 1930s, clean coal was slurried with recycled oil, and tin compounds were added as catalysts. The mixture was passed through large forgings that constituted the reactors. Compressed hydrogen was passed through the mixture to provide the hydrogen required chemically and to control the resultant rise in temperature. The final mixture was cooled and the pressure reduced. After the unconverted coal and ash were separated, the clean heavy oils were distilled. The heaviest fraction was recycled with fresh coal to the first stage, and the lighter fractions were passed forward to the next stage, where they were further hydrogenated, again at high pressure but now in the vapour phase, over a catalyst (tungsten sulfide). The products from this stage were fractionated and worked up to gasoline.

The difficult stage was the first: quantities of expensive hydrogen were required, and corrosion and erosion caused much trouble. In practice this stage was often bypassed by starting with tar or creosote.

Nevertheless, Germany produced about 75 percent of its synthetic oil in this way in 1941–44. The economics of hydrogenation, however, are unattractive under normal conditions. The most advantageous conditions are to use cheap lignite in circumstances where natural petroleum is expensive. The possibility of such a plant in the U.S. has been envisaged for the future, working on the scale of several million tons of oil annually. The lignite would be carbonized, the resultant tar being hydrogenated and the residual char being used as boiler fuel or for making hydrogen.

In the 1950s, a U.S. firm studied hydrogenation of coal for the production of aromatic hydrocarbons, phenol, and other products. Although these were obtained in good yield, the costs, including those of separating the mixtures of products, were too high for the process to be economic at that time.

Underground gasification. In underground mining, men work under conditions that are often unpleasant and sometimes dangerous. A method of obtaining energy from coal without actually mining it therefore has attractions. One possibility is to pass air and steam through a burning seam of coal underground, recovering the producer gas so formed and using it as a source of energy to generate steam and electricity. Much effort was devoted to this method in the period from 1940 to 1960 in the U.S.S.R., in Western Europe, the U.K., and the U.S. In practice, two boreholes, spaced up to 60 metres (197 feet) apart, were sunk from the surface into the seam. Sometimes a length of seam was prepared in some way; *e.g.*, by forcing hot air through it at pressure. Air and steam were then alternately blown through.

Energy from unmined coal

Desulfurization and other processes. In most of its uses, the sulfur content of coal is objectionable in varying degrees. Part of the sulfur is associated with ash, and coal washing removes some sulfur along with the ash. Much sulfur, however, is more intimately associated with the coal substance itself and cannot be removed by washing. Since carbonization removes some sulfur, coke usually contains a lower percentage of sulfur than did the coal from which it was made. During total gasification most of the sulfur is converted into hydrogen sulfide, in which form it can be readily separated from the gas. Extraction of coal with solvents produces an extract of relatively low sulfur content. Despite the use of these methods and considerable effort, no effective method has been devised to reduce the sulfur content substantially, particularly the portion closely associated with the coal substance.

Sulfur content

Active carbons, akin to charcoal, are produced when coal is subjected to rapid heating followed by steaming. They can be used for the same purposes as active charcoal. One process in commercial operation in the U.K. and in the U.S.S.R. starts from granular anthracite with no briquetting stage involved.

Particularly after calcining, small anthracite is useful as a source of carbon in certain metallurgical processes in which it acts chemically to form carbon electrodes for certain electrochemical and electrometallurgical processes. Granular anthracite, after sizing and drying, is useful as a filter material, particularly in the treatment of water for domestic use and for process water, because it is chemically inert in water.

COAL AS A CHEMICAL

Coal has a complicated chemical structure based on carbon and hydrogen, with a certain amount of oxygen, nitrogen, sulfur, and other elements. One way to use coal as a source of chemical compounds is to break it down by heating or by reaction with steam and oxygen or with hydrogen. It then forms relatively simple compounds such as carbon, carbon oxides, benzene, naphthalene, and phenol. This process of degradation is expensive, however, particularly if chemicals subsequently have to be synthesized again from these simple compounds. Attempts therefore have been made to use coal as a macromolecule without breaking it down unduly.

Methods of degradation

Coal reacts readily with a wide range of chemicals. The effects of mild oxidation and treatment with nitric acid and caustic soda have been well explored. With low-rank coals particularly, this yields complex organic acids, called humic acids. Unfortunately, these acids have still not found a useful outlet of any size. Ammonia also reacts with coal to form substances containing up to 16 percent nitrogen which could be used as a fertilizer, providing both nitrogen and humus-forming material, but so far production appears to be uneconomic. Coal reacts readily with phenol in the presence of catalysts to form products that are soluble in solvents and of potential value to the plastics industry.

In another approach, coal is digested with solvents. Particularly effective are tetralin, phenanthrene, and tetrahydroquinoline. More than 80 percent of bituminous coal can be dissolved by digestion at 300° C (572° F) for a few hours. Such digests still contain undissolved coal and ash that can be separated only with some difficulty. The extract so obtained is essentially coal in solution. The solvent itself can be separated and recycled, leaving behind solid coal extract. Chemically this is rather similar to coal but

has a lower molecular weight, is free from ash, and is low in sulfur. The extract, after hydrogenation, can be used to make synthetic oil or even power-station fuel where sulfur is particularly objectionable, but the practicability of the application depends upon the price of coal. More attractive is the use of coal extract as a source of carbon or graphite for electrodes or artifacts. Furthermore, the ability of coal extracts to mix with and react with simple polymers, such as polyvinyl chloride, may open the way to less expensive plastics for building and civil engineering.

(R.J.M./Ed.)

BIBLIOGRAPHY

Extraction and processing of metals and other nonorganic minerals. *Ore deposits:* F.D. ADAMS, "Origin and Nature of Ore Deposits: An Historical Study," *Bull. Geol. Soc. Am.*, 45:375–424 (1934), a good review of the development of theories concerning ore genesis; D.F. STRONG (ed.), *Metallogeny and Plate Tectonics* (1976), and D.W. STRANGWAY (ed.), *The Continental Crust and Its Mineral Deposits* (1980), two volumes of symposia papers on tectonics and the origin of ore deposits; A.M. BATEMAN (ed.), *Economic Geology* (1955), symposium papers on specific ore deposits; R.L. BATES, *Geology of the Industrial Rocks and Minerals* (1960), a comprehensive reference on nonmetallic minerals; P.T. FLAWN, *Mineral Resources* (1966), a useful reference for mineral economics; C.A. LAMEY, *Metallic and Industrial Mineral Deposits* (1966), a good reference to use in conjunction with Bates's text; W. LINDGREN, *Mineral Deposits,* 4th rev. ed. (1933), an older textbook regarded as a classic, and particularly useful to compare with Adams' paper; T.S. LOVERING, *Minerals in World Affairs* (1943), a reference work on the historical development of mineral resources and economics; H.E. MCKINSTRY, *Mining Geology* (1948), a good textbook on structural and other physical characteristics of ore deposits; D.R. DERRY, *A Concise World Atlas of Geology and Mineral Deposits* (1980), a reference work that locates major economic deposits on geologic maps of all the continents; C.F. PARK, JR., and R.A. MACDIARMID, *Ore Deposits,* 3rd ed. (1975), an excellent textbook and general reference source for students; F.G. SMITH, *Physical Geochemistry* (1963), a comprehensive textbook of particular value as a reference to chemical reactions that are involved in the genesis of ore deposits.

(C.E.B.C.)

Mining: AGRICOLA, *De re metallica* (1556; Eng. trans. by H.C. and L.H. HOOVER, 1950), is a classic work on early mining in Europe. ROBERT PEELE and JOHN A. CHURCH (eds.), *Mining Engineers' Handbook,* 3rd ed., 2 vol. (1941), is a standard U.S. work on basic surface and underground mining to 1941. SOCIETY OF MINING ENGINEERS, *SME Mining Engineering Handbook,* 2 vol. (1973), gives a contemporary treatment. ROBERT S. LEWIS and GEORGE B. CLARK, *Elements of Mining,* 3rd ed. rev. (1964), is a general text that treats both theory and practice. GEORGE B. CLARK (ed.), *Symposium on Mining Research,* first to sixth (1956–61), deals with mining research, explosives, and blasting. WILLIAM C. PETERS, *Exploration and Mining Geology* (1978), is a comprehensive text. C.G. DOWN and J. STOCKS, *Environmental Impact of Mining* (1977), examines all aspects of this controversial subject.

Additional works on surface mining include: EUGENE P. PFLEIDER (ed.), *Surface Mining* (1968); INSTITUTION OF MINING AND METALLURGY, *Symposium on Opencast Mining, Quarrying, and Alluvial Mining* (1965); and *Pit and Quarry Handbook and Buyers Guide* (annual). MELVIN A. COOK, *The Science of High Explosives* (1958), is a comprehensive treatment of theory and experimentation. E.I. DUPONT DE NEMOURS AND COMPANY, *Blasters' Handbook,* 16th ed. (1977), deals with applied blasting techniques. U. LANGEFORS and B. KIHISTROM, *The Modern Technique of Rock Blasting,* 3rd ed. (1978), gives an analysis of blasting theory and practice. Two books on mine ventilation are A. ROBERTS (ed.), *Mine Ventilation* (1960); and HOWARD L. HARTMAN, *Mine Ventilation and Air Conditioning*, 2nd ed. (1982).

Valuable journals and reports on mining theory, policy, research, and practice include: *Mining Magazine, Mining Journal, Mining Engineering, Canadian Mining Journal, Institution of Mining and Metallurgy Transactions, Mining Congress Journal,* U.S. Bureau of Mines publications (*Bulletins, Information Circulars,* and *Reports of Investigations*), *Annales des Mines, South African Mining and Engineering Journal,* and the *Engineering and Mining Journal.*

(G.B.Cl.)

Metallurgy: J.G. PARR, *Man, Metals and Modern Magic* (1958), a fascinating and informative story of discoveries and developments in metals going back to 6000 BC; C.S. SMITH (ed.), *Sorby Centennial Symposium on the History of Metallurgy* (1963), a series of papers containing a wealth of historical information by outstanding metallurgists.

Works that focus on metallurgical technology include: A.M. GAUDIN, *Principles of Mineral Dressing* (1939), a text containing useful information on extractive metallurgy; E.J. PRYOR, *Mineral Processing,* 3rd ed. (1966), a modern text on methods for treating ores for concentration; INSTITUTE OF MINING AND METALLURGY, *Advances in Extractive Metallurgy* (1968), a collection of contemporary papers on this subject; W.H. DENNIS, *Extractive Metallurgy: Principles and Application* (1965), on the basic principles involved in extraction and refining of metals, pyrometallurgy, hydrometallurgy, and electrometallurgy; *Metallurgy of the Nonferrous Metals* (1954), a concise summary of metallurgical factors, including principles, properties, methods of working, and heat treatment, relative to the metals; D.M. LIDDELL (ed.), *Handbook of Nonferrous Metallurgy,* 2nd ed., 2 vol. (1945), a comprehensive treatment prepared by experts in each field; C.R. HAYWARD, *An Outline of Metallurgical Practice,* 3rd ed. (1952), on the occurrence and sources of metals, with a description of practices used in the recovery of metals from ores; A.H. COTTRELL, *An Introduction to Metallurgy* (1967), an excellent source for information on progress in contemporary research in the field of metals; E.R. PETTY, *Physical Metallurgy of Engineering Materials* (1968), on the basic properties of the important metals and their alloys, and on the properties that determine the utility of various metals and alloys for specific purposes; AMERICAN SOCIETY FOR METALS, *Metals Handbook,* 8th ed., 5 vol. (1961–70), specialized reports on all aspects of this topic. JAMES A. MULHOLLAND, *A History of Metals in Colonial America* (1981), is a study of the role of metallurgy, especially that of iron, in the early American colonies.

Transactions of the following professional societies are additional sources of information: American Society for Metals, Metallurgical Society of the American Institute of Mining Engineers, British Iron and Steel Institute, American Electrochemical Society, American Iron and Steel Institute, American Foundrymen's Society, the Society of Automotive Engineering, and the American Society for Testing and Materials. The proceedings from these societies provide data and descriptions of present-day technology in almost every area of metallurgy and metal application.

The latest advances in metal technology are covered in the following technical journals: *Metal Progress, Journal of Metals, Iron Age, Foundry,* and *Modern Casting.*

(C.H.L.)

Production of important industrial metals: General information on the production and consumption of iron ores, iron, and steel may be found in publications of the United Nations, the International Iron and Steel Institute, and the United States Bureau of Mines. The best general source on the technology of iron and steelmaking in the United States is H.E. MCGANNON (ed.), *The Making, Shaping and Treating of Steel,* 8th ed. (1964). Excellent coverage of the properties of steels, methods of processing, metallography, and testing and inspection may be found in TAYLOR LYMAN (ed.), *Metals Handbook,* 8th ed. (1970). Treatment of the current technology is covered in the Swedish periodical *Scandinavian Journal of Metallurgy* (bimonthly); the *Journal* of the British Iron and Steel Institute (monthly); the *Journal* of the Japanese Iron and Steel Institute (monthly); *Tetsu to Hagane* (bimonthly); various publications of the American Mining, Metallurgical, and Petroleum Engineers—in particular, the annual proceedings of this society dealing with the making of iron and steel and with mechanical working; and G.J. DERGE (ed.), *Basic Open Hearth Steelmaking,* 3rd ed. (1964); and C.E. SIMS (ed.), *Electric Furnace Steelmaking* (1962). Leading journals of the trade are *Iron Age, Iron and Coal Traders Review, Stahl und Eisen, Magazine 33, Stahl in English,* and *Iron and Steel Engineer.* A.K. OSBORNE, *Encyclopaedia of the Iron and Steel Industry* (1956), will be helpful to the student. Earlier publications that may also be of interest are L. AITCHISON and W.I. PUMPHREY, *Engineering Steels* (1953); H.P. TIEMANN, *Iron and Steel,* 3rd ed. (1933); R. FORSYTHE, *The Blast Furnace and the Manufacture of Pig Iron,* 3rd ed. (1922); and A.W. and H. BREARLEY, *Ingots and Ingot Moulds* (1918). JEANNE MCHUGH, *Alexander Holley and the Makers of Steel* (1980), is, in effect, an account of the early steelmaking industry and processes.

General references on aluminum include: K.R. VAN HORN (ed.), *Aluminum,* 3 vol. (1967), a comprehensive treatment; J.D. EDWARDS, F.C. FRARY, and Z. JEFFRIES, *The Aluminum Industry,* 2 vol. (1930), a general survey; V.N. TIKHONOV, *Analytical Chemistry of Aluminum* (1973); P.V. FARAGHER, *Fundamentals of Aluminum* (1959, reprinted 1967); LUCIO F. MONDOLFO, *Aluminum Alloys* (1976); AMERICAN METAL MARKET, *Aluminum Supplement* (annual); ALUMINUM ASSOCIATION, *Aluminum Statistical Review* (annual); and ALUMINUM COMPANY OF AMERICA, *Aluminum Handbook Series* (1967). Other suggested readings are: S.L. BRUBAKER, *Trends in the World Aluminum Industry* (1967); and U.S. DEPARTMENT OF COMMERCE, BUSINESS AND DEFENSE SERVICES ADMINISTRATION, *Aluminum Factbook* (1963).

Articles on this subject appear regularly in the periodical *Metals Week.*

Information about copper production is provided in the following works: METALLURGICAL SOCIETY OF AIME, PAUL E. QUENEAU (ed.), *Extractive Metallurgy of Copper, Nickel, and Cobalt* (1961), a collection of technical papers with a valuable bibliography of more than 1,000 references, the majority referring to copper; J.N. ANDERSON and PAUL E. QUENEAU (eds.), *Pyrometallurgical Processes in Nonferrous Metallurgy* (1967), technical articles confined to pyrometallurgy with about one-third of them concerned with copper; ALLISON BUTTS (ed.), *Copper: The Science and Technology of the Metal, Its Alloys and Compounds* (1954), contains authoritative chapters on all aspects of copper, each written by an expert, though somewhat outdated; ROBERT BOWEN and ANANDA GUNATILAKA, *Copper: Its Geology and Economics* (1977), a technical survey of principal copper deposits of the world; ALEXANDER SUTULOV, *Copper Production in Russia* (1967); RALPH HULTGREN and P.D. DESAI, *Selected Thermodynamic Values and Phase Diagrams for Copper and Some of Its Binary Alloys* (1971); statistical data also may be found in the publications of the U.S. Department of the Interior, Bureau of Mines; RAYMOND F. MIKESELL, *The World Copper Industry: Structure and Economic Analysis* (1979), a comprehensive treatment of the world copper industry, including the operation of international copper markets and the economic prospects of the copper industry through the year 2000.

References on zinc include: C.H. MATHEWSON (ed.), *Zinc: The Science and Technology of the Metal, its Alloys and Compounds* (1959), a monograph on all aspects of zinc and its compounds; E.J. COCKS and B. WALTERS, *A History of the Zinc Smelting Industry in Britain* (1968), a modern text including an introduction to the world history of zinc; C.G. MAIER, "Zinc Smelting from a Chemical and Thermodynamic Viewpoint," *Bull. U.S. Bur. Mines 324* (1930), a classic work on physicochemistry. Basic accounts of specific processes include: E.H. BUNCE and E.C. HANDWERK, "New Jersey Zinc Company Vertical Retort Process," *Trans. Am. Inst. Min. Metall. Engrs.,* 121:427–440 (1936), an authoritative account of this process, also included in this volume is a review of zinc history and process technology through 1936; S.W.K. MORGAN, "The Production of Zinc in a Blast Furnace," *Trans. Inst. Min. Metall.,* 66:553–565 (1956–57); and A.C. JEPHSON, A.Y. BETHUNE, and R.C. KELAHAN, "Zinc Recovery at Canadian Electrolytic Zincs' New Valleyfield Plant," *J. Metals,* 18:947–956 (1966).

Lead is treated in LEAD INDUSTRIES ASSOCIATION, *Lead in Modern Industry* (1952), a comprehensive text treating the history, mining, production, and uses of lead; E.R. THEWS, *Metallurgy of White Metal Scrap and Residues* (1930), a work on the chemistry and metallurgy of the treatment of scrap metals, excluding iron and steel; and UNITED NATIONS, *Report of International Lead-Zinc Study Group* (annual), an international statistical report on production and use.

The following works treat calcium: C.L. MANTELL, "The Alkaline Earth Metals: Calcium, Barium and Strontium," in C.A. HAMPEL (ed.), *Rare Metals Handbook,* 2nd ed. (1961), describes recent technological processes involving calcium, barium, and strontium; C.L. MANTELL, *Industrial Electrochemistry,* 3rd ed. (1950), includes operating details for calcium-products processes; W.A. HAMMOND, "Calcium Compounds," *Kirk-Othmer Encyclopedia of Chemical Technology,* 2nd rev. ed., vol. 4, pp. 1–27 (1964), is a general technical approach.

Two works covering all aspects of chromium production are A.H. SULLY, *Chromium,* 2nd ed. (1967); and M.J. UDY (ed.), *Chromium,* 2 vol. (1956–57). Information may also be found in C.A. HAMPEL (ed.), *The Encyclopedia of the Chemical Elements* (1968), dealing with ores, thermal and electrolytic production, chemical and physical properties, alloys, and chemical and refractory uses; and *The Encyclopedia of Electrochemistry* (1964), a description of electrolytic production; and R.E. KIRK and D.F. OTHMER (eds.), *Encyclopedia of Chemical Technology,* 2nd ed., vol. 5 (1964), a comprehensive treatment, including economics.

The production and uses of cobalt, together with general information on the element in all its important forms, are detailed in ROLAND S. YOUNG (ed.), *Cobalt: Its Chemistry, Metallurgy, and Uses* (1961), and his *Analytical Chemistry of Cobalt* (1966), which discuss all aspects of the analytical chemistry of this metal. Statistics and notes of current interest on cobalt may be found in the *United States Bureau of Mines Mineral Yearbook* (annual). The Centre d'Information du Cobalt, Brussels, Belgium, issues *Cobalt* (quarterly), containing papers and abstracts and a variety of other technical literature.

Gold production and the uses of the metal in its various forms are covered by E.M. WISE in "Gold and its Compounds," *Encyclopedia of Chemical Technology,* 2nd ed., vol. 10 (1966), and in *Gold: Recovery, Properties, and Applications,* ed. by WISE (1964). An estimate of the mode of consumption of gold throughout the world has been made by D.O. LLOYD-JACOB and P.D. FELLS in *Gold* (1969). A tabulation of the important physical properties of pure gold is given in J. PENTON, *Properties of High Purity Noble Metals* (1965). The chemistry of gold and silver is presented by J.W. LAIST, "Copper, Silver and Gold," in *Comprehensive Inorganic Chemistry,* vol. 2 (1954). Analytical methods are summarized in E.E. BUGBEE, *Textbook of Fire Assaying,* 3rd ed. (1940), and in F.E. BEAMISH, *The Analytical Chemistry of the Noble Metals* (1966). Also useful is W.S. RAPSON and T. GROENEWALD, *Gold usage* (1978).

Magnesium is dealt with in *Encyclopedia of Chemical Technology,* 2nd ed., vol. 12 (1963), a comprehensive text detailing the production methods of primary magnesium metal; C.S. ROBERTS, *Magnesium and Its Alloys* (1960), a thorough review of the history of magnesium, production methods, alloys, and uses; W.H. GROSS, *The Story of Magnesium* (1949), historical data on magnesium presented in an interesting manner; AMERICAN SOCIETY FOR METALS, *Magnesium* (1946), semitechnical data on the progress of magnesium production and its uses at that time period, *Metals Handbook,* 8th ed., vol. 1 (1961), excellent illustrations on magnesium applications and property data on the material; M. CAMPBELL and H. HATTON, *Herbert H. Dow: Pioneer in Creative Chemistry* (1951), a classic story of the man who pioneered magnesium in the U.S.; J.H. RIZLEY, *Magnesium: A Material of the Space Age* (1960), a paper on the use of magnesium in space presented at the Magnesium International Conference of May 1960.

Data on mercury can be found in R.C. WEAST (ed.), *Handbook of Chemistry and Physics,* 50th ed. (1969), a compilation of the chemical and physical properties of elements and compounds; C.J. SMITHELLS (ed.), *Metals Reference Book,* 4th ed., 3 vol. (1967), a comprehensive reference work on the properties and uses of metals and alloys; D.M. LIDDELL, *Handbook of Nonferrous Metallurgy,* vol. 2, *Recovery of the Metals,* pp. 473–515, a textbook describing the metallurgical processing of ores and concentrates to recover nonferrous metals. See also J.W. PENNINGTON, *Mercury: A Materials Survey,* Bureau of Mines Information Circular 7941 (1959), a useful summary of resources, processing, and uses of mercury; S.H. WILLISTON, "Mercury in the Atmosphere," *J. Geophys. Res.,* 73:7051–7055 (1968), a discussion of a modern geophysical technique for the discovery of ore deposits.

The development of the world nickel industry is described by J.F. THOMPSON and N. BEASLEY in *For the Years to Come* (1960). A summary of present world sources and a description of the processes employed for recovery may be found in J.R. BOLDT, JR., and P. QUENEAU, *The Winning of Nickel* (1967). See also E.J. PRYOR, *Mineral Processing* (1965), for an excellent discussion. The properties of nickel and many of its alloys are discussed in the *Metals Handbook,* 8th ed., vol. 1 (1961); and in *Monograph 106* of the NATIONAL BUREAU OF STANDARDS (1968). The development of nickel steels from the earliest time to the present is covered by R.B.G. YEO and O.O. MILLER in *The Sorby Centennial Symposium on the History of Metallurgy* (1965). Nickel plating was an early use of nickel that has grown substantially over the years. Much practical information on this aspect may be found in a publication of the INTERNATIONAL NICKEL CO., *Nickel Plating Processes and Properties* (1967). World production statistics plus other information on developments may be found in the *Annual Reports of the U.S. Bureau of Mines.*

The history of the production of platinum is well covered in D. MCDONALD, *A History of Platinum from the Earliest Times to the Eighteen-eighties* (1960). The nickel ores that are the principal sources of platinum metals in the U.S. are discussed in J.R. BOLDT and P. QUENEAU, *The Winning of Nickel: Its Geology, Mining and Extracting Metallurgy* (1967). A large world source of platinum metal in South Africa is well described in "Platinum Mining at Rustenburg," *Platin. Metals Rev.,* 13:15–22 (1969). An excellent summary with a discussion of areas of future demand is given in L.B. HUNT and F.M. LEVER, "Availability of the Platinum Metals," *ibid.,* 13:126–138 (1969). The properties and uses of important platinum metal alloys and compounds are described in "Platinum Group Metals," *Kirk-Othmer Encyclopedia of Chemical Technology,* 2nd ed., 15:832–860 (1968); a pioneering summary of properties is presented in R.F. VINES and E.M. WISE (eds.), *Platinum Metals and Their Alloys* (1941). The properties of palladium and its alloys are presented in E.M. WISE, *Palladium* (1968). See also D.W. RHYS and P. TAIMSALU, "Effect of Alloying Additions on the Thermoelectric Properties of Platinum," *Engelhard Inds. Tech. Bull.,* 10:41–47 (1969).

Silver is treated in A. BUTTS and C. COXE, *Silver: Economics, Metallurgy, and Use* (1967; reprinted 1975); J. NEWTON, *An Introduction to Metallurgy* (1938); and AMERICAN SOCIETY FOR METALS, *Metals Handbook,* 8th ed. (12 vol., 1961–).

Coverage of sodium and potassium can be found in the *Kirk-Othmer Encyclopedia of Chemical Technology,* 2nd ed. rev., vol. 16, pp. 361–400 (1968), and vol. 18, pp. 432–515 (1969), an extensive, authoritative, and modern treatise of elements

and compounds that includes some theory, ample properties, and rather complete industrial processes. Other works recommended include: AMERICAN CHEMICAL SOCIETY, *Handling and Uses of Alkali Metals* (1957), a collection of papers comprising a symposium on industrial problems in the handling and application of the alkali metals; IRVING FATT and MARIE TASHIMA, *Alkali Metal Dispersions* (1961), a rather extensive work with details of the methods of preparation; D.W. KAUFMANN (ed.), *Sodium Chloride* (1960), an exceedingly thorough treatise on salt; and MARSHALL SITTIG, *Sodium: Its Manufacture, Properties and Uses* (1956), a full and exhaustive account of the details of manufacture, handling, uses, and the properties and reactions of sodium metal.

Good discussions on tin include: E.S. HEDGES (ed.), *Tin and Its Alloys* (1960), individual chapters written by specialists, covering the physical metallurgy, chemical behaviour, practical metallurgy in the alloy field, tin and tin-alloy coating practices, and applications for tin products; W.E. HOARE, E.S. HEDGES, and B.T.K. BARRY, *The Technology of Tinplate* (1965), a profusely illustrated guide for manufacturers and users of tinplate; C.L. MANTELL, *Tin: Its Mining, Production, Technology and Applications,* 2nd ed. (1949, reprinted 1970), a comprehensive coverage of tin involving the history, the geology and sources of tin deposits, mining and metallurgy, properties of tin and tin alloys, analysis, and applications; P.A. WRIGHT, *Extractive Metallurgy of Tin* (1966), an up-to-date treatise on the physical chemistry of tin-melting processes linked by a discussion of the economics involved; NATIONAL PRODUCTION AUTHORITY, U.S. DEPT. OF COMMERCE, *Materials Survey: Tin* (1953), a detailed textual and statistical study of the tin industry; D. HANSON and W.T. PELL-WALPOLE, *Chill Cast Tin Bronzes* (1951), a text dealing with the history, constitution, melting and casting, and heat treatment of tin bronzes; T.S. MACKEY, "The Electrolytic Tin Refining Plant at Texas City, Texas," *J. Metals,* 2:32–43 (1969), a description of a modern refinery for tin metal recovery from low-grade concentrates including by-product recovery; and manuals of the TIN RESEARCH INSTITUTE, Columbus, Ohio, on the subject of tin coating and alloying processes, electroplating, analysis, corrosion behaviour, research and development, uses and applications for tin products. JOHN THOBURN, *Multinationals, Mining and Development* (1981), is a study of the tin industry.

For information about titanium, consult, A.D. and M.K. MCQUILLAN, *Titanium* (1956), a classic text describing the metallurgy of titanium and its alloys, relating properties to metallurgical phenomena; J. BARKSDALE, *Titanium: Its Occurrence, Chemistry and Technology* (1949), on the geology, extraction, and chemistry of titanium minerals, metal, and compounds; S.C. WILLIAMS, *Report on Titanium, The Ninth Major Industrial Metal* (1965), a modern review of the titanium industry with emphasis on technical-economic factors involved in the production and uses of titanium; S. ABKOWITZ, J.J. BURKE, and R.H. HILTZ, JR., *Titanium in Industry* (1955), a general review of the properties and applications for titanium and its alloys in the aerospace and other industries; H. CONRAD (ed.), *Applications Related Phenomena in Titanium Alloys* (1968), a compilation of papers describing metallurgical phenomena that influence the performance characteristics of titanium alloys in practical applications; see also *Industrial Applications of Titanium and Zirconium: A Symposium* (1981).

Studies on tungsten are K.C. LI and C.Y. WANG, *Tungsten,* 3rd ed. (1955), a comprehensive textbook covering history, geology, ore dressing, metallurgy, chemistry, analysis, applications, and economics; C.J. SMITHELLS, *Tungsten,* 3rd rev. ed. (1952), discusses the metallurgy, properties, and applications, with particular reference to production; G.D. RIECK, *Tungsten and Its Compounds* (1967), contains useful data; C. AGTE and J. VACEK, *Wolfram und Molybdän* (1959; Eng. trans., *Tungsten and Molybdenum,* 1963), all aspects of the metals are included, most useful in detailing conditions for sintering and working; N.E. PROMISEL (ed.), *The Science and Technology of Tungsten, Tantalum, Molybdenum and Niobium and Their Alloys* (1964), report of a conference on the positions of the metallurgy of tungsten and other metals, with information on alloying, properties, protection, and fabrication; LOS ALAMOS SCIENTIFIC LABORATORY, *Bibliography on Tungsten, its Alloys and Compounds,* LAMS-2401, 2 vol. (1960), very detailed listing; P. SCHWARZKOPF and R. KIEFFER, *Refractory Hard Metals* (1953), deals with the preparation and properties of tungsten carbide, nitride, and silicide; and *Cemented Carbides* (1960), details the preparation, composition, properties, and application of composites formed by sintering refractory materials with a binder metal.

Full coverage of uranium is provided by J.H. GITTUS, *Uranium* (1963); N.P. GALKIN and B.N. SUDARIKOV *et al., Technology of Uranium* (1966; orig. pub. in Russian, 1964); R.G. BELLAMY and N.A. HILL, *Extraction and Metallurgy of Uranium, Thorium and Beryllium* (1963); E.H.P. CORDFUNKE, *The Chemistry of Uranium* (1969); and W.D. WILKINSON, *Uranium Metallurgy,* 2 vol. (1962).

(H.A.W./J.F.E./K.R.V.H./A.Bu./A.W.S./A.W.R./E.J.M./C.L.M./F.E.B./R.S.Y./E.M.W./J.H.R./J.E.S./S.L.Sm./W.L.J./J.B.Pa./R.McK.MacI./H.R.O./G.L.M./J.Wi./Ed.)

Extraction and processing of petroleum and natural gas. *Recovery:* T.C. FRICK (ed.), *Petroleum Production Handbook,* 2 vol. (1962), comprehensive coverage of all phases of petroleum production technology, with diagrams, charts, and tables; F.W. COLE, *Reservoir Engineering Manual,* 2nd ed. (1969), a general treatise, including descriptions of petroleum reservoir evaluation and methods of increasing the production of oil and gas from a reservoir; B.C. CRAFT and M.F. HAWKINS, *Applied Petroleum Reservoir Engineering* (1959), and S.J. PIRSON, *Oil Reservoir Engineering,* 2nd ed. (1958), standard textbooks covering the technical side of petroleum reservoir analysis and production predictions; C.R. SMITH, *Mechanics of Secondary Oil Recovery* (1966), a textbook on the methods by which oil to be recovered may be increased by the injection of other fluids into the reservoir; D.R. SKINNER, *Introduction to Petroleum Production* (1981–), covering reservoir engineering, drilling, and well completions.

(B.H.C.)

Refining: The literature on petroleum refining is extensive; each of the publications listed gives further references for study. W.L. NELSON, *Petroleum Refinery Engineering,* 4th ed. (1958), an authoritative textbook on all aspects of the industry; R.J. HENGSTEBECK, *Petroleum Processing* (1959), a highly readable account; A.E. DUNSTAN *et al.* (eds.), *The Science of Petroleum,* 6 vol. (1938–53), an encyclopaedic work prepared by worldwide authorities; W.A. GRUSE and D.R. STEVENS, *Chemical Technology of Petroleum,* 3rd ed. (1960); INSTITUTE OF PETROLEUM, *Modern Petroleum Technology,* 3rd ed. (1962); THE BRITISH PETROLEUM COMPANY LTD., *Our Industry Petroleum,* 4th ed. (1970); and SHELL INTERNATIONAL PETROLEUM COMPANY LTD., *The Petroleum Handbook,* 5th ed. (1966). A valuable reference work is the *Refining Processes Handbook* (pub. at two-yearly intervals by the journal *Hydrocarbon Processing*), containing current technical details and flow diagrams of commercial processes available to the petroleum refiner.

Books treating specific aspects of petroleum refining include: C.S. ROBINSON and E.R. GILLILAND, *Elements of Fractional Distillation,* 4th ed. rev. (1950); and V.A. KALICHEVSKY and K.A. KOBE, *Petroleum Refining with Chemicals* (1956). K.A. KOBE and J.J. MCKETTA, JR. (eds.), *Advances in Petroleum Chemistry and Refining,* 10 vol. (1958–65), contains articles useful for more detailed insight into specific topics.

The utilization of petroleum products is treated in the following works: H. ABRAHAM, *Asphalts and Allied Substances,* 5th ed., 2 vol. (1945); L.C. DENNY, L.L. LUXON, and B.E. HALL, *The Handbook of Butane-Propane Gases,* 4th ed. (1962); W. FRANCIS, *Fuels and Fuel Technology,* 2nd ed. (1980); E.M. GOODGER, *Hydrocarbon Fuels* (1975); V.B. GUTHRIE (ed.), *Petroleum Products Handbook* (1960); and H.H. ZUIDEMA, *The Performance of Lubricating Oils* (1952). The manufacture and utilization of chemicals derived from petroleum is extensively discussed in A. LAWRENCE WADDAMS, *Chemicals from Petroleum,* 4th ed. (1978). The transportation of petroleum is treated in the publication of the ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Pipelines in the United States and Europe and Their Legal and Regulatory Aspects* (1969).

Techniques in petroleum refining are advancing continuously; accounts of recent developments appear regularly in these technical journals: *Hydrocarbon Processing, Oil and Gas Journal, Industrial and Engineering Chemistry Fundamentals, Petroleum Engineer, International Petroleum Economist, International Petroleum Times,* and *International Oil News.*

(J.E.C./A.L.W.)

Processing, distribution, and storage of natural gas. AMERICAN GAS ASSOCIATION, *Historical Statistics of the Gas Industry* (1956), complete historical record of statistics relating to gas production, transmission and distribution, underground storage, customers, sales, revenues, utilization, finances, labour, and prices for U.S. and Canada; and ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, *Impact of Natural Gas on the Consumption of Energy in the OECD European Member Countries* (1969), describes natural gas in the energy economy of the western European countries, its availability, the policies of the countries in the field of natural gas, competition of natural gas with other fuels, and the future role of natural gas in energy supply and its impact in practice.

(H.Pe./Ed.)

Extraction and processing of coal. *Mining:* E.S. MOORE, *Coal,* 2nd ed. (1940), is a concise treatise on the physical properties, geology, and geographic distribution of coal. *The Coal Mine Mechanization Year Books,* ed. by G.B. SOUTHWARD (1928–29), include reports and analyses of early underground studies

on coal-mining operations. J.D.A. MARROW presented a paper in 1962 to the Illinois Mining Institute entitled "History of Development of Mining Machines," containing not only his own experiences but also very substantial research; this article was published in the *Proc. Ill. Mining Inst.*, pp. 62–81 (1962). H.L. HARTMAN (ed.), *Case Studies of Surface Mining* (1969), includes papers concerning unique mining conditions throughout the world, together with procedures in arriving at final solutions to the various problems. It is a companion piece to *Surface Mining*, ed. by E.P. PFLEIDER (1968), the authoritative reference in this field. A biennial report of the NATIONAL COAL ASSOCIATION to keep the industry apprised of progress is *Bituminous Coal Facts. Surface Mining and Our Environment*, issued by the UNITED STATES DEPARTMENT OF INTERIOR (1967), is a beautifully illustrated text describing the degradation of the environment in the recovery of some resources, showing what has been done and what can be done to correct the abuse. SYD S. PENG, *Coal Mine Ground Control* (1978), is a comprehensive text of the room-and-pillar, longwall, and shortwall mining systems; KOEHLER S. STOUT, *Mining Methods and Equipment* (1980), is an illustrated text of conventional and continuous mining methods and machinery.

(M.A.E./Ed.)

Processing: D.W. VAN KREVELIN, *Coal: Typology, Chemistry, Physics, and Constitution* (1961), an authoritative survey of scientific aspects of coal; J.W. LEONARD and D.R. MITCHELL (eds.), *Coal Preparation*, 3rd ed. (1968), a modern text dealing with all aspects of coal preparation; R.A. MOTT (ed.), *The History of Coke Making* (1936), a brief historical account of coke making; *The Triumphs of Coke* (1965), a brief account of some processes for making and using coke; O. GROSSKINSKY (ed.), *Handbuch des Kokereiwesens*, 2 vol. (1955–58), a German text dealing with all aspects of coke manufacture; D. MCNEIL, *Coal Carbonization Products* (1966), a brief text with emphasis on the refining of tar and benzole; G. CLAXTON (ed.), *Benzoles: Production and Uses* (1961), an authoritative and comprehensive survey.

(R.J.M./Ed.)

Manufacturing Industries

Manufacturing—in the modern sense of organized activity devoted to the conversion of raw materials into more useful forms and eventually into finished consumer goods—is too widespread and diverse to be comprehensively described and analyzed in a single article. In the late 20th century, more than 450 separate industries were identified in the United States alone; their products were classified in about 20 major groups belonging to two principal categories, durable and nondurable goods.

Several major industries are treated directly or indirectly in the *Macropædia* articles INDUSTRIES, CHEMICAL PROCESS; INDUSTRIES, TEXTILE; INDUSTRIAL GLASS AND CERAMICS; FOOD PROCESSING; and BEVERAGE PRODUCTION and sections of the articles FORESTRY AND WOOD PRODUCTION; FURS, LEATHERS, AND HIDES; INDUSTRIES, EXTRACTION AND PROCESSING; PRINTING, TYPOGRAPHY, AND PHOTOENGRAVING; and PUBLISHING. This article deals with six industries or industry groups more or less typical of late 20th-century manufacturing activity, making products that range in complexity from floor coverings to spacecraft. Each section of this article includes a historical sketch of the particular industry; a broader treatment of the development of modern industry appears in the article TECHNOLOGY. The techniques of integrating human effort in industrial enterprises are covered in the article WORK AND EMPLOYMENT.

This article is divided into the following sections:

Automotive industry 467
- History 468
  - Developments before World War I
  - Mass production
  - Large-scale organization
  - The automotive industry in World War II
  - The automotive industry after 1945
- The modern industry 471
  - Concentration
  - Diversity of products
  - New car development
  - Manufacturing processes
  - Sales and service organization
  - International operations
  - Economic and social significance

Aerospace industry 474
- Development of the aerospace industry 475
- Aerospace products 476
  - Flight vehicles
  - Other products
- Research and development 478
  - New commercial aircraft
  - Missiles and spacecraft
  - Other areas of research
  - Cooperative programs
- Manufacturing facilities and techniques 481
  - Basic fabrication techniques
  - Product assurance
  - Assembly methods and facilities
  - Maintenance facilities

Ship design and construction 484
- History 484
  - From antiquity until the 19th century
  - Developments since the 19th century
- Ship design 485
  - Principles of naval architecture
  - Hydrodynamics and hydrostatics
  - Structural strength
  - The naval architect
- Shipbuilding 489
  - The shipyard
  - Planning
  - Fabrication and assembly
  - Launching, outfitting, and trials
  - Modern trends
- Propulsion 491
  - Steam installations
  - Internal-combustion engines
  - Gas turbines
  - Nuclear propulsion
  - The engine room

Clothing and footwear industry 493
- History 493
  - Social aspects
  - Modern developments
- Modern materials and design considerations 494
  - Raw materials
  - Quality in apparel and allied products
  - Design in clothing and footwear
- Modern manufacturing processes and equipment 495
  - Cutting processes
  - Sewing production
  - Pressing and molding processes
  - Special footwear processes
- Production control and plant considerations 498
  - Division of labour
  - Unit flow and multiple flow
  - Plant layout and materials handling

Floor coverings 498
- History 499
  - Early floor coverings
  - Carpet and rug weaving
  - Smooth-surfaced floor coverings
- Handmade carpets and rugs 499
  - Orientals
  - Other handmade carpets and rugs
- Machine-made carpets and rugs 500
  - Nomenclature and types
  - Manufacturing methods
- Smooth-surfaced floor coverings 502
  - Linoleum
  - Printed felt base
  - Asphalt tile
  - Vinyl floor coverings
  - Rubber floorings
  - Cork tiles and carpet
  - Epoxy resins
- Economic aspects 503

Furniture industry 503
- History 503
- Modern furniture manufacturing 504
  - Materials
  - Storage and transport
  - Woodworking machinery
  - Period furniture
  - Modern factory layout
  - The production process
  - The art of chairmaking
  - Fully upholstered furniture

Automotive industry

The automotive industry includes all those companies and activities involved in the manufacture of motor vehicles, including most components, such as engines and bodies, but excluding tires, batteries, and fuel. The industry's principal product is passenger automobiles; commercial vehicles, though important, are secondary. The vehicles themselves are discussed in the article TRANSPORTATION.

The history of the automobile industry, while brief compared with that of many other industries, has exceptional interest because of its effects on 20th-century history. Although the automobile originated in Europe, the United States completely dominated the world industry for the first half of the 20th century through the invention of mass-

production techniques. In the second half of the century the situation altered sharply as western European countries and Japan became major producers and exporters.

HISTORY

Although steam-powered road vehicles were produced earlier, the origins of the automotive industry are rooted in the development of the gasoline engine in the 1860s and 1870s, principally in France and Germany. By the beginning of the 20th century, German and French manufacturers had been joined by British, Italian, and U.S. makers.

Developments before World War I. Most early automobile companies were small shops, hundreds of which each produced a few handmade cars, and nearly all of which abandoned the business soon after going into it. The handful that survived into the era of large-scale production had certain characteristics in common. First, they

Types of early shops

fell into one of three well-defined categories: they were makers of bicycles, such as Opel in Germany and Morris in Great Britain; builders of horse-drawn vehicles, such as Durant and Studebaker in the United States; or, most frequently, machinery manufacturers. The kinds of machinery included stationary gas engines (Daimler of Germany, Lanchester of Britain, Olds of the United States), marine engines (Vauxhall of Britain), machine tools (Leland of the United States), sheep-shearing machinery (Wolseley of Britain), washing machines (Peerless of the United States), sewing machines (White of the United States), and woodworking and milling machinery (Panhard and Levassor of France). One company, Pierce of the United States, had made birdcages. Two notable exceptions to the general pattern were Rolls-Royce in Britain and Ford in the United States, both of which were founded by partners who combined engineering talent and business skill.

In the United States almost all of the producers were assemblers who put together components and parts that were manufactured by separate firms. The assembly technique also lent itself to an advantageous method of financing. It was possible to begin building motor vehicles with a minimal investment of capital by buying parts on credit and selling the finished cars for cash; the cash sale from manufacturer to dealer has been integral in the marketing of motor vehicles in the United States ever since. European automotive firms of this period tended to be more self-sufficient.

The pioneer automobile manufacturer not only had to solve the technical and financial problems of getting into production but also had to make a basic decision about what to produce. After the first success of the gasoline engine, there was widespread experimentation with steam and electricity; for a brief period the electric automobile actually enjoyed the greatest acceptance because it was quiet and easy to operate, but the limitations imposed by battery capacity proved competitively fatal.

Steam power, a more serious rival, was aided by the general adoption, after 1900, of the so-called flash boiler, in which steam could be raised rapidly. The steam car was easy to operate because it did not require an elaborate transmission. On the other hand, high steam pressures were needed to make the engine light enough for use in a road vehicle; suitable engines required expensive construction and difficult maintenance. By 1910 most manufacturers of steam vehicles had turned to gasoline power. The Stanley brothers in the United States, however, continued to manufacture steam automobiles until the early 1920s.

Patent controversies

As often happens with a new technology, the automotive industry experienced patent controversies in its early years. Most notable were two long, drawn-out court cases in Britain and the United States, in each of which a promoter sought to gain control of the new industry by filing comprehensive patents. In Britain the claim was rejected by the courts in 1901, five years after the patent application. In the United States there was a legal battle between Ford and the Association of Licensed Automobile Manufacturers over the Selden patent, which the association claimed as a basic patent on the gasoline-powered car. In 1911 the courts held the patent "valid but not infringed" by Ford. The main consequence of the decision was the formation in 1913 of the predecessor of the Motor Vehicle Manufacturers Association to supervise an agreement for cross-licensing patents, which was ratified in 1915.

Mass production. The outstanding contribution of the automotive industry to technological advance was the introduction of full-scale mass production, a process combining precision, standardization, interchangeability, synchronization, and continuity. Mass production was a U.S. innovation. The United States, with its large population, high standard of living, and long distances, was the natural birthplace of the technique, which had been partly explored in the 19th century. Although Europe had shared in the experimentation, the U.S. role was emphasized in the popular description of standardization and interchangeability as "the American system of manufacture." The fundamental techniques were known, but they had not previously been applied to the manufacture of a mechanism as complex as a motor vehicle (see the article WORK AND EMPLOYMENT: *Mass production*).

The kind of interchangeability achieved by the "American system" was dramatically demonstrated in 1908 at the British Royal Automobile Club in London: three Cadillac cars were disassembled, the parts were mixed together, 89 parts were removed at random and replaced from dealer's stock, and the cars were reassembled and driven 500 miles (800 kilometres) without trouble. Henry M. Leland, founder of Cadillac Motors and the man responsible for this feat of showmanship, later enlisted the aid of a noted electrical engineer, Charles F. Kettering, in developing the electric starter, a significant innovation in promoting the acceptability of the gasoline-powered automobile.

Ford and the assembly line. The mass-produced automobile is generally and correctly attributed to Henry Ford, but he was not alone in seeing the possibilities in a mass market. Ransom E. Olds made the first major bid for the mass market with a famous curved-dash Oldsmobile buggy in 1901. Although the first Oldsmobile was a popular car, it was too lightly built to withstand rough usage. The same defect applied to Olds's imitators. Ford, more successful in realizing his dream of "a car for the great multitude," designed his car first and then considered the problem of producing it cheaply. The car was the so-called

The Model T

Model T, the best known motor vehicle in history. It was built to be durable for service on the rough U.S. country roads of that period, economical to operate, and easy to maintain and repair. It was first put on the market in 1908, and more than 15,000,000 were built before it was discontinued in 1927.

When the design of the Model T proved successful, Ford and his associates turned to the problem of producing the car in large volumes and at a low unit cost. The solution was found in the moving assembly line, a method first tested in assembling magnetos. After more experimentation, in 1913 the Ford Motor Company displayed to the world the complete assembly-line mass production of motor vehicles. The technique consisted of two basic elements: a conveyor system and the limitation of each worker to a single repetitive task. Despite its deceptive simplicity, the technique required elaborate planning and synchronization.

The first Ford assembly line permitted only very minor variations in the basic model, a limitation that was compensated for by the low cost. The price of the Model T touring car dropped from $950 in 1909 to $360 in 1916, and still lower to an incredible $290 in 1926. By that time Ford was producing half of all the motor vehicles in the world.

Spread of mass production. Ford's success inspired imitation and competition, but his primacy remained unchallenged until he lost it in the mid-1920s by refusing to recognize that the Model T had become outmoded. More luxurious and better styled cars appeared at prices not much higher than that of the Model T, and these were increasingly available to low-income purchasers through a growing used-car market. Abroad, William R. Morris (later Lord Nuffield) in Britain undertook to emulate Ford as early as 1912, but found British engineering firms reluctant to commit themselves to the large-scale manufacture of automotive parts. Morris in fact turned to the United States for his parts, but these early efforts were

cut short by World War I. In the 1920s Morris resumed the production of low-priced cars, along with his British competitor Herbert Austin and André-Gustave Citroën and Louis Renault in France. British manufacturers had to face the problem of a tax on horsepower, calculated on a formula based on bore and the number of cylinders. The effect was to encourage the design of small engines that had cylinders with narrow bore and long stroke, in contrast to the wide-bore, short-stroke engines favoured elsewhere. This design handicapped the sale of British cars abroad and kept production from growing. It was not until 1934 that Morris Motors finally felt justified in installing a moving assembly line; the Hillman Company had preceded Morris in this by a year or two.

The effect of the horsepower tax on design

Large-scale organization. Although the appearance of mass production in the automotive industry coincided with the emergence of large-scale business organization, the two had originated independently. They were related, however, and influenced each other as the industry expanded. Only a large firm could make the heavy investment in plant and tooling that the assembly line required, and Ford was already the largest single U.S. producer when it introduced the technique. The mass producer in turn enjoyed a cost advantage that tended to make it increasingly difficult for smaller competitors to survive. There have been exceptions, but the trend has been consistent.

General Motors. General Motors, which ultimately became the world's largest automotive firm, and in fact the largest privately owned manufacturing enterprise in the world, was founded in 1908 by William C. Durant, a carriage manufacturer of Flint, Michigan. In 1904 he assumed control of the ailing Buick Motor Company and made it one of the principal U.S. producers. Durant developed the idea for a combination that would produce a variety of models and control its own parts producers. As initially formed, General Motors included four major vehicle manufacturers—Buick, Cadillac, Oldsmobile, and Oakland—and an assortment of smaller firms. The combine ran into financial trouble in 1910 and was reorganized by a financial syndicate. A similar combination, the United States Motor Corporation, formed in 1910, collapsed in 1912 and was reorganized as the Maxwell Motor Company. General Motors survived. A new reorganization took place after Durant, with backing by E.I. du Pont de Nemours and Company, regained control in 1916.

Rise of the Big Three. At the end of World War I, Ford was the colossus, dominating the automotive scene with the Model T not only in the United States but also through branch plants throughout the world. British Ford was the largest single producer in the United Kingdom. General Motors was emerging as a potential major competitor in the United States. No other automotive firms of comparable size existed.

During the next decade there was a striking transformation. The depression of 1921 had far-reaching effects on the U.S. automotive industry. General Motors was plunged into another financial crisis. Alfred P. Sloan became president of the corporation in 1923 and raised it to its unchallenged first place in the industry. Among other steps he gave General Motors a staff-and-line organization, with autonomous manufacturing divisions, which facilitated management of a large corporate structure and became the model for other major automotive combinations. Ford also went through a crisis because the crash caught him involved in the construction of a large new plant (River Rouge) and in the process of buying out his stockholders. Ford weathered the storm (though many of his dealers, unable to sell cars and not permitted to return them, went out of business), but the Ford Motor Company had reached its crest.

Chrysler's entry

The third member of the "Big Three" automotive manufacturers in the United States was created at this same time. When the Maxwell Motor Company failed in the 1921 depression, Walter P. Chrysler, formerly of General Motors, was called in to reorganize it. It became the Chrysler Corporation in 1925 and grew to major proportions with the acquisition of the Dodge firm in 1928. When Ford went out of production in 1927 to switch from the Model T to the Model A (a process that took 18 months), Chrysler was able to break into the low-priced car market with the Plymouth.

The independents. By 1929 the Big Three supplied three-fourths of the U.S. market for motor vehicles; most of the remainder was divided among the five largest independents—Hudson, Nash, Packard, Studebaker, and Willys-Overland. In less than 10 years the number of automobile manufacturers in the United States dropped from 108 to 44. Some of the minor firms had technological or personal interests, including Nordyke and Marmon, makers of luxury cars, and E.L. Cord, who marketed a front-wheel-drive car in 1929. The depression years of the 1930s eliminated all but the largest independent manufacturers and increased still further the domination of the Big Three. Motor vehicle production declined from a peak of more than 5,000,000 in 1929 to a low of just over 1,000,000 in 1932. It rose again slowly but had not returned to the 1929 figure when World War II broke out.

While these years were difficult economically, they saw some significant developments within the industry. Greater emphasis was placed on style in passenger-car design, with the general trend in the direction of incorporating the body, bumpers, and mudguards into a single pattern of smoothly flowing lines. A number of technical features came into general use: the V-8 engine, introduced by Ford in 1932; three-point engine suspension; free-wheeling (permitting the car to coast freely when the accelerator was released); overdrive (a fourth forward speed); and, on a limited scale, automatic transmission.

Growth in Europe. The period from 1919 to 1939 also brought significant growth in automobile manufacturing in Europe, though on a considerably smaller scale than in the United States. The European industry was moving in the same directions as the U.S. industry, toward a mass market for motor vehicles, but it made slower progress for a variety of reasons: lower living standards with less purchasing power, smaller national markets, and more restrictions in tax and tariff policies. Still, the same trend toward concentration was discernible. British automotive production rose from 73,000 in 1922 (both private and commercial vehicles) to 239,000 in 1929, while the number of producers declined from 90 to 41. Three firms—Austin, Morris, and Singer—controlled 75 percent of the British market in 1929.

The apparent analogy to the U.S. experience was temporary. British production had not yet reached the level at which the economies of scale gave the larger firms as commanding a lead as in the United States. There were other factors that created a somewhat different situation. During the 1930s British automotive production continued to increase steadily, in contrast to U.S. production, and so the smaller companies were not forced to compete for a shrinking market. With output reaching almost half a million in 1937, at the end of the decade there were six major British producers instead of three: Morris, Austin, Standard, Rootes, Ford, and Vauxhall. The last two represented entry by U.S. firms. Vauxhall had been bought by General Motors in 1925; Ford had been in Britain since 1911, had lost ground in the 1920s, and had later recovered. The Rootes Group, based on Hillman and Humber, was a combine formed by a family that had built a large automobile sales concern and then moved from sales to production. The replacement of Singer by Standard was simply the rise of one company and the decline of another, as evidence that open competition could still change the structure of the British automotive industry.

British production gains in the 1930s

In France, three major firms—Peugeot, Renault, and Citroën—emerged in the 1920s. Citroën accounted for 40 percent of French automotive production in 1925, but had reached that dominating position at the cost of financial stability. When André Citroën died before the decade ended, his company came into the hands of Michelin Tire. A new French firm, Simca, rose to prominence in the 1930s. The German automobile industry suffered from the dislocation of World War I and Germany's subsequent economic difficulties. The major developments of the 1920s were the merger of Daimler and Benz in 1926, after the founders of those firms had died (their bitter rivalry for the distinction of being the inventor of the gasoline

automobile made any such union during their lifetimes unthinkable), and the entry of General Motors onto the German scene through the acquisition of the Opel company in 1929. The Germans were ardent admirers of Henry Ford and his methods, which they termed *Fordismus,* but Ford never succeeded in becoming a power in the German automotive world. During the 1930s the Nazi regime sought to emulate Ford by undertaking mass production of a low-priced car, the Volkswagen, but the onset of war interrupted this project. Italian automobile manufacturers gained a reputation for highly engineered sports cars and racing cars, but Italy had no mass market and therefore achieved only small-scale production at that time.

The automotive industry in World War II. During World War I the productive capacity of the automotive industry first demonstrated its military value. Motor vehicles were used extensively for transport and supply. In addition, automotive plants could readily be converted into facilities for manufacturing military equipment, including tanks and aircraft. For all of the belligerents the conversion of automotive facilities was an afterthought, improvised after the beginning of hostilities, and the U.S. industry, involved only for a short time, never fully utilized its capacity.

"Shadow factories"

More preparation was made for using the resources of the various automotive industries as World War II approached. The British government built "shadow factories" adjacent to their automotive plants, equipped to go into military production (principally aircraft) when war came, with managerial and technical personnel drawn from the automotive industry. France attempted conversion, but belatedly and inefficiently. The German automotive industry, which built the military vehicles needed for blitzkrieg, was not fully converted to military production until 1943. In the United States the preparation for industrial mobilization was negligible until 1940; in fact, there was no serious effort even to restrict civilian automobile production until after the attack on Pearl Harbor. Still, the U.S. automotive industry represented such a concentration of productive capacity and skill that, once its resources had been harnessed to war production, its contribution was tremendous. Between 1940 and 1945 automotive firms made almost $29,000,000,000 worth of military materials, a fifth of the entire U.S. output. The list included 2,600,000 military trucks and 660,000 jeeps, but production extended well beyond motor vehicles. Automotive firms provided one-half of the machine guns and carbines made in the United States during the war, 60 percent of the tanks, all the armoured cars, and 85 percent of the army helmets and aerial bombs.

It had been assumed that automotive facilities could be readily converted for aircraft production, but this proved more difficult than anticipated. Automobile assembly plants did not readily accommodate airframes, nor could an automobile engine factory be converted without substantial modification. These problems were eventually resolved, and automobile companies contributed significantly to aircraft production.

Britain was better prepared to use the resources of its automotive industry, at that time the world's second largest. The shadow factories became operative, and Austin, Morris, Standard, Daimler, Ford, and Rootes participated in filling the wartime demand for aircraft and aircraft engines. Leyland Motors and Vauxhall built tanks. Lord Nuffield made a notable contribution to the production effort by establishing a system for repairing aircraft, employing the sales and service organization of Morris Motors, and it was subsequently extended to a large number of small contractors.

The automotive industries of the other belligerents were smaller in scale, and their facilities for armaments manufacture were proportionately greater than in the United States or Great Britain. Consequently, the automotive firms in these countries were concerned chiefly with meeting the insatiable demand for vehicles. The various Ford properties that came under German control, along with Volkswagen, which turned out the German equivalent of the jeep, were employed in this manner. Renault, a tank manufacturer since World War I, built tanks for France and later for Germany.

The automotive industry after 1945. After World War II there was a striking expansion of motor vehicle production. During a 35-year period the total world output increased almost 10-fold. The most significant feature of this increase was that most of it occurred outside the United States. Although U.S. production continued to grow, its share of world automotive production fell from about 80 percent of the total to 20 percent. Among individual nations the United States was the leading producer until the recession of the early 1980s. In 1980 Japan became the leading producer, with the European Economic Community (EEC) ranking second.

The industry in the United States. At the end of World War II the U.S. automobile industry had intact facilities, somewhat enlarged by construction for military needs. There was also a great demand for automobiles. This situation invited several attempts by newcomers to enter the industry, but all proved unsuccessful. The most promising, Kaiser-Frazer, lasted for some 10 years but lacked the financial, technical, and sales resources to compete when the automobile market returned to normal. By the mid-1950s Kaiser-Frazer had stopped producing everything but Willys Jeeps, an operation that had been acquired by buying Willys-Overland. The manufacture of Jeeps continued as a subsidiary of Kaiser Industries until 1970, when the division was sold to American Motors Corporation (AMC) in a transaction that gave Kaiser financial interest in AMC.

Kaiser-Frazer

The trend of the automotive industry to mergers and large-scale organization, and to a situation in which each producer could affect but not control the market, continued unchecked. In 1954 Nash and Hudson joined to form AMC. The company enjoyed temporary prosperity in the late 1950s when it introduced the first U.S. compact car, the Rambler, in response to growing imports of small foreign cars. A merger of Studebaker and Packard in 1954 was less successful. The new company stopped production in the United States in 1964 and in Canada two years later.

Declining sales and heavy financial losses pushed the Chrysler Corporation close to bankruptcy in the late 1970s. Attempting to avert the company's collapse, the U.S. government granted Chrysler $1.5 billion in loan guarantees. In return Chrysler surrendered supervisory control to the government's Chrysler Loan Guarantee Board. By exercising such control, the U.S. government became a de facto partner of the country's third largest auto maker.

By the early 1980s the automotive industry in the United States was concentrated in four major firms—General Motors, Ford, Chrysler, and AMC—and two important manufacturers of commercial vehicles, the Jeep Corporation and International Harvester. A few producers of specialized vehicles remained, along with an assortment of companies that made automotive parts and components.

Technically, the decades after World War II had been marked by improvement and refinement rather than by important innovation. Diesel engines were increasingly used on trucks and buses. Automatic transmissions became virtually standard equipment for passenger cars, and power brakes and power steering found widespread acceptance. In the early 1960s Chrysler experimented with a gas turbine engine for passenger automobiles, but it had too many technical problems for general use.

Styling became increasingly important in automotive design as a marketing device. The general trend in styling became established late in the 1920s when cars began to lose their square, boxlike lines and to develop flowing curves. In time the new design encompassed both body and chassis, integrating such formerly separate features as mudguards, running boards, and bumpers. A combination of pressures made U.S. cars of the 1950s high-powered and ornate, with extravagant use of chrome and exaggerated tail fins; these features were abandoned when the public found the simpler lines of imported cars more attractive.

Growth in Europe. In Europe motor vehicles were recognized as an export item that could help restore war-shattered economies. Britain, for example, earmarked more than half of its automotive output for export and

restricted domestic purchases for several years after the war. In addition, the horsepower tax was abandoned to enable British manufacturers to build profitably for the world market. The most popular British designs (excluding such a specialized luxury vehicle as the Rolls-Royce) continued to be lightweight cars, including several models with an ingenious front-wheel drive. The trend to concentration led in 1952 to the merger of Morris and Austin to form the British Motor Corporation, Ltd., a combine that accounted for about two-fifths of Britain's motor vehicle production. Another British combine was formed around Leyland Motors, which had grown into the nation's largest manufacturer of commercial vehicles and became a power in the passenger-car field by acquiring Standard-Triumph and Sunbeam in the 1950s. Leyland and the British Motor Corporation united in 1968 as the British Leyland Motor Corporation (later British Leyland Ltd. and, after 1978, BL Ltd.); this move, sanctioned by the government, was intended to forestall possible U.S. domination of the British automobile industry. Except for Rolls-Royce, whose automobile production was only a very small part of the company's business, British automobile output was then largely controlled by four firms: British Leyland, Ford, Vauxhall, and Rootes, which came under Chrysler control in 1967. When British Leyland had financial difficulties in the early 1970s it was taken over by the government.

Introduction of the Volkswagen

The revival of the German automobile industry from almost total destruction was a spectacular feat, with most emphasis centring on the Volkswagen. At the end of the war the Volkswagen factory and the city of Wolfsburg were in ruins. Restored to production, in a little more than a decade the plant was producing one-half of West Germany's motor vehicles and had established a strong position in the world market. Breaking away from what had become standard design, the Volkswagen used a four-cylinder air-cooled engine at the rear of the car. The Volkswagen also dispensed with the annual model change that had become customary with other automobile manufacturers. Although the company had been founded by the German government, in the 1960s the government divested itself of 60 percent of its interest by selling stock to the public, an unusual case of denationalization in an era when nationalization of industry was far more common.

Fiat (Fabbrica Italiana Automobili Torino), a firm founded in 1899 but without a mass market until the 1950s, dominated Italian automotive production. The French industry was centred on Renault, Peugeot, Citroën, and Simca. Renault was nationalized in 1944 and merged with Peugeot in the 1960s. Simca became a Chrysler property in 1958, and Fiat acquired a 15 percent interest in Citroën. Although Sweden was a relatively small producer, Swedish-built Volvos and Saabs became an important factor in the world market during the 1960s and 1970s.

Japan. The most spectacular increases in automotive production after World War II occurred in Japan. From a negligible position in 1950, Japan in 30 years moved past West Germany, France, Great Britain, and the United States to become the world's leading automotive producer. Steadily growing export sales of Japan's small, fuel-efficient cars played a major role in this achievement. During the late 1970s and early 1980s, Japan's principal auto makers—Toyota, Nissan, Honda, and Toyo Kogyo—enjoyed such impressive export gains in North American and western European markets that some governments threatened to impose trade restrictions on Japanese imports. In 1981 the Japanese government responded to these pressures by announcing a cutback on the export of automobiles to Canada and the United States.

THE MODERN INDUSTRY

The modern automotive industry is huge. In the United States it is the largest single manufacturing enterprise in terms of total value of products, value added by manufacture, and number of wage earners employed. One of every six U.S. businesses is dependent on the manufacture, distribution, servicing, or use of motor vehicles; sales and receipts of automotive firms represent more than one-fifth of the country's wholesale business and more than one-fourth of its retail trade. For other countries these proportions are somewhat smaller, but Japan and the countries of western Europe have been rapidly approaching the U.S. level.

Concentration. The trend toward concentration in the industry has already been traced. In each of the major producing countries the output of motor vehicles is in the hands of a few very large firms, and small independent producers have virtually disappeared. The fundamental cause of this trend is mass production, which requires a heavy investment in equipment and tooling and is therefore feasible only for a large organization. Once the technique is instituted, the resulting economies of scale give the large firm a commanding advantage, provided of course that the market can absorb the number of vehicles that must be built to justify the investment. Although the precise numbers required are difficult to determine, the best calculations, considering both the assembly operation and the stamping of body panels, place the optimum output at between 200,000 and 400,000 cars per year for a single plant.

Significance of size of production run

The structural organization of these giant enterprises, despite individual variation, resembles the pattern first adopted by General Motors in the 1920s. There is a central organization with an executive committee responsible for overall policy and planning. The operating divisions are semi-autonomous, each reporting directly to the central authority but responsible for its own internal management. In some situations the operating divisions even compete with each other. The Ford Motor Company was consciously reorganized on the General Motors pattern after World War II; other American automotive firms have similar structures, as do BL Ltd. and Rootes.

In addition, the largest producers decentralize their manufacturing operations by means of regional assembly plants. These permit the central factory to ship frames and components rather than complete automobiles to the areas served by the assembly plants, effecting substantial savings in transportation costs. This system was developed for the Ford company in 1911. It was particularly well-suited to the extensive market areas of the United States and Canada. The local assembly plant has also been generally employed by U.S. and European firms to penetrate foreign markets.

Diversity of products. The immense resources of the automotive industry in production facilities and technical and managerial skills have been devoted predominantly to the building of motor vehicles, but there has been a consistent and strong incentive to extend into related products and occasionally into operations whose relationship to automobiles is remote. The Ford Motor Company, for example, has long manufactured tractors.

Ford and General Motors have both been involved in building aircraft. Ford manufactured the famous Ford trimotor, an all-metal transport, from 1924 to 1932. For about 20 years General Motors had an interest in North American Aviation and Bendix Aviation, and General Motors continues to be an important manufacturer of aircraft engines. It also makes refrigerators, diesel engines, diesel-powered railway locomotives, and earth-moving equipment. All these operations resulted from the acquisition by General Motors of companies that were actively or experimentally engaged in these fields.

Along with Nash, AMC acquired the Kelvinator Corporation, a manufacturer of electric refrigerators and other household appliances, but it sold Kelvinator in 1970. Ford moved into electronics and astronautics in 1961 by acquiring the Philco Corporation, a manufacturer of radios, televisions, and electronic equipment, and by organizing an Aeronutronic Division. In Europe the same trend has occurred. Fiat is an industrial giant that not only makes nearly 80 percent of all Italian cars but also builds tractors, earth-moving equipment, locomotives, and atomic reactors. In addition, it has a division engaged in major civil engineering and construction works. Renault builds tractors and machine tools. BL Ltd. makes construction equipment, commercial refrigeration products, and material-handling equipment. The Japanese automobile manufacturer Nissan produces rockets, textile machinery, and boats.

Market surveys

New car development. The process of putting a new car on the market has become largely standardized. If a completely new model is contemplated, the first step is a market survey. Since there may be an interval of five years between this survey and the appearance of the new car in the dealers' showrooms, there is a distinct element of risk, as illustrated by the Ford Motor Company's Edsel. (Market research had indicated a demand for a car in a relatively high price range, but by the time the Edsel appeared, both public taste and economic conditions had changed.) Conferences then follow among engineers, stylists, and executives to agree on the basic design. The next stage is a mock-up of the car, on which revisions and refinements can be worked out.

Manufacturing processes. The bulk of the world's new cars come from the moving assembly line introduced by Ford, but the process is much more refined and elaborated (see Figure 1). The first requisite of this process is an accurately controlled flow of materials into the assembly plants. No company can afford either the money or the space to stockpile the parts and components needed for any extended period of production. Interruption or confusion in the flow of materials quickly stops production. Ford envisioned an organization in which no item was ever at rest from the time the raw material was extracted until the vehicle was completed, a dream not yet realized.

The need for careful control over the flow of materials is an incentive for automobile firms to manufacture their own components, sometimes directly but more often through subsidiaries. Yet complete integration does not exist, and is not desirable. Tires, batteries, and dashboard instruments are generally procured from outside sources. In addition, and for the same reasons, the largest companies support outside suppliers even for items of in-house manufacture. First, it may be more economical to buy externally than to provide additional internal facilities for the purpose. Second, the supplier firm may have special equipment and capability. Third, the outside supplier provides a check on the costs of the in-house operation. U.S. companies rely more than others on independent suppliers. Only General Motors in the United States produces more than half of its own parts.

Production of a new model also calls for elaborate tooling, and the larger the output the more highly specialized the tools. For example, it is expensive to install a stamping press exclusively to make a single body panel for a single model, but if the model run reaches several hundred thousand, the cost is amply justified. Ford Motor Company has been successful in developing computerized numerical controls for body tooling and dies, reducing the lead time between mock-up and manufacture.

The assembly process itself has a quite uniform pattern throughout the world. As a rule there are two main assembly lines, body and chassis. On the first, the body panels are welded together, the doors and windows installed, and the body painted and trimmed (upholstery, interior hardware, wiring). On the second line the frame has the springs, wheels, steering gear, and power train (engine, transmission, drive shaft, and differential) installed, plus the brakes and exhaust system. The two lines merge at the point at which the car is finished except for minor items and necessary testing and inspection. A variation on this process is "unitized" construction, whereby the body and frame are assembled as a unit. In this system the undercarriage still goes down the chassis line for the power train, front suspension, and rear axle, to be supported on pedestals until they are joined to the unitized body

Adapted from *Automotive Industry* (October 1970)

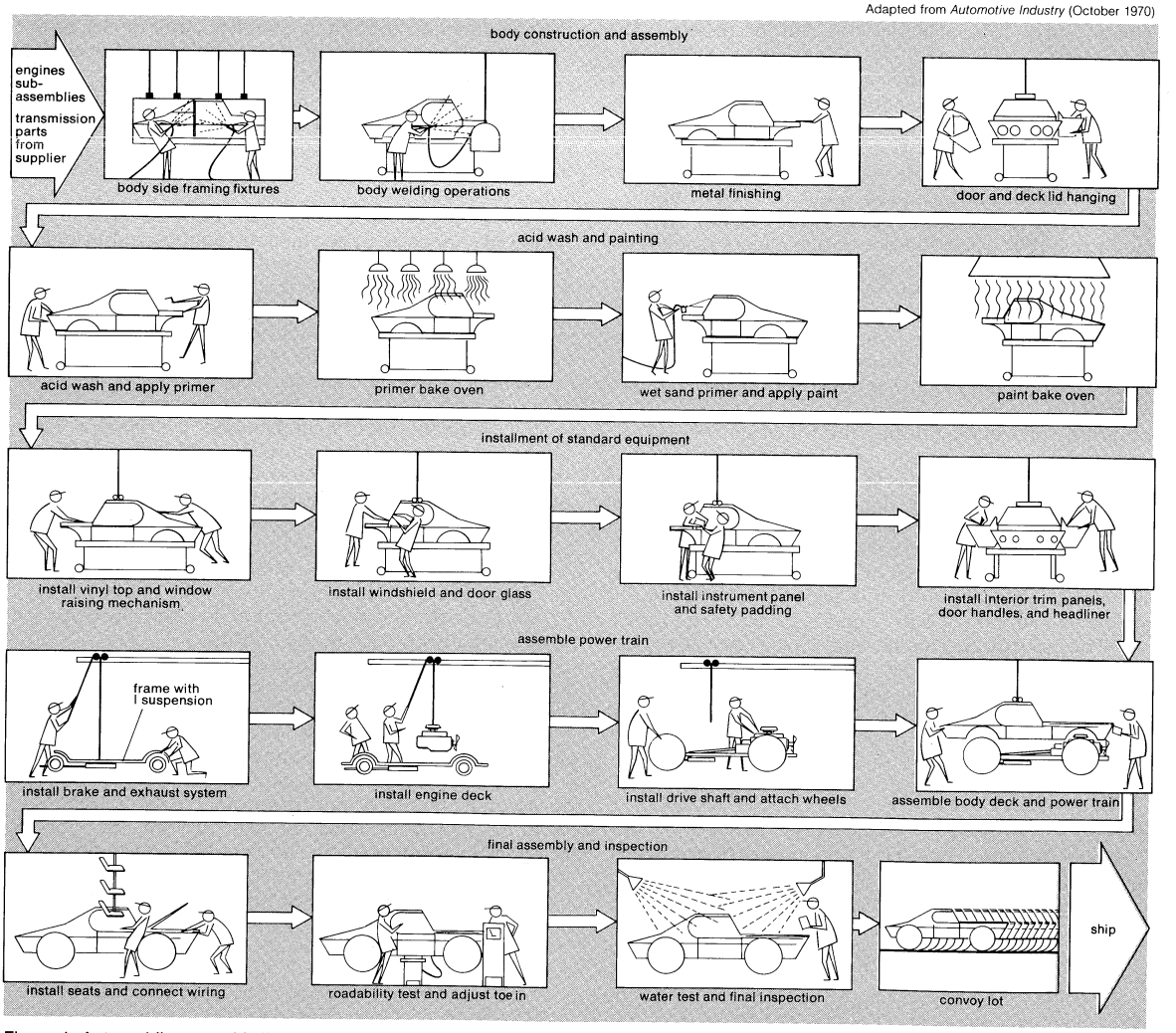


Figure 1: Automobile assembly line.

structure. Most engines are constructed in the same way in separate assembly plants.

Automated assembly lines

Assembly lines have been elaborately refined by automatic control systems and transfer machines, which have replaced many manual operations when volume is high. Austin Motors in Britain pioneered with its automatic transfer machines in 1950. The first large-scale automated installation in the United States was a Ford Motor Company engine plant that went into production in 1951. A universal form of automatic control has used punch cards and computers to schedule assembly operations so that a variety of styles can be programmed along the same assembly line. Customers can be offered wide choices in body styles, wheel patterns, and colour combinations.

Sales and service organization. Mass production implies mass consumption, which in turn requires an elaborate distributive organization to sell the cars and to develop confidence among customers that adequate service will be available. In the early days of the industry, cars were sold directly from the factory or through independent dealers, who might handle several different makes. Many bicycle manufacturers simply used their existing sales outlets when they added horseless carriages to their line. When sales in large quantities became the objective, however, more elaborate and better organized techniques of distribution became essential.

In the United States the restricted franchise dealership became the uniform and almost exclusive method of selling new cars. In this system dealers may sell only the particular make of new car specified in their franchise, must accept a quota of cars specified by the manufacturer, and must pay cash on delivery. In return the dealers receive some guarantee of sales territory and may be assisted in various ways by the manufacturer: financing or aid in advertising, for example. Contracts also specify that dealers must maintain service facilities according to standards approved by the manufacturer.

Seemingly weighted in favour of the manufacturer, the system has been subjected to periodic dealer complaints, producing state legislation and a federal statute in 1956 to protect dealers from arbitrary actions by manufacturers. Yet dealers have never been united in these attitudes, and no effective substitute for the restricted franchise has yet been found. On the contrary, it is becoming the general practice in other parts of the world where large-scale markets for motor vehicles have developed.

The market in used cars is an important part of the distribution system for motor vehicles in all countries with a substantial motor vehicle industry because it affects the sale and styling of new cars. The institution of the annual model was adopted in the United States during the 1920s to promote new-car sales in the face of used-car competition. The new model must have enough changes in styling or engineering to persuade prospective buyers that it is indeed an improvement. At the same time it must not be so radically different from its predecessors as to give the buyer doubts about its resale potential.

Like all machinery, motor vehicles wear out. Some become scrap metal to feed steel furnaces; some go to wrecking yards where usable parts are salvaged. Throughout the world, however, the disposal of discarded motor vehicles has become a problem without a completely satisfactory solution. In too many areas landscapes are disfigured by abandoned wrecks or unsightly automobile graveyards.

International operations. The automotive industry is multinational in both its organization and its operation. Reference has already been made to the international expansion of the principal U.S. automobile companies. The major European and Japanese producers have developed a comparable pattern, with worldwide networks of subsidiaries and affiliates. Toyota has a controlling interest in Australian Motors, after Holden, the second largest Australian producer. Volkswagen's success in the U.S. market led to the opening of a factory in Pennsylvania in 1978. In 1979 Renault agreed to buy 46 percent of American Motors; in return, AMC agreed to manufacture and market Renault cars in the United States. Fiat has an interest in Citroën, and in an unusual transaction the government of Libya bought nearly a 10 percent share in Fiat in 1976. In 1978 Sweden and Norway reached an agreement whereby Volvo transferred some of its operations to Norway. A state-owned Norwegian holding company acquired 40 percent of Volvo's stock, with the remaining 60 percent held by a Swedish holding company, putting Volvo in the unique position of being owned by two governments.

The most promising markets for motor vehicles have traditionally been developed countries with the purchasing power to create a demand for automobiles; these have included North American and European countries and such nations as Australia, New Zealand, South Africa, and Japan. Since 1950 there has been a significant shift in market prospects, however, as developing countries have shown greater growth in vehicle registrations than the highly developed nations. Consequently, there has been an intensification of both assembly and distribution in parts of the world not previously important in the automotive industry.

International vehicle registrations

The great bulk of this production is assembly, done in plants affiliated with and usually operated by U.S., European, or Japanese automotive firms. In order to stimulate their own automotive industries, most developing nations have tariff policies that make imported cars prohibitively expensive and, in addition, requirements that a substantial portion of the components used in local assembly plants be of domestic origin. A certain percentage of local ownership, public or private, is also a normal requirement. The rest of the financing and most of the initial managerial and technical skill come from the parent company.

Economic and social significance. The automotive industry has become a vital element in the economy of the industrialized nations: motor vehicle production and sales are one of the major indexes of the state of the economy in those countries. For such countries as the United Kingdom, Japan, France, Italy, Sweden, and West Germany, motor vehicle exports are essential to the maintenance of healthy international trade balances.

The effect of motor vehicle manufacturing on other industries is very great. Almost one-fifth of U.S. steel production and nearly three-fifths of its rubber output go to the automotive industry, which is also the largest single consumer of machine tools. Moreover, the special requirements of automotive mass production have had a profound influence on the design and development of highly specialized machine tools and have stimulated technological advances in petroleum refining, steelmaking, paint and plate-glass manufacturing, and other industrial processes.

The indirect effects are also considerable through the many auto-related businesses, such as motor-freight operators and highway construction firms. In addition, truck transportation has grown steadily throughout the world.

Highway development. Before the advent of the motor vehicle, roads in most parts of the world were generally poor. The available methods of road transport were so costly and inefficient that, unless there were special considerations such as military movements, it was not worthwhile to maintain roads for other than local traffic. The general use of automobiles created a strong demand for better highways. The first response was to provide for the improvement of existing road networks. Experience subsequently demonstrated that roads for automobile traffic needed to be differentiated functionally, depending on whether they were intended for through traffic or local traffic. Main arteries are best designed as freeways (motorways, *autostrade, Autobahnen); i.e.,* divided highways with complete control of access and no intersections at grade.

Highways functionally differentiated

Social effects. A historian has said that Henry Ford freed common people from the limitations of their geography. The statement cogently summarizes the social transformations still proceeding throughout the world as a result of the motor vehicle. It has created mobility on a scale never known before, and the total effect on living habits and social customs is still incalculable. The automobile has radically changed urban life by accelerating the outward expansion of population into the suburbs. As with other automobile-related phenomena, the trend is most conspicuous in the United States but is rapidly appearing elsewhere. The decentralizing trend is accentuated by the fact that highway transportation encourages

business and industry to move outward to sites where land is cheaper, where access by car and truck is easier than in crowded cities, and where space is available for the one-story structures that permit optimum use of modern materials-handling techniques. Yet the effect on rural life has been, if anything, more pronounced than the effect on cities. In the days of horse-drawn transport the economical limit of wagon transportation was 10 to 15 miles, so that any community or individual farm more than 15 miles from a railroad or navigable waterway was isolated from the mainstream of economic and social life.

In both urban and rural areas the automobile is credited with causing drastic changes in the sexual values of young people, who have found in it a privacy not formerly attainable.

Motor vehicles and paved roads have narrowed much of the gap between rural and urban life. Farmers can ship easily and economically by truck and can drive to town when convenient. In addition, such institutions as regional schools and hospitals are now accessible by bus and car.

It would be impossible to list all of the specific effects of motor vehicle production, but two are especially illustrative. First, the marketing of automobiles has stimulated a great expansion in the use of credit. Installment buying existed before the automobile but in a limited scope. The technique was introduced into the U.S. automobile industry in 1916 by manufacturers of medium-priced cars to help meet the competition of the low-priced Model T. It became a universal practice in all countries in the purchase of motor vehicles, and it accustomed people to buying other durable consumer goods in the same way. Second, there has been a striking development of drive-in businesses, again first in the United States but spreading rapidly everywhere.

Recreational travel. One of the conspicuous effects of the automobile has been to permit nearly everyone in the automotive countries to travel for recreation. The motor vehicle allows for such auxiliary devices as trailers (caravans in Europe), campers, and boat trailers, which broaden the scope of recreational opportunities.

Congestion, pollution, and accidents

Adverse effects. The mass use of motor vehicles was bound to have some unforeseen and undesirable consequences, of which three can be singled out: traffic congestion, air pollution, and highway accidents. The approach to each of these problems illustrates a common propensity to blame the technology, rather than the way in which the technology has been used.

City streets were congested long before the automobile existed, but the problem has been compounded enormously by the masses of motor vehicles that enter or leave cities at peak traffic hours. The constantly growing number of automobiles throughout the world adds to the difficulty of finding remedies for congestion. The heart of the problem is that few city street systems have been designed for automobile traffic. Reliable estimates are that some two-thirds of the vehicles in central business districts are passing through and should have been routed on circumferential highways. Remedying this situation is difficult and expensive. It calls for modern highways to provide both ready access into downtown areas and ways to avoid them. Programs for this purpose encounter vigorous opposition, frequently justified, on the ground that building freeways in cities disrupts neighbourhoods and destroys scenic or historic areas.

The widespread use of automobiles for business travel has also led in many cities to a decline in public-transit systems, and the need to develop and use mass transit has been much discussed. Given the trend toward dispersal of people and businesses in urban areas, it seems doubtful that mass transit will appreciably diminish motor vehicle traffic. Still, in most cities bus systems can provide the needed capacity for public transportation and are the most economical way of doing so.

Atmospheric pollution antedates the automobile, but the concentration of many thousands of motor vehicles in large cities has given the problem a new dimension. Automobile exhausts commonly contribute half the atmospheric pollutants in large cities and even more in cities where atmospheric and topographic conditions are peculiarly conducive to smog formation. In the 1960s federal and state legislation in the United States required the installation of controls on motor vehicles to restrict the emission of pollutants.

Highway accidents create a distressing toll of fatalities and injuries wherever there is widespread use of automobiles, and the social and economic cost of such accidents is incalculable. Efforts to improve highway safety have been successful in most countries, but a reduction in the ratio of fatalities and injuries per distance travelled is often offset by increases in numbers of accidents because of the ever-growing use of motor vehicles.

Beginning with the National Traffic and Motor Vehicle Act of 1966, the United States has required automobiles to have safety features such as seat belts. In 1973–74 the U.S. Congress, attempting to reduce fatalities and conserve fuel, passed a law making 55 miles per hour (mph) the national maximum speed limit. According to the Department of Transportation, the law resulted in a slower, more uniform flow of traffic that saved 18,000 lives between 1974 and 1979. By the late 1970s and early 1980s, however, officials in some Western states raised strong opposition to the national speed limit, arguing that it imposed economic and personal hardship on people who had to travel long distances in short periods of time. The Department of Transportation in the late 1970s ordered that all new cars sold in the United States after 1983 have automatic seat belts or airbags that inflate on impact. In 1981 the federal government rescinded the order, maintaining that the financial burden of complying with safety regulations would hamper the auto industry's recovery from recession. (J.B.Ra.)

Aerospace industry

A combination and contraction of "aeronautics" and "space flight," "aerospace" covers all vehicular flight, whether within or beyond the Earth's atmosphere. The aerospace industry is the complex of manufacturing concerns engaged in the production of flight vehicles, including unpowered gliders and sailplanes, lighter-than-air craft, ground-effect machines, heavier-than-air craft of both fixed-wing and rotary-wing varieties, military missiles, space-launch vehicles, and manned or unmanned spacecraft; propulsion systems and other thrusting devices; on-board equipment essential to the design purpose of the flight vehicle; and ground-based support equipment needed for the operation and maintenance of the flight vehicle. An additional area of the industry's effort, one that is expanding rapidly, is fabrication of nonaerospace products and systems that have been derived from aerospace technology.

The aerospace industry, which of necessity devotes a considerable portion of its effort to research and development in the broadening realm of flight, is a leader of world technological advance. Despite the lack of statistical detail from the Communist nations, it would appear that the aerospace industry is one of the world's largest manufacturing industries in terms of employment and the monetary value of product output.

Nations with extensive industrial development

The two greatest aerospace industrial complexes are those of the United States and the Soviet Union. France and the United Kingdom round out the Big Four, those nations that have achieved a high degree of aerospace sophistication and have the ability to manufacture home-designed systems embracing the entire range of aerospace vehicles. Japan, whose aircraft industry was entirely destroyed in World War II, did not resume production activity until 1952; it expanded rapidly thereafter, producing few home-designed products, but relying for the most part on U.S. licenses. Canada has a relatively small but advanced industry. West Germany has little aerospace production capacity but has developed a strong research and development base. China had been building aircraft of Soviet design since the early 1950s and by the 1970s was producing home-designed aircraft; its demonstrated ability to launch space satellites indicated an advancing technology in other aerospace areas. Sweden's aerospace industry has been very small, but the national policy of

selective specialization has brought about a high degree of competence in a narrow focus of effort. All told, about 30 nations engage in some form of aerospace production activity, for the most part the manufacture of light and medium aircraft of home design or production of heavier aircraft and propulsion systems under license from major-nation companies.

For the major aerospace nations of the world, their own governments and in some cases other governments constitute by far the largest customer. Next in importance as buyers are the world's commercial airlines, a market dominated by the United States since World War II, although the French and British have a large share. Many general-aviation aircraft, private, business, and nonairline commercial planes are produced, but these aircraft are much less costly than jetliners; as a result, the sale of general-aviation equipment represents a small part of the overall volume. In the United States, nonaerospace systems, primarily for state and municipal governments, utilities, surface-transportation companies, hospitals, and other segments of the medical community, represent a significant portion of total sales.

DEVELOPMENT OF THE AEROSPACE INDUSTRY

The aerospace industry may be said to trace its origin to December 17, 1903, the date on which Orville and Wilbur Wright made the first powered flights of a heavier-than-air machine. An industry, however, is by definition a business activity, and for several years after 1903 the Wrights were engaged in a purely experimental effort. A more accurate date for the industry's origin is therefore perhaps 1908, when the Wright Company sold its patent rights for France to a syndicate called the Astra Company, the first recorded commercial transaction of the aerospace industry. In the same year, the Wrights contracted with the U.S. Army for the first military airplane; the contract was successfully concluded the following year after army acceptance tests.

The first aerospace transaction

In 1909 the list of airplane producers was increased by Glenn Curtiss in the United States and by Henri Farman, Louis Blériot, Gabriel and Charles Voisin, and Léon Levavasseur in France. The Wrights licensed production of their airplane in England, and experimental groups that would start regular production within a year or two were active in Germany and Russia. The planes being built differed in design but were similar in construction: wing and fuselage frames were built of wood, usually spruce or fir, and covered by a coated fabric. This kind of airplane construction predominated until after World War I.

Over the next five years, France and Germany, both aware of the war potential of the flying machine, began aircraft construction on what was then a large scale. By the outbreak of war in August of 1914, France had built some 2,000 airplanes of all types, 1,500 of them military. Germany started the war with about 1,000 military machines. Great Britain was third in production, and the United States had completely abdicated its former role as aeronautical leader; in the years before 1914 the United States had built fewer than 100 airplanes, fewer than half of them military. The French industry, assisted by rapidly expanding facilities in Britain, carried the Allied aircraft-production load through most of the war. When the United States joined the Allies, two types of British-designed bombers were assigned it for production, along with one American-designed plane (the Curtiss Jenny trainer). The U.S. did not develop enough production by the war's end to have much effect militarily, but the 400-horsepower Liberty engine, a very advanced power plant for its day, provided the foundation of an aero-engine industry.

Worldwide aircraft production went into a sharp decline after the war, until commercial air transportation, started in 1919, gained public acceptance. The airplane itself was considerably refined in terms of design and performance during the '20s, the most notable change being the gradual ascendancy of the monoplane over the biplane. From the standpoint of manufacturing, the largest change was a swing to metal construction, a wartime innovation introduced in Germany by Hugo Junkers, who also introduced the cantilever wing, which required no external struts or braces. The combination of cantilever wing and metal construction met with rapidly widening acceptance. Metal (steel and aluminum alloys) initially was used only in the airframe structure, but, in the latter years of the decade, it was extended to the wing and fuselage skins, and by 1930 the all-metal airplane was appearing with increasing frequency.

In the 1930s a tremendous acceleration of aeronautical research brought the airplane to design maturity. There were hundreds of innovations, among the more important being large increases in aircraft-engine power, new propeller efficiency, cabin pressurization, retractable landing gear, landing flaps, the autopilot, better instruments and controls, and new aids to navigation for all-weather flying. Performance increased; a major reason was the trend toward "cleaning up," or streamlining the airplane for reduced air resistance. Except for light private planes, all-metal construction became standard.

The most notable influence on manufacturing processes in the 1930s was aircraft size. The growing need for greater capacity in commercial transports, coupled with military requirements for longer range and heavier bomb loads, brought a generation of large-size aircraft. Britain and Germany led the way in the early 1930s with transports capable of carrying as many as 40 passengers; the U.S. took over in the latter half of the decade and introduced the first really modern, highly productive airliners, which were later to see almost universal service. The British, French, Germans, and Russians all produced four-engine bombers of varying efficiency, but again it remained for the U.S., in the latter '30s, to develop and begin production of the airplane that was to become the standard daylight strategic bomber of World War II.

Coming of the large transports

In World War II the production situation of the earlier war was reversed; this time it was the United States that predominated among the Allies. From a 1939 base of fewer than 6,000 planes a year, the U.S. industry, once again bolstered by nonaviation firms, doubled production in 1940 and redoubled it in 1941 and 1942. In the six-year period 1940 through 1945, the U.S. turned out an incredible 300,718 military aircraft.

The postwar years were marked by a rapid shift to jet propulsion as the power source for military aircraft. The first successful turbojet engine had been tested in Britain in 1937 and a rival engine first flown in Germany in 1939. Both nations introduced the jet in war service, though not in significant numbers, and wartime development paved the way for the great transition of the 1940s and 1950s. Jet power was applied first to military fighters, then to bombers, finally to commercial airliners. Military aircraft progressed from the subsonic (below the speed of sound), 600-mile-per-hour level of the late '40s to supersonic flight in the early '50s, to double sonic (twice the speed of sound; Mach 2) capability in the late '50s, then to Mach 3, which represented the upper speed limit of the fastest operational aircraft in the early 1970s. Each increment of performance gain brought an attendant increase in aircraft complexity, producing an across-the-board impact on the manufacturing process, from tools, materials, and facilities to composition of the labour force. At the same time, it started a trend in which research and development became a much more significant part of the total industry effort. The most notable effect, because greater performance permitted fewer airplanes for the assigned mission, was the transition from mass production of many types to short-run production of fewer but much more complex types. Never since the introduction of jets has aircraft production approached World War II levels.

The transition to jets

While the jet transition was moving into high gear, the industry fell heir to a new responsibility, the development and production of guided missiles. It was a logical inheritance, because some of the missiles were aerodynamic, lift-supported like an airplane, and the others shared with the airplane at least a similar set of major components—an airframe, a propulsion system, a method of guidance and control, and a payload.

Wartime Germany had pioneered long-range missiles, introducing as operational weapons two types of surface-to-surface missiles and developing a wide range of other

types. German experience provided a departure point from which the victorious nations advanced their own experimental programs in the latter 1940s. Initial missile production got under way in the early 1950s; first-generation missiles included artillery-like battlefield weapons, antiaircraft weapons, pilotless tactical bombers, and air-launched weapons. In the United States and the Soviet Union the 1950s witnessed the state of production of long-range strategic missiles. Missile sophistication increased through successive generations of automated weapons in the 1960s, and production capability spread to other nations; by 1971 there were 14 missile-building nations, most of them confining their efforts to short-range battlefield- and air-defense systems.

The impact of missilery on manufacturing processes was felt in two primary areas: completely new facilities were required, and the labour-force composition, already changing markedly as the result of the jet transition, was further altered as production of automated weaponry dictated greater emphasis on highly skilled workers—scientists, engineers, and technicians—and less on unskilled personnel.

Impact of Sputnik

Before missilery had reached full maturity, a third manufacturing responsibility was thrust upon the aircraft industry, when the Soviet Union sent the first man-made object into orbit, on October 4, 1957. Their success accelerated the American space effort. In 1959 the U.S. Aircraft Industries Association, Inc. changed its name to Aerospace Industries Association of America, Inc., which, the organization stated, "more properly reflected the changing nature of the effort." Initially confined to the United States and the Soviet Union, manufacture of space systems quickly became a significant portion of the total work load, particularly in the rapid evolution of ever more advanced manned spacecraft.

Space-equipment fabrication compounded all the previously established industrial change trends, particularly the trend toward heightened emphasis on research and development. Once simply a prelude to production of a flight vehicle, research and development soon became an integral part of the manufacturing process, extending to all of the major systems that comprise the flight vehicle, to the tools and methods of fabrication and assembly, and to the multitude of testing devices designed to insure an unprecedented level of equipment reliability.

AEROSPACE PRODUCTS

Flight vehicles. The product line of the aerospace industry is necessarily a broad one, because each of the primary products—that is, the flight vehicles—requires hundreds, thousands, or in a few cases millions of systems, subsystems, components, and individual parts.

Aircraft. Among the primary products, manufacture of aircraft has continued throughout the years as the major work load of the world's aerospace industry. A majority of those nations with such industries build only aircraft and aircraft engines. In the United Kingdom and France, missile-space efforts are relatively small in comparison to aircraft output. Even in the United States and the Soviet Union, both heavily engaged in missile and space work, aircraft manufacture holds first position by any measure of productivity.

Airplane builders fall into two categories: producers of general-aviation planes and producers of heavy aircraft. General aviation embraces planes and light helicopters used for private pleasure flying, personal transportation, business travel, and short-haul commercial operations, such as air taxis and commuter airliners with low take-off weights (specifications vary internationally). Included in the general-aviation bracket are specialized aircraft such as agricultural sprayers, acrobatic craft, sailplanes and gliders, air ambulances, pipeline-patrol craft, geophysical-survey platforms, and a broad variety of other civil applications. Also included are certain light military aircraft used for observation, liaison, and utility duty and incorporating only minor military modifications to their basic civilian-type design. Although many modern general-aviation planes are jet- and turboprop-powered, all are subsonic. General-aviation aircraft constitute the bulk of airplane manufacture in terms of units but little in terms of sales volume.

Types of heavy aircraft

The heavy-aircraft category includes nearly all military planes, commercial airliners, and large corporate aircraft adapted from airline designs. Among the more technologically demanding craft in this category are the supersonic military planes, huge military and commercial transports weighing up to 800,000 pounds (360,000 kilograms), and supersonic jetliners. Also belonging to this group are specialized craft such as the vertical takeoff and landing (VTOL) vehicle, initially a military type but one with broad potential in commercial operations; large helicopters, employed as combat weapon systems, as intratheatre military transports, and as passenger-carrying commercial carriers; unmanned reconnaissance drones and target aircraft used for training of gun and missile crews; and ground-effect machines, or air-cushion vehicles, which operate close to the surface, supported by a column of high-velocity air generated by the machine's engines. The term heavy, as applied to this category, is not entirely a matter of weight but refers as well to the heavier demands of work load per unit to produce the stronger structures, more powerful engines, and more complex ancillary equipment required for the vehicles' design missions.

Guided missiles. Second to aircraft in worldwide industry productivity is manufacture of guided missiles and related products. From the standpoint of manufacturing work load, missiles can be grouped in two basic categories. The first group embraces battlefield, tactical, and air-defense weapons and short- to medium-range systems that carry either conventional high-explosive or low-yield nuclear warheads and that have speed capabilities ranging from subsonic to about Mach 3. Production of such weapons is relatively straightforward, within the capabilities of some of the less sophisticated aerospace industries of the world.

Far more technologically demanding are the strategic missiles that carry multimegaton-yield nuclear or thermonuclear warheads at hypersonic speeds (far beyond the speed of sound; above Mach 3) over distances ranging from more than 1,000 miles to intercontinental. Although it is a defensive rather than an offensive system, the antiballistic missile must be included in this category because of its complexity.

Strategic missiles and space-launch vehicles

A close relative of the strategic missile is the space-launch vehicle. (In space terminology, the launch vehicle is the rocket system that lifts a payload into orbit; the payload itself, whether manned or unmanned, is called a spacecraft, and the combined launch vehicle-spacecraft is known as a space vehicle.) Only in the United States and the Soviet Union have launch vehicles been produced on a quantity basis. The Soviet Union has followed the practice of modifying long-range missiles for space-launch work; the U.S. used converted missiles for the majority of its space launches but additionally developed families of launch vehicles designed from the beginning for space-launch duty. France, Britain, China, Japan, and the European Launcher Development Organization, a consortium of several European nations plus Australia, have all built launch vehicles on a limited basis.

Another missile-related product is the sounding rocket, known in Europe as the probe rocket. Used for upper-atmosphere and near-space research, the sounding rocket is a small launch vehicle that does not orbit a research payload but sends it instead on a brief, parabolic flight above the atmosphere and back to Earth. Sounding rockets generally have one to four rocket stages that fire in sequence to achieve the altitude desired for the experiment; payloads usually run from 50 to about 500 pounds (23 to 230 kilograms), and altitude capability ranges upward to about 1,000 miles (1,600 kilometres). Many countries are engaged in sounding-rocket programs, but the vehicles are primarily of American origin.

Spacecraft. The third major area of aerospace-industry effort is fabrication of spacecraft. Spacecraft are individually tailored to specific research programs and, therefore, do not fall into logical groupings of systems with similar construction characteristics. Unmanned spacecraft, called satellites when they operate in Earth orbit and probes when launched on a deep-space trajectory, have weighed from three pounds (1.4 kilograms) to almost 19 tons (17,000

kilograms) and have taken so many different forms that no basic description is possible. Manned spacecraft, naturally, impose far greater fabrication problems than unmanned systems, because of the equipment necessary to sustain human crews in space and bring them back to Earth. Modern manned spacecraft, weighing up to 100 tons (91,000 kilograms), are the most complex of all aerospace vehicles, and their production has been confined to the United States and the Soviet Union. In addition to these two nations, unmanned spacecraft have been built and orbited (in most cases by means of U.S.-provided launch services) by France, the United Kingdom, Canada, Italy, Australia, Germany, Japan, China, and the European Space Agency, a 10-member combine. A number of other countries have built individual systems and experiments for cooperative space programs sponsored by the United States and the Soviet Union.

Propulsion systems. Rounding out the primary-product line are the propulsion systems for all of the flight vehicles described. There are three basic types of power plants: piston, or reciprocating, engines; turbine engines; and rocket engines. No longer used in current-model heavy aircraft, the piston engine still has wide utility as a power plant for light single- and twin-engine general-aviation planes. The turboprop (turbine spinning a propeller) finds its primary applications in light-medium commercial and business transports and in military planes when greater range is more desirable than speed. The turboshaft (turbine driving a rotor system) engine is becoming the standard propulsion system for the larger helicopters, particularly those of the twin-engine variety. The pure jets, turbojet and turbofan engines, power all supersonic aircraft except research vehicles and a wide range of subsonic types from light single-engine trainers to the mammoth jumbo jetliners.

Types of propulsion systems

The primary employment of the rocket power plant is in propulsion for guided missiles and space-launch vehicles. There are two basic types of rocket engines: solid-fuelled and liquid-fuelled. Because of the requirement for long storage, the great majority of missiles are now powered by solid-fuel systems, as most liquid fuels have to be stored at cryogenic (very low) temperatures. Conversely, most space-launch vehicles employ liquid-propellant rocket systems, although large solid rockets are frequently used as boost-stage auxiliaries for greater thrust in the initial phase of a launch. An additional usage of the rocket engine is for spacecraft on-board propulsion, permitting changes of orbital path. These engines, relatively low-thrust systems, are used only on manned spacecraft and some of the most complex unmanned craft and consequently do not figure prominently in the production work load. Engine manufacturers also produce, in limited quantities, a variety of specialized rockets, including small monopropellant or bipropellant thrusters for adjusting the attitude of an orbiting spacecraft; solid rockets for explosive separation of launch-vehicle states; rockets for high-speed track-borne research sleds; stabilizing thrusters to control roll in missiles and launch vehicles; and aircraft auxiliary rockets.

Nations with substantial aerospace-engine industries, including those that build other nations' designs under license, include the U.S., the Soviet Union, France, Britain, Australia, Canada, China, Czechoslovakia, Germany, Israel, Italy, Japan, Poland, and Sweden.

Other products. *Secondary aerospace products.* The secondary-product line of the aerospace industry is composed of the great many items of on-board equipment required by the various design missions of the flight vehicles. An airframe and a propulsion system sufficed for the initial flights of the Wright brothers, but the modern aerospace vehicle admits of no such simplicity; depending on design purpose, a flight vehicle may have thousands of separate functional equipment units. The list of individual items is too lengthy for even a catalog listing, but a sampling of the more important systems and subsystems illustrates the breadth of the secondary-product line.

Aircraft equipment includes aerodynamic controls and actuating devices; engine controls; cockpit instruments and displays, including instrument-landing systems and other aids to bad-weather approaches; communications systems; radar for weather avoidance, target acquisition, search missions, and identification signals to ground stations; navigation equipment other than basic instruments, including radio, optical, and electronic systems and, in the most advanced aircraft, computer-directed inertial navigators; air-conditioning and pressurization equipment; landing-gear brakes and steering controls; data recorders; trouble-warning systems; radar, photographic, infrared-imaging, and other sensory devices for intelligence gathering, together with on-board intelligence-processing equipment; armament for military planes, including armour systems, launchers for airborne missiles, high-speed cannons and gun turrets, bomb and rocket fuses, and arming devices; electronic countermeasure equipment for confusing enemy defenses; engine starters and auxiliary power units for on-board electricity; seats, galleys, and passenger-service equipment for commercial airlines; and escape devices such as parachutes, surface-survival aids, and, for higher speed aircraft, ejection seats and separable cabins.

Missile and launch-vehicle equipment includes automatic guidance systems, ranging from heat-seeking homing devices to computerized navigation systems; on-board power systems; controls, including actuating devices for course-changing fins, thrusters, or gimballing engines; communications systems, including receivers for reception of command signals from Earth and transmitters for relaying vehicle-condition data to Earth, sensors to assay vehicle conditions, and trouble-warning devices; abort-destruct systems; engine-ignition and cut-off systems; automatic fuel-flow systems for liquid-fuelled vehicles; tank-pressurization systems; warhead systems; and, for strategic missiles, re-entry vehicles protecting the destructive payload from the friction heat of re-entry into the atmosphere.

The category of spacecraft equipment includes an extremely broad variety of experiment packages, including instruments ranging from simple micrometeoroid counters to complex automatic telescopes; communications systems, including antenna arrays, telemetry transmitters for reporting data to Earth, recorders to store data for later transmission, encoders to translate data into coded signals, receivers for acceptance of Earth commands, decoders to interpret command signals, and command-distribution devices for translating commands into compliance; power sources, including solar-cell arrays, batteries, fuel-cell power plants, and nuclear generators; stabilization systems and station-keeping devices to maintain a desired spacecraft position or attitude; spacecraft-condition sensors; radio and optical beacons to facilitate Earth tracking; ejection and descent systems for payloads recoverable from orbit; photographic, radar, infrared, and other types of sensory devices for military intelligence gathering and for civil use in meteorological and Earth-resources-survey spacecraft. Additionally, for manned spacecraft there are environmental-control systems; guidance and navigation systems, including sophisticated computers for acceptance and processing of Earth-acquired data; cabin instrumentation and displays, including trouble-warning systems; special-purpose radars, such as those used for in-space rendezvous or lunar landing; pressure suits for extravehicular activity; and galleys, water dispensers, and waste-management systems designed for operation under zero-gravity conditions.

Experimental equipment

Tertiary aerospace products. The industry's tertiary-product line includes those ground-based items necessary for support of flight vehicles, either directly, such as engine starters, or indirectly, such as training devices. A major area of effort is the production of ground radars and antennas with associated data-processing systems. This equipment is employed for air-traffic control and airport control-tower operations; for detection and tracking of potentially hostile flight vehicles; for the Earth-command type of missile guidance and for interception guidance of air-defense aircraft; and for tracking spacecraft. Another major area is production of automatic checkout equipment for complex aircraft, missiles, and space vehicles; such equipment electronically "interrogates" each of the vehicle's systems and verifies that the system is in proper working condition or pinpoints faults for quick correction. A third area involving considerable work load is fabrication of simulation systems, used for training aircraft, missile and spacecraft

Ground-based equipment

crews, and research and development processes. Among other examples of ground-based aerospace equipment are radio transmitters for aircraft navigation; airport lighting and landing aids; ground-based air-conditioners, power units, and airplane-servicing equipment; transporters, erectors, launchers, and fire-control systems for mobile battlefield missiles; and the systems that make up a space mission-control centre, such as communications circuitry, computers, and display consoles.

Nonaerospace products. Nonaerospace products and systems, general-usage offshoots of aerospace technology, constitute a fourth area of industry work load, not significant as a proportion of total work load except in the United States and presumably in the Soviet Union. Examples of new applications of aerospace-developed technology include many medical aids, ranging from a thimble-sized cardiac-pacemaker power source to a computerized hospital-management system; in transportation, hydrofoil ships and boats, vastly improved railroad cars, truck and bus transmissions, motor-vehicle traffic-monitoring systems, inflatable crash-bag systems for autos, and turbine propulsion systems for both land and water vehicles; in construction, prefabricated low-cost housing and lightweight, superstrong structural materials; in electric power, a number of types of nuclear and nonnuclear power-generating systems; and, in environmental control, a broad range of antipollution systems.

RESEARCH AND DEVELOPMENT

The ultimate aim of aerospace research and development is the creation of flight vehicles more advanced than their predecessors, but so succinct a statement of goal oversimplifies the enormous scope of research and development activity. It is a fact of aerospace life that expansion of the operating capacity—the "envelope" of a flight vehicle—almost invariably entails corollary advances in onboard systems, ground-support equipment, and even the methods of construction. Frequently, desired performance gains dictate not merely improvement of existing ancillary equipment but entirely new inventions. Aerospace research and development is a continual process of across-the-board thrusting into the unknown and frequently moves beyond the traditional areas of investigation into new byways embracing practically all of the technological disciplines.

Types of research and development

Aerospace research and development is of four types. Basic research involves investigations that have no immediate practical objective; it is the advancement of knowledge for potential rather than direct interest. Applied research is investigative effort aimed at the practical application of knowledge. Development, by definition, is "the use of scientific knowledge directed toward the production of useful materials, devices, systems or methods, including design and development of prototypes and processes"; it is the translation into hardware of the results of applied-research studies. Operational research is concerned with the solution of problems encountered by flight vehicles already in regular service. The aerospace industry engages in a certain amount of basic research, but it is negligible as a percentage of total research and development work load. Operational research is conducted to a greater degree, but the primary focus in the industry is on applied research and development related to the betterment of existing products and the emergence of completely new products.

The level of research and development activity naturally varies a great deal among companies and countries, the determining factor being the types of flight vehicle to be built. Light, general-aviation aircraft, for instance, are not overly demanding of research and development. Like automobile manufacturers, light-plane builders offer their products in a number of price categories and cost is a restricting influence on size, weight, and performance. General-aviation manufacturers, whether airframe, engine, or equipment builders, do conduct considerable research, but it is aimed at small increments of improvement that can be obtained without cost increases that would take the plane out of its competitive category.

New commercial aircraft. Development of a new commercial airliner demands extensive research and development, but the requirement is substantially reduced by the fact that the jetliner operates in an already-explored flight regime. Traditionally, military aircraft reach a given plateau of performance a decade or more before the airliner achieves the same capability. Thus, designers of commercial aircraft have available at the outset a vast storehouse of aerodynamic, propulsive, and other knowledge relative to the flight regime in which they propose to operate. This is not to suggest that they can bypass those areas in which prior research has been accomplished; rather, they can use the knowledge available as a departure point for a research and development program in which the emphasis is on utmost efficiency within a known flight bracket rather than initial penetration of an unknown regime.

Jetliner developmental teams focus on two primary areas of research and development: safety and economic viability. Safety involves such major considerations as stronger structures, ultrareliability of engines and on-board equipment, and provision of the most advanced navigation and all-weather landing systems. Economy of operation embraces capacity; maximum flight efficiency, which is translatable into greater range and payload; maximum propulsive efficiency, for reduced fuel costs; lowest possible turnaround time for best utilization of the airplane; longer equipment life for extended periods between overhaul; and ease of maintenance.

Expanding the aircraft envelope

The industry's major research and development effort is expended on those programs that seek major expansion of one or more facets of a flight vehicle's envelope, the bounds of its performance capabilities. More often than not, a speed greater than its predecessor's is the primary performance gain sought for a new vehicle, but the research effort may also be directed toward other elements of improvement. The goal, for instance, might be an increase in range or payload, as is the case in global-ranging military bombers and transports; it might be manoeuvrability, a vital consideration in fighter-aircraft design; it might be greater destructive capability, as in multiple-warhead strategic missiles; it might be duration, as in an extended-stay orbiting vehicle; it might be extremely high altitude, for a reconnaissance airplane; or it might be a combination, as in the case of a new bomber under development, of supersonic speed for long periods at minimal altitudes, a particularly demanding area of research because of different shock-wave characteristics at sea level as opposed to those of high altitude. There are other factors involved, and rarely are the sponsors of a new project willing to settle for improvement in only one area.

High-technology programs of this nature are almost always government sponsored. Because they probe unknown areas of flight, they require exceptionally high funding outlays, millions of research and development man-hours, and an elaborate complex of research and development facilities, some of which may have to be created especially for the particular project. A brief outline of the step-by-step process in the development of a new high-technology vehicle—a high-performance military aircraft, for instance—serves to illustrate the extraordinary human skills required and the complex array of tools and facilities needed to bring such a program to fruition.

Computer design. The design phase is largely theoretical. Drawing upon the general fund of aeronautical knowledge and their own preliminary research, scientific and engineering specialists produce a mathematical model of an airplane that they think will meet all of the project sponsor's specifications. The team's primary tool is the electronic computer, used in a variety of ways. It provides rapid and accurate answers to the millions of mathematical calculations required. It stores in its memory banks the design input fed into it and creates an electronic image of the airplane. Coupled with an automatic plotter, it displays on a screen pictures of the complete airplane or any section of it, such as the wing, flaps, tail section, or nacelles, affording the team a visual presentation of current analysis. Programmed with atmospheric data, it can "fly" the hypothetical airplane and report to the team the aerodynamic response of the design to any condition of flight, from takeoff to landing, from straight-and-level cruise to combat manoeuvres. Through a lengthy series

of such simulated flights, the team refines its design, until the electronic airplane meets or exceeds all specifications.

Wind-tunnel testing. The next step is wind-tunnel testing of a physical model, a carefully constructed scale replica of the computer-aided design, for verification of the performance predictions. The wind tunnel, which actually predates powered flight by more than 30 years, is an apparatus in which air is blown over a model mounted in a test section, creating an effect comparable to airplane flight. There is wide variance in the size of wind-tunnel test sections, depending upon airflow-velocity requirements. Some low-speed tunnels have test sections large enough to accommodate a complete small airplane or a wing-nacelle section of a larger plane; in a higher speed system the test model may be only the size of a desk model, and in the very high velocity tunnels used for missile and space testing the model may be as small as a match. Tunnels are classified according to airflow-velocity range in five categories: subsonic, up to Mach 0.8, or 80 percent of the speed of sound; transonic, Mach 0.8 to Mach 1.2; supersonic, Mach 1.2 to Mach 6; hypersonic, Mach 6 to Mach 12; and hypervelocity, above Mach 12.

Methods of operation of wind tunnels

There are two basic methods of feeding air to the test section. In the continuous-flow tunnel, airflow is generated by a battery of fans or compressors, forced through the tunnel to the test section, then directed through a curving tunnel section back to the starting point for reuse. The alternate method involves storing high-pressure air in a tank, releasing it by valve controls to pass through the tunnel and test section into a capture tank. This type of system, the intermittent-flow tunnel, is generally used for very brief bursts of high-velocity airflow.

Some wind tunnels are almost as complex as the flight vehicle under development. They have variable throats to permit changing airflow velocity, conditioning systems to change air characteristics for simulation of different altitudes, and heating devices to simulate the friction heating encountered by high-speed flight vehicles. Mounted within the model are tiny transducers that acquire a variety of data and transmit it to recording systems outside the tunnel. Specially designed colour cameras photograph shock waves and other airflow phenomena. The computer digests the data from the test run, analyzes it, and presents a report on the flight vehicle's performance.

Most major manufacturers have some type of wind tunnel, usually of the intermittent-flow type, and for more demanding programs they have access to the complex government-operated supertunnels that no single company could afford. After a series of wind-tunnel "flights," the design may be further refined to incorporate the findings of tunnel research; then the airplane is ready for prototype construction.

The prototype. In the prototype-construction phase, emphasis shifts from research to testing. A customary procedure is to build one or more static test airplanes, designed not to fly but to verify the structural integrity of the design. This requires another array of facilities, including ovens for applying high temperatures to materials; acoustic chambers, in which engineers can study the effect of high-frequency engine noise on structures; rigs for measuring landing impacts; and rapid-cycle vibrators for investigations of wing-panel flutter. A major facility is the load-test rack. A flight load is the force experienced by the airplane in its various manoeuvres; designers provide for an ultimate load factor, which represents the greatest load the airplane may encounter in the most violent manoeuvre plus a comfortable safety margin. The load-test rack consists of a series of instrumented hydraulic jacks that push, bend, and twist the airplane to the point of failure, verifying that the ultimate load factor called for in the design has been met in construction.

While the prototype airframes are being built, the manufacturers of on-board equipment are going through a parallel process to devise and test the airplane's many systems and subsystems. The broad variety of ancillary equipment precludes description of the testing process, since it differs with each system. Generally, equipment is checked out, repetitively and exhaustively, by a battery of electronic test equipment tailored to the system being tested. In the laboratory, the test set electronically monitors the system as it is run through its performance cycle and signifies okay or detects and isolates faults for correction. In many cases, complete systems are further checked in altitude chambers that simulate the environment in which they will later be operating. Engines are tested in the propulsion research equivalent of the wind tunnel, the test cell capable of simulating the conditions of flight; test engines are heavily instrumented, and the data findings are transmitted to a computer for processing. To qualify for installation in a prototype airplane, a new engine is normally put through several hundred hours of cell testing, embracing the entire range of speed and altitude capacity of the airplane for which it is intended.

Flight testing. As a general rule, prototype flight testing is conducted at government-operated installations in lightly populated areas, due to the possibility of accident to an experimental airplane. Flight testing is necessary to validate what has been to this stage really only theory, but so thorough and extensive are modern procedures of computerized design and tunnel testing that the flight-test phase rarely dictates a major design change. Rather, it serves to debug the airplane, to correct the many minor faults that somehow find their way into a new airplane despite the most painstaking efforts to avoid them.

Modern flight testing bears little resemblance to the once-familiar motion-picture portrayal of the daredevil pilot trying to fly the wings off the airplane. The modern test pilot is a thoroughly experienced flyer who is also an engineer; he is selected for his flying ability and for his ability to observe and report in technical detail on the handling qualities and peculiarities of his airplane. He is backed by a forest of sensitive test instruments, gathering volumes of data, which are relayed by radio telemetry to computers on the ground. The flight program is carefully planned, again with the help of a computer. It begins with ground-taxi tests and progresses to brief flights at the lower end of the performance spectrum, then to longer flights of gradually increased performance until the limits of the design envelope have been reached or surpassed. The test program for an advanced airplane may take several months and complete evaluation of the data additional weeks or months.

The modern test pilot

Missiles and spacecraft. *General considerations.* The research and development effort that goes into development of missiles, launch vehicles, and spacecraft parallels that of the airplane in the design and ground-test stages but differs when the flight-test stage is reached. The shorter range battlefield and tactical weapons pose no major testing problems; they are fired more or less like artillery systems at large-area land ranges. Major launch vehicles and strategic missiles are more demanding as regards test techniques. The absence of man on board, the great expense of a single launch, and the inability to recover and re-use the test vehicle all combine to exert two major influences: more elaborate instrumentation, both in the vehicle and on Earth, and extraordinarily intensive preflight checkout to prevent a costly abort.

Additionally, the distance covered in a major unmanned flight-vehicle test, which is several thousand miles in the case of an intercontinental ballistic missile, requires the use of spread-out test ranges with down-range tracking and data acquisition stations. These ranges are necessarily government operated, but generally contractors have their own facilities at the launch centre for prelaunch testing. In the United States the U.S. Air Force Eastern Test Range includes both the John F. Kennedy Space Center and the Cape Canaveral Air Force Station, launch bases for civil space flights and military missile testing; surface stations extend 10,000 miles down the South Atlantic. A second major United States facility, the Western Test Range, centred at Vandenberg Air Force Base, California, extends into the Pacific. The Soviet counterpart of Cape Canaveral is the missile-space-launch centre at Tyuratam, on the Aral Sea, with impact areas for long-range flights in the Pacific Ocean. The Soviet Union operates two other major facilities, one at Plesetsk, near the White Sea port of Arkhangelsk, the other at Kapustin Yar, north of the Caspian Sea. France operates a range that extends into the

Major world testing ranges

Atlantic from a base at Kourou, French Guiana. Other major ranges are operated by Australia and Japan, with launch centres at Woomera and Kagoshima, respectively.

Unmanned spacecraft are unique in that they undergo no prototype test flights; if their launch vehicles perform successfully, they go into orbit, so they carry their full complement of operational instrumentation. In a sense, then, all unmanned spacecraft are test vehicles, since they report details of their own operation as well as scientific-data findings, providing the basis for design refinement in later models of the same family. The substitute for prototype testing is Earth-based simulation, usually in two types of simulators: the space simulator, which duplicates all the environmental conditions in which the spacecraft will operate, and the mission simulator, which permits conduct of the entire range of manoeuvres and systems operations that would be performed on an actual flight.

Manned spacecraft are not only prototype tested but are subjected to the most rigorous ground-test program experienced by any flight vehicle. The program defies detailed description, since it involves literally millions of separate tests. Some idea of the intensity of the test effort and the broad variety of research and development facilities needed may be obtained by a recounting of the major ground-test operations of a manned spacecraft, such as the United States' Apollo mooncraft.

Testing. Although most of its lunar mission took place outside the atmosphere, Apollo did fly within the atmosphere briefly en route to orbit and during re-entry and parachute descent to the ocean. To verify design predictions as to aerodynamic loads, acoustic effects, friction heating, and lift-to-drag ratios at the hypersonic speed of re-entry, Apollo was put through more than 7,500 hours of wind-tunnel testing.

The environmental-control system was tested at conditions ranging from sea-level pressure to hard vacuum.

In impact tests a full-scale command module was slammed into a water tank by a huge pendulum arm, simulating various angles of splashdown and impacts much greater than would be encountered on a normal descent.

The complete spacecraft was placed in a large facility capable of simulating the total space environment, including hard vacuum, solar radiation, and temperature extremes, to determine effects on systems and materials.

In space, Apollo was at times positioned so that one side was in the undiluted sunlight while the other was in cold shadow. For simulation of these simultaneous extremes of temperature and observation of structural effects, the spacecraft was electrically roasted on one side at 600° F (320° C), while the other side was sprayed with liquid nitrogen at −320° F (−200° C).

In a facility that is the spacecraft equivalent of the hydraulically operated airplane load test rack, Apollo was subjected to immense loads far greater than any normally anticipated.

Dynamic tests included acoustic, vibration, shock, and acceleration investigations. A huge "cocktail shaker" rapidly shook the complete spacecraft to simulate the vibration and buffeting of the launch phase.

The above are a few samples of a broad range of ground tests conducted repetitively over a period of several years. Running concurrently was a series of unmanned suborbital and orbital flights. For the initial series the contractor built a number of replicas of the actual spacecraft in size, shape, structure, mass, and centre of gravity but not equipped with the myriad functional systems. The replicas were used on 10 flights to test the spacecraft's compatibility with its launch vehicle, the launch escape system, and the general stability and structural integrity of the spacecraft in orbital flight. These tests were followed by six additional flights in which completely equipped but unmanned spacecraft were flown in orbit to check out all systems under actual space conditions and to verify the performance of the re-entry heat shield. For these flights, the stand-in for the human crew was a robot controller capable of directing systems operation either by preprogrammed instructions or by command from Earth monitors. The meticulous preparation for each flight, coupled with the evaluation of the vast amount of data acquired, stretched the 16 unmanned flights over a period of four and a half years.

The boilerplate test series

Other areas of research. In addition to project-oriented research and development, the effort involved in a specific flight-vehicle program, the aerospace industry conducts a tremendous amount of research and development of a general nature designed for competitive readiness for the next generation of aerospace products. Among the more active areas of general research are:

Propulsion. Anticipating continuing demand for greater thrust, particularly in the airplane field, engine manufacturers are designing and developing new families of superpowered propulsion systems. In addition to thrust ratings, research is aimed at lower engine weight in relation to power output, better combustion efficiency, and further extension of the time between overhauls. A special line of research in recent years has been concerned with reduction of engine noise and emissions.

Electrical power. Continually increasing flight-vehicle performance means more on-board systems and greater capability of the individual system, hence a requirement for more electrical power to operate the systems. Manufacturers are paying particular research attention to fuel-cell power plants and thermoelectric converters, along with a variety of conventional generators of more advanced design.

Materials. Probably the greatest single influence on aerospace design is weight, and researchers are constantly looking for new materials that offer weight reductions without sacrifice of strength. The quest extends to materials for all parts of all types of flight vehicles, but the greatest emphasis is on temperature-resistant materials for high-performance vehicles. Conventional materials weaken under extended exposure to combustion-induced heat, solar-induced heat, friction-induced heat, and hot-cold extremes. Seeking simultaneous improvement in temperature resistance and strength-to-weight ratio, researchers are exploring in particular such exotic metals as columbium, molybdenum, and tantalum; ceramics; composites, combinations of two or more materials in which the resultant has better properties than the individual constituents; and vehicle-skin coatings for prevention of oxidation.

Significance of weight in aerospace design

New systems. "New," in this context, means not developmental improvements of existing equipments but completely new systems that do not exist in operational status and for which a need is clearly indicated. An example is an automatic collision-avoidance system for aircraft.

Miniaturization. A traditional method for reducing weight and at the same time cutting down on systems volume requirements is making things smaller. The missile-space era brought such great strides in this area that manufacturers call it microminiaturization. This type of research is continual, first because the trend in high-performance flight vehicles is generally toward more equipment and second because the initial version of a system, designed primarily for performance, usually leaves considerable room for later miniaturization.

New manufacturing processes. A large portion of general research and development work is devoted to research in new manufacturing processes, in order to combat ever-mounting costs, reduce fabrication time, and assure greater product reliability. Additional effort goes into development of machines and techniques for shaping and forming the new materials being introduced to production status, which almost invariably are more difficult to handle than their predecessors.

Cooperative programs. As aerospace flight vehicles and systems grow in performance capability and consequently in complexity, research and development costs mount steadily. Even the United States and the Soviet Union, with their vast resources, have found it difficult to finance essential research and development, and among the smaller nations financial limitations would make it necessary to bypass promising areas of research and development were it not for the cooperative-program approach.

In western Europe the cooperative approach has been growing for two decades and has reached significant proportions. In these programs, two, three, or more nations pool their financial resources and their facilities on projects that would be beyond the means of any one of them.

In some cases, developmental responsibility and financing are divided equally among the participants; in others, one nation assumes the major burden but is backed by associate participants. Project management may be assigned to a single major company, or the project may be directed on a consortium basis, with alternating national leadership.

In this fashion, European countries are accomplishing considerable research that would otherwise not have been undertaken. Several have banded together as the European Space Agency; a parallel group is the European Launcher Development Organization, which seeks to develop and produce space-launch vehicles. Other projects are more directly oriented toward commercial markets; the British-French Concorde supersonic plane is the prime example, but there are many others, involving jetliners, military aircraft, helicopters, missiles, turbine engines, and a communications satellite for a contemplated European network. Participants in cooperative programs include France, the United Kingdom, the Federal Republic of Germany, Italy, Sweden, Belgium, Switzerland, The Netherlands, Spain, and Denmark; Australia, which supplies a launch range, is also a participant in the European Launcher Development Organization.

The Concorde project

MANUFACTURING FACILITIES AND TECHNIQUES

No one company builds an entire flight vehicle, not even the simplest light plane, because of the exceptional range of skills and facilities required. Aerospace manufacturers are specialists, operating as either independent companies, divisions of very large corporations, or bureaus of a nationalized industry (among the major producers the Soviet Union is completely nationalized, France almost exactly half and half as regards the personnel division between the *secteur nationalisé* and the *secteur privé,* and the United States and United Kingdom entirely private industry). The primary areas of specialization are airframes, propulsion systems, airborne electric and electronic systems, other airborne systems, and ground-based equipment. Within these broad areas there is sharper focus of specialization; for instance, one electric-electronic company has separate divisions manufacturing guidance and navigation systems, control systems, electric-power systems, instruments and life-support equipment, and scientific-data-acquisition systems.

A major aerospace production program may involve literally thousands of individual suppliers operating as subcontractors to a team leader, usually the airframe builder, known variously as the prime contractor, principal constructor, "mother" manufacturer, systems integrator, or design bureau, the last a Soviet term that embraces not only design but also supervision of the complete manufacturing process. Subcontracting covers not only the onboard equipment but, in most major programs, the airframe itself. Production of the British-French Concorde serves as an example: the airframe is broken down into 21 separate components built by 11 different companies and divisions, then delivered to the team leader's plant for assembly. Work-load sharing of this nature takes advantage of additional pools of skill and facilities and helps reduce production time. On commercial projects it is also employed as a method of spreading the requisite financing for materials, tooling, and other production preliminaries.

Basic fabrication techniques. The variety of the product line, from individual parts to major systems to the complete flight vehicle, precludes description of a typical manufacturing plant, since each factory is tailored to the items produced. In general, the primary work areas of any plant are the fabrication facilities, in which the individual parts and components are made, and the assembly facilities, ranging from moderate-size rooms to huge, high-bay buildings covering several acres.

Metal processing. In the fabrication stage, each of the thousands of parts must be precisely shaped, and there are a number of ways of doing this. Fabrication tools and techniques fall into two major categories: chipless operations, performed without cutting tools, and material-removal operations. Chipless operations include die-casting, or pouring molten metal into a mold; molding, which is casting under pressure; and a number of methods of pressure forming. Aircraft skins, for instance, are made by rolling machines that squeeze and flatten hot sheet metal to the desired thinness; other parts are made from metal sheet, plate, foil, ingot, and bars by stamping; by forging, or pressing the metal over dies; by extruding, or forcing heated metal through a die, like squeezing toothpaste; by stretch forming, giving the metal a permanent bend by stretching it over a curved die; or by braking, pressure bending a clamped metal workpiece. A relatively new method is explosive forming, in which an explosive charge is detonated against a water-submerged workpiece and the resulting blast slams the metal against a shaped mold.

Metal-removal operations

Metal-removal operations are primarily those in which metal is trimmed away from a workpiece by tools such as milling machines, lathes, shearing machines, borers, punch presses, knives and saws, sanders, grinders, and polishers. In addition to these cutting and friction tools, metal removals are accomplished by electrical devices, such as an electrical-discharge borer, and by chemical means; in the latter method, an acid bath eats away the excess metal while a coating protects the desired shape. In either chipless or metal-removal operations, the particular technique selected for a given part is dependent upon a number of factors; for example, the type of basic material, the intricacy of the shape, requisite strength, and, when it becomes a part of the flight vehicle, the degree of exposure to load forces and temperatures.

Joining methods. Individual parts, once fabricated, must be mated, or joined, with other parts. The most familiar method is mechanical joining by means of rivets, bolts, screws, pins, and other special-purpose fasteners. In cases in which a tighter seal is necessary or where the assembly will be subjected to great loads, manufacturers employ welding, which in most cases is accomplished by melting the metal at the joint, applying molten metallic filler, and allowing the joint to cool. Electric-arc welding, in which heat is supplied by an electrical circuit created between the workpiece and the welding tool, is in wide aerospace usage. Other methods include cold welding, or joining by pressure without high temperature, and forge hammering with heat application. A more advanced technique, diffusion bonding, involves the application of both heat and pressure in a nitrogen atmosphere, creating an exceptionally strong, lightweight joint. Another joining method in wide use is adhesion bonding, or simply gluing the parts together by means of supercapable adhesive coatings, such as plastic resins.

In the production of components that must bear high loads yet be as light as possible, aerospace fabricators have evolved a technique for changing the properties of a material by engineering. The most notable example is the so-called honeycomb sandwich. The core of the sandwich is the honeycomb, a structure composed of row upon row of framed cells, or holes, resembling the honey-storage facility of a beehive. To each side of the core are bonded extremely thin sheets of metal, creating the sandwich, which is far lighter yet has greater resistance to bending than a comparable thickness of metal plate. Aluminum is the most used metal, in both the core and the facing sheets, but the technique is applicable to a great many metallic and nonmetallic materials. Sandwich construction is a major part of aerospace manufacturing, one used to some degree in almost every type of flight vehicle.

The honeycomb sandwich

Most of the basic fabrication techniques have been employed for years in aerospace manufacture, and the difference today is in the degree of sophistication of the tools and techniques; like the product line itself, tools and techniques have progressed through several plateaus of technological advance. For example, simple hand-torch welding would not suffice for the stringent joining requirements of a current spacecraft; it is necessary to enclose the welding station in a pressurized, temperature-controlled, contamination-free portable room. Similarly, while forming machines perform the same functions as their forebears, they differ considerably because of new and more difficult to handle materials, such as titanium, magnesium, and beryllium in airframe construction and the superalloys of iron, nickel, and cobalt in engine production. Additionally, the modern machine must work to

incredible tolerances to produce parts of greater precision than the finest watch movement. Not long ago, machining to a thousandth of an inch was considered a tremendous accomplishment; some modern systems demand tolerances measured in millionths of an inch.

Use of electronic computers. The primary difference between today's manufacturing facility and its World War II or immediate postwar counterpart is the machining precision and efficiency made possible by the electronic computer. In a process broadly known as computer-aided manufacturing (CAM), computers are employed to direct the operations of machine tools. CAM embraces a number of types of computing equipment and various methods of controlling the machine. In general, the computer serves as a translator between the human programmer and the machine tool. The design of the part to be manufactured is reduced to symbolic computer language, and the programmer feeds the computer instructions as to the necessary machining operations. The computer in turn relays the instructions, in the form of command signals punched on tape, to an automatic controller, which directs the machining with extreme precision. Advanced CAM systems include multiple-operation machines, equipped with a battery of different tools; by computer-direction, the machine can perform any combination of operations in a prescribed sequence and even change the tools as required. Some huge tools can turn out large single-piece components and structures of intricate shape, eliminating the need for joining several individual parts. Computer-aided manufacturing shortens production time, thereby reducing cost, and in addition makes possible greater accuracy in parts fabrication.

The CAM system

Product assurance. Reliability is an important factor in the manufacture of any type of vehicle, but it assumes obvious urgency when the vehicle is one that flies. In aerospace production, exhaustive measures are taken to ensure utmost reliability and to guarantee that the completed vehicle or system or subsystem meets all the specifications prescribed by the end user of the product. The terms vary by company and by nation, but generally the collective precautionary measures are known as product assurance, which includes two subdivisions of effort: reliability engineering and quality assurance.

Reliability engineering, which begins in the design phase, essentially is a process of intensive study of alternative methods of performing a particular function and selection of the approach with the best predictable reliability. Reliability-engineering teams are frequently at odds with other designers, because the most reliable system is often one that demands compromise in such matters as cost and weight. An example is the practice of assuring reliability by means of redundant systems; a particular unit or system is provided an identical backup, together with a monitoring package that automatically activates the backup when it detects a malfunction of the primary. In some flight vehicles triple redundancy (two backups and a primary) is employed.

Reliability through redundancy

Quality assurance encompasses the entire manufacturing effort. It begins, as a management function of the project-team leader, with selection of the multitude of suppliers. The team leader studies the suppliers' own capability estimates, inspects their facilities, and investigates their history of performance before assigning a production task.

Quality-assurance groups focus a good deal of their attention on people. Despite the high degree of automation and the sophistication of fabrication tools, human effort still plays the dominant role in aerospace production. Quality assurance reaches into the processes of selecting skilled personnel and training employees. By lectures, bulletins, and posters on the potential consequences of carelessness, the quality-assurance group seeks to reduce human error. Some plants adopt extraordinary error-reduction measures; for example, some producers, to guard against the possibility of a potentially damaging tool being left behind in an engine assembly, provide workers with fitted toolboxes that are inspected after each shift.

The major aspect of quality assurance is quality control, the process of monitoring every step of the manufacturing cycle, from reception of raw material through delivery of the finished product. Quality-control technicians inspect new material coming into the plant to make sure that it is the specified type and has adequate properties. As the material is formed into parts and then subassemblies and assemblies, items that are considered critical are inspected after each work step, sometimes visually, sometimes by X-ray, microscopic, chemical, and ultrasonic devices that uncover invisible or subsurface cracks, voids, and other defects. Noncritical items, those whose failure could not influence the safety of the flight vehicle, are inspected on a spot-check basis.

Functional systems are subjected to repetitive tests after each workstep to insure that they operate exactly as they are supposed to when incorporated into the flight vehicle. There are hundreds of different types of testing equipments, each specially designed for a particular product, the bulk of them electronic testers that check out the operation of the product and report any deviation from specified performance. There are tests of individual parts, of assemblies formed by a number of parts, then of complete systems. Systems are checked out separately in their assembly facilities, later in their actual position in the flight vehicle.

Assembly methods and facilities. Assembly of an aerospace vehicle and the systems that go into it is a pyramidal process. The base of the pyramid corresponds to the mating of two or more individual parts designed to work together. Moving up the levels of the pyramid, these parts assemblies are joined with other assemblies to form subsystems, then systems. The terms system and subsystem are not sharply defined, other than that either is an assemblage of parts or equipment that *in toto* perform a particular function in the operation of the flight vehicle. Thus, in military terminology, a strategic missile is a system, and its inertial navigator is a subsystem; in the plant of the guidance-equipment manufacturer, however, the inertial navigator is considered a system, composed of such subsystems as the computer, the gyroscopic sensing platform, and the electrical-power source and circuitry. Few on-board systems are manufactured in a single plant; in most cases, parts and subsystems are produced elsewhere and delivered to the prime-system contractor for assembly.

At the top of the pyramid is assembly of the complete flight vehicle in the plant of the team leader, to which flow from a great many plants the airframe components or subassemblies, on-board systems and subsystems, and a variety of individual items that do not merit classification as systems.

The types of facilities employed in the assembly of on-board equipment differ considerably, as do assembly-line methods. Generally speaking, this type of aerospace product, small by comparison with airframe components, is assembled in low-bay buildings containing a number of separate work stations at which specific steps in the assembly process are accomplished. The growing assemblage is moved from station to station by conveyor or by wheeled cradling fixtures, frequently inspected or tested along the way. Some relatively simple equipment needs no assembly line; it is put together, inspected, and tested all at one work station.

In extensive use among aerospace manufacturers is the "clean room," where assembly is conducted in a contamination-free atmosphere that rivals the surgical theatre. Clean-room assembly is imperative for a great many systems that go into missiles, launch vehicles, and spacecraft (and, to a lesser extent, aircraft), complex equipments that are especially sensitive to environmental contamination; a particle of dust or a temperature outside certain rigid limits could cause a malfunction.

"Clean room" assembly

A typical clean room has an atmospheric-control system that rigidly controls temperature and humidity and bars entrance, by means of filters, of all but the tiniest mote of dust. Walls and ceilings are of one-piece plastic with no cracks where dust might collect and are washed and vacuumed daily. Maintenance cannot be done within the room: plumbing, wiring, and lighting are so arranged that maintenance can be handled in crawl spaces above the ceiling. The room has no sharp corners; they are rounded off to forestall dust collection.

Before entering, workers don special clean suits, including head covering and boots, and pass under an "air shower" that removes all loose particles of matter. The parts that make up the assembly are thoroughly cleaned and polished before delivery to the clean room, which they enter through an air lock.

Clean-room assembly is expensive and is used only when necessary, in most cases for the late stages or the final phase of the assembly procedure. The size of the room varies with the product; it may be from living-room size to a huge area considerably larger than a football field. Introduced in the latter 1950s with the development of sophisticated strategic missiles, clean-room assembly expanded over the years as systems became more complex.

Final assembly

The final assembly of the complete aerospace vehicle usually takes place in a large structure with a very lofty ceiling and no interior partitions or columns. Major aircraft require the tallest bays; some have tail sections reaching more than 65 feet (20 metres) above the floor. Additionally, the facility must accommodate a network of overhead rails over which ride heavy-lift cranes capable of moving large segments of the airplane.

Although flight-vehicle assembly bears little resemblance to systems assembly, the technique is the same in terms of movement through progressive work stations. At each work station, equipped with its own tools and workstands that provide access to all parts of the segment being worked upon, a certain number of assembly operations are performed according to a carefully considered manufacturing plan.

Assembly of an airplane begins with subassemblies. A typical subassembly for an airline-type plane is the rear fuselage section, itself composed of several segments. The segments are dollied or craned to the subassembly area, fitted together, and joined into a unit, and certain of the interior equipment is installed. In similar manner, plant teams put together other subassemblies; that is, other fuselage sections, wing sections, the tail section, and engine nacelle. The various subassemblies then join the main assembly line.

Mainline operations normally start with the joining, or mating, of two fuselage sections craned onto a supporting jig, or fixture. When all of the attendant joining and interior operations are completed, the embryo airplane moves to the next work station for mating with another subassembly. Light- and medium-weight aircraft may be moved on wheeled fixtures; heavier aircraft may be craned; and movement of modern supersize planes is accomplished by an adaptation of the air-cushion technique. The supporting fixture is a massive steel structure with a great many holes in its underside. Compressed air is pumped into the structure to escape downward through the holes. The thrusting force of the escaping high-pressure air lifts the entire structure and assembly off the floor a fraction of an inch, enough to permit movement of the assembly by an ordinary tractor. In this fashion, the budding airplane moves through a series of work stations, acquiring additional subassemblies one at a time and gradually accumulating its on-board systems and the maze of wiring, ducting, control cables, and other interior plumbing. Major steps include the matings of nose and tail sections, the wing, the engines, and the landing gear. Upon completion of work at the last station, the airplane is rolled out of the assembly plant to the flight line for production flight test, a process that by no means compares with the exhaustive research and development testing but that nonetheless involves a thorough check-out of specified performance.

There are many types of small missiles that require no such elaborate techniques or facilities. Composed basically of a cylindrical shell, a warhead, a guidance system, and a rocket motor, they are readily assembled in a simple low-bay plant. The larger missiles of the ballistic type and space-launch vehicles are assembled aircraft fashion in a high-bay facility, although most of them, horizontally mounted on the line, do not require the extreme bay heights of modern jumbo jet planes.

Spacecraft are unique among the flight vehicles because final assembly takes place under clean-room conditions. Preliminary assembly covering the mating of the structural elements is carried out in a conventional work area, after which the spacecraft goes through an air lock into an atmosphere-controlled chamber, where it is vacuumed and, in some cases, machine tumbled to shake loose dust particles. It then moves on to the final assembly clean room. Some spacecraft are assembled at progressive work stations; others remain in a fixed position while teams of specialists successively install the myriad on-board systems. Since the spacecraft gets no flight test, the equivalent, an intensely detailed acceptance checkout, is handled in the clean room by automatic test equipment. Then the spacecraft is encased in a steel container for shipment.

Maintenance facilities. The maintenance support provided by aerospace-industry firms revolves mostly around aircraft of the corporate, airline, or military types. Light-plane maintenance is generally handled by local fixed-base operators, not considered a part of the aerospace industrial complex. Launch vehicles and spacecraft, although they are maintained throughout their prelaunch life by constant checking and correction, are single-use systems. The smaller military missiles are maintained in the field by specialists in the operating unit. Major missiles similarly undergo routine maintenance at their field installation or silo base, but certain types of work are beyond the capabilities of field units. Overhaul of an inertial guidance system, for instance, requires special equipment and clean-room assembly, so it is usually returned to the plant of the original builder for re-manufacture.

Routine maintenance, normally accomplished by the civil or military operator, includes frequent inspections, sometimes after every flight, sometimes after a series of flights or a stated interval of time, and minor maintenance, such as replacement of a part or repair of a faulty item of equipment. This type of maintenance can be handled at most airline terminals and military bases.

Major maintenance of aircraft

Major maintenance involves complete rework of an airplane or engine that has had considerable service time. Some of the larger airlines have their own extensive facilities for major overhaul, and major military air forces have similar facilities. Even in these cases, however, some major maintenance work is contracted, as is the work of all other operators who do not have adequate facilities, to maintenance-specialist firms operating either independently or as a division of a major aerospace company.

These firms require facilities almost as elaborate as those of the original airplane or engine manufacturer: high-bay plants or hangars, machine tools, many types of inspection and test systems, engine-test cells, disassembly and assembly tools, parts- and components-cleaning equipment, equipment for overhaul of many of the on-board systems, and a complete inventory of spare parts. Maintenance companies are closely monitored as to capability and safety standards by appropriate government agencies. A maintenance firm must be certified by the agency for each type of product it overhauls, and a repair operation must be supervised by an agency-certified mechanic. Maintenance companies operate under two additional guidelines: the original manufacturer of the equipment provides a manual outlining maintenance requirements and detailing specifications that must be met, down to the tolerance of an individual part, and the operator provides a similar manual, which lays down his own demands as to maintenance procedures and standards.

Airframe and engine overhaul. In a major airframe overhaul, the entire structure is disassembled, and each component is visually inspected for wear or damage. Additionally, structures are examined by X-ray, fluorescent, ultrasonic, and dye-penetrant methods to detect defects not visible to the eye. Corrosion is removed by sandblasting or vacuum blasting. Defective components are repaired or replaced, sometimes requiring machining operations to make a part. The airplane is reassembled, in much the same manner as it was originally assembled, repainted, then rolled out to the flight line for post-maintenance ground and air test, which is as thorough as the original production test.

Engine overhaul procedure

The procedure is similar in engine overhaul: complete disassembly, down to individual parts; removal of grease or carbon from each part; inspection of individual parts

and measurement by precision equipment to insure conformance with the tolerance limits tabled in the manufacturer's manual; repair or replacement of any part that is even marginal; then reassembly via the progressive-workstation technique. After assembly the engine is mounted in a test cell and run through a lengthy series of tests as directed by the manuals. In all maintenance and overhaul operations, whether on airframe, engine, or accessories, technicians follow the same quality-control procedures that were in effect during original manufacture.

Remanufacture process. The most elaborate type of program under the general heading of maintenance is the remanufacture process, a military measure that combines a general overhaul with an updating of the airplane's performance capability. It is customary practice in aircraft manufacture to progress through several models of a basic airplane type, the later models incorporating design changes and better on-board systems dictated by service experience with the original model. Thus, older aircraft differ considerably from advanced models of the same type. If it is decided that a particular model, in service for several years, still has years of useful life, it is cheaper to upgrade its performance by remanufacture than to build additional aircraft of the advanced model.

Remanufacture is usually handled by a division of a major airframe company—not necessarily the one that built the plane originally. The remanufacture facility needs additional capabilities; it must, for example, have design and engineering groups for the extensive modifications needed, and it must have the ability to manufacture large structural segments of the airplane if design changes so dictate. In this type of work the airplane is disassembled and stripped of its interior equipment and its engines, which almost always are replaced by improved models. After inspection and repair, the airframe segments that can be retained move through progressive work stations for step-by-step mating with new components. Similarly, the on-board equipment that is still usable is overhauled and delivered to the work stations for reincorporation, along with the new systems that design progression has specified. Reassembled, painted, and production-tested, the airplane begins life anew, a machine in most respects superior to the original. (J.J.Ha.)

Ship design and construction

Ship design and construction today is a complicated compound of art and science. In the great days of sail, vessels were designed and built on the basis of practical experience; ship construction was predominantly a skill. With the rapid growth and development of the physical sciences, beginning in the early 19th century, it was inevitable that hydrokinetics (the study of fluids in motion), hydrostatics (the study of fluids at rest), and the science of materials and structures should augment the shipbuilder's skill. The consequence of this was a rapid increase in the size, speed, commercial value, and safety of ships.

HISTORY

From antiquity until the 19th century. Man has been a boatbuilder since remotest antiquity. It is probable that primitive man first took to the water on a log or a raft. Later, the invention and application of paddles enabled man to progress from rafts to dugout canoes. The paddle led eventually to the lever and so to the use of oars. Much later came the harnessing of wind and the application of sails. Excavations in Crete have shown that seagoing craft were constructed in the eastern Mediterranean as early as 5000 BC. In ancient Egypt, ships built of planks were used on the Nile. By about 2800 BC the basic principles of boatbuilding had evolved, with transverse and longitudinal strengths provided by different parts of the structure. The largest Roman ships, which exceeded 100 feet (30 metres) in length, were carvel-built; that is, the oak planks were laid flush at the seams over closely spaced ribs. These craft were propelled by oars and by a primitive form of rectangular sail. Later, the triangular lateen sail on its long slantwise yard appeared on Roman ships; this had first been used by Arab dhows sailing the Red Sea and Indian Ocean. A striking development in Scandinavian waters was the Viking ship, of fine form, but clinker-built—that is, with the planks on its sides overlapping, and propelled mainly by oars and carrying a simple sail. The ship was open and undecked, of wood-plank construction with cross beams. A seagoing ship peculiar to Chinese and Japanese waters was the junk, a flat-bottomed boat with square sails of matting.

Wooden seagoing vessels

Progress continued from the Middle Ages onward, with considerable advances in sail design. Sails entirely supplanted oars in all large European vessels, except in the Mediterranean. As ships grew larger, they were decked over, and more decks were fitted with "castles" at bow and stern. In Portugal, Henry the Navigator (1394–1460) encouraged the building of new types of ships. Portugal's achievements in navigation were due in part to Henry's enlightened policy of experiment and trial on actual vessels. Sporadic attempts to test ship forms were made in Italy, France, Sweden, and England but to little practical effect. Ships remained essentially unchanged from the 15th to the 19th century, when a fast new vessel of revolutionary design was constructed in North America in 1812. The clipper ship followed, a triumph of empiricism and craftsmanship; it achieved very high speeds because its hull offered little water resistance, and it carried a great area of sail.

Developments since the 19th century. A more profound revolution in shipbuilding followed the substitution of steam engines for sails and the substitution of iron for wood in hull construction. Iron, in turn, was replaced by steel, with a substantial saving in weight. In the decades-long period of transition from sail to steam, side-paddle wheels were first used but were gradually ousted by the screw propeller. The first propeller-driven ship was put into service in 1839, but it was not until about 1850 that steam navigation entered upon its era of unprecedented development, during which the steam engine evolved into a propulsive force of great power and economy. At the turn of the 20th century, the steam turbine appeared and soon became a rival to the reciprocating steam engine, eventually supplanting it. More powerful than even the largest reciprocators, the steam turbine, because of its comparatively small size, greatly changed the size and proportions of ships, and cut costs of operation.

Steam navigation

The first decade of the 20th century also saw the heavy-oil engine come into use. Though the first of these was not notably successful, by 1912 the marine diesel engine was established. From that time onward, the steam turbine and the heavy-oil engine enjoyed a keen rivalry. They were joined later by the gas turbine and the nuclear power plant.

Effect of science and technology. From early days shipbuilding was practiced in most of the maritime European nations, notably in Britain, The Netherlands, Germany, Italy, and in the countries of the Baltic Sea. Before 1850, U.S. shipbuilders were supreme in the building of sailing vessels. The U.K. industry was at that time handicapped by the shortage of suitable timber supplies; moreover, it could not compete in price with either U.S. or Baltic shipbuilders. The introduction and development of iron and steel ships, however, gave the U.K. an enormous, if temporary, advantage because of Britain's wealth of coal and iron, together with its early and successful application of steam propulsion.

Modern shipbuilding

For a time old and new practices were mingled; sails were used to drive iron ships, wooden hulls were propelled by steam engines, and composite ships—framed with iron sections and planked with wood—were driven by both systems. The size of wooden ships had been limited by timber's inherent shortcomings as a structural material; ships made of iron and steel could be larger.

Despite the rapid progress in the design and construction of ships, the greatest technical innovations were in their engines. After a century of development, the steam engine was eclipsed by the direct-driven steam turbine, at first, only for the largest and fastest vessels; later, by the aid of gearing, for vessels of moderate size and speed. It was natural that in Britain, the home of steam, the steam turbine should be the favoured engine. In Europe, notably

in Germany, Switzerland, and Denmark, the heavy-oil engine was preferred.

In the period from 1850, the shipbuilding industry in Britain grew so rapidly that for a time it produced more than 80 percent of the world output of ships, and even from 1910 to 1914, 60 percent. During the next half century other countries increasingly moved into the world's shipbuilding industry, notably Japan and the U.S.S.R.

Since World War II, scientific research has resulted in a general improvement, both in the propulsive performance of ships and in their structural qualities. The widespread adoption of welding, the introduction of prefabrication, as well as technical planning and improvements in managerial efficiency, have combined to revolutionize the industry.

Trend to larger ships. Since the early 1960s, the dimensions and complexity of ships have increased remarkably. Ore carriers, oil tankers, and other types have also grown in size, while specialized ships have replaced the general bulk-cargo vessels. Probably the most significant innovation has been the container ship, which carries prepacked cargo in standard containers eight feet square (2.4 metres) in cross section and 20 feet (6.1 metres) or 40 feet (12.2 metres) in length. The governing principle has been that the cost of transportation decreases as ship size increases. The unit cost of carrying oil in a 200,000-ton tanker, for example, can be 25 percent lower than in a 16,000-ton vessel. Moreover, a larger ship requires a proportionally smaller crew than does a smaller ship, thus saving on the wages bill. Modern ships not only are larger but are capable of attaining higher average service speeds, which permit these ships to spend more days out at sea and fewer days in port.

Container ships

SHIP DESIGN

Principles of naval architecture. Naval architecture is the branch of engineering concerned with the design and construction of ships. Merchant vessels range in size from mammoth tankers to tugs and trawlers; naval ships range from aircraft carriers to patrol craft. The basic principles of naval architecture apply to all vessels, irrespective of size and type: a ship design must guarantee safety, stability, strength, and speed completely adequate for the requirements and conditions of the vessel's operations.

The measurements of ships are given in terms of length, breadth, and depth. The length, L, between perpendiculars, is the distance on the summer (maximum) load waterline, from the forward side of the stem at the extreme-forward part of the vessel to the after side of the rudder post at the extreme rear, or to the centre of the rudder stock, if there is no rudder post. The beam, B, is the greatest breadth of the ship. The depth, D, is measured at the middle of the length, from the top of the keel to the top of the deck beam at the side of the uppermost continuous deck. Draft is the summer draft, measured from the keel. These terms, together with several others of importance in ship design, are illustrated in Figure 2.

Form coefficients

Certain form coefficients are useful as indications of a vessel's shape characteristics. The commonest is the block coefficient, which is the ratio of the actual displacement of the vessel to the displacement of a rectangular block with the same length, breadth, and draft as the vessel. The block coefficient thus indicates the relative fineness or slenderness of the vessel and varies from about 0.5 in a fine, fast ship to 0.85 in a full, slow-speed oil tanker. The midship area coefficient is the ratio of the immersed area of a vessel's midship section (the transverse section at the middle of the length) at any draft to the area of a rectangle of which the breadth is equal to the breadth of the vessel and the depth is equal to the given draft.

The prismatic coefficient is the ratio of the actual displacement of a vessel to the displacement of a prism of length equal to the vessel's length and of the same cross section as the vessel's immersed midship area. The water plane of a ship is the horizontal section of the ship at the waterline. The water plane area coefficient is the ratio of the actual area of a vessel's water plane to the area of a rectangle having the length of the ship and a breadth equal to the widest part of the water plane.

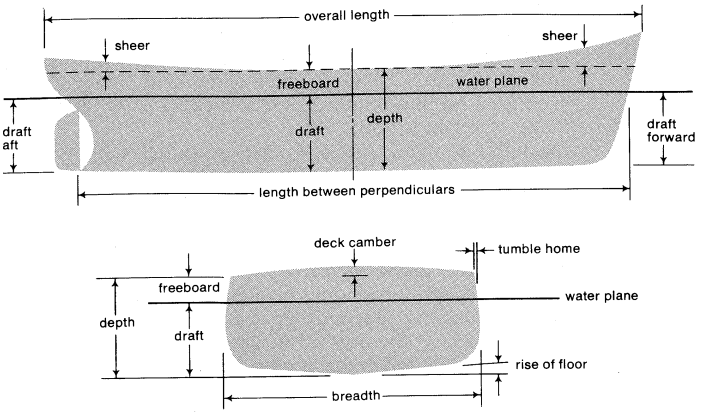


Figure 2: Terms used in ship construction.

Hydrodynamics and hydrostatics. When a body is immersed in a liquid, the upthrust of the water that holds up the body equals the weight of the fluid displaced by the body; this is Archimedes' principle. This upthrust is called buoyancy. Hence, when a body is floating freely in a fluid, the weight of the body equals the buoyancy, which equals the weight of the fluid displaced. The watertight volume of a ship above the waterline is called the reserve buoyancy. This is important in the assignment of the minimum service freeboard, freeboard being the distance between the waterline and the deck below which all bulkheads are made watertight. Minimum service freeboard determines the maximum permissible draft, the vessel's ability to survive flooding after damage, and its weather qualities in a seaway. The centre of gravity of the volume of water displaced by a ship floating at rest is called the centre of buoyancy. Its position is largely determined by the underwater shape of the vessel and is specified by its distance above the keel and from amidships. For equilibrium, the centre of gravity of the ship and its contents must lie in the same vertical plane as the centre of buoyancy. If a weight already on board is moved aft, for example, the ship's centre of gravity will move aft, but for equilibrium the vessel will adjust its position until the new centre of buoyancy is in the same vertical plane as the new centre of gravity.

Centre of buoyancy

Stability and trim. Figure 3 shows a transverse section of a ship floating at a waterline W_1L_1, displaced from its original waterline WL. One condition of equilibrium has

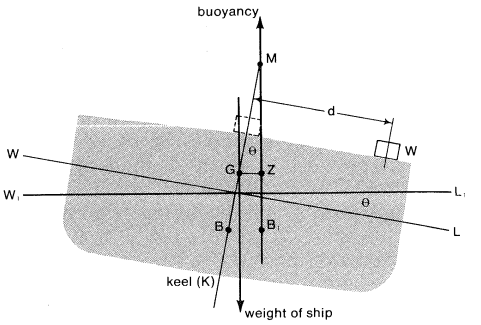


Figure 3: Stability shown in a transverse section of a floating ship (see text).

been defined above. A second condition is that the centre of gravity of the ship must be in such a position that, if the vessel is inclined, the forces of weight and buoyancy tend to restore the vessel to its former position of rest. At small angles, vertical lines through B_1, the centre of buoyancy when the vessel is inclined at an angle θ, intersect the centre line at M, the metacentre, which means "change point." If M is above G (the centre of gravity of the ship and its contents), the vessel is in stable equilibrium. When M coincides with G, there is neutral equilibrium. When M is below G, the forces of weight and buoyancy tend to increase the angle of inclination, and the equilibrium is unstable.

The distance GM is termed the metacentric height and

the distance GZ, measured from G perpendicular to the vertical through B_1, is termed the righting level or GZ value. Weight and buoyancy are equal and act through G and B_1, respectively, to produce a moment (tendency to produce a heeling motion) ΔGZ, where Δ is the displacement or weight in tons.

Stability at small angles, known as initial stability, depends upon the metacentric height GM. At large angles, the value of GZ affords a direct measure of stability, and it is common practice to prepare cross-curves of stability from which a curve of GZ can be obtained for any particular draft and displacement.

Transverse stability should be adequate to cover possible losses in stability that may arise from flooding, partially filled tanks, and the upward thrust of the ground or from the keelblocks when the vessel touches the bottom on being dry-docked.

The case of longitudinal stability, or trim, is illustrated in Figure 4. There is a direct analogy with the case of transverse stability. When a weight originally on board at

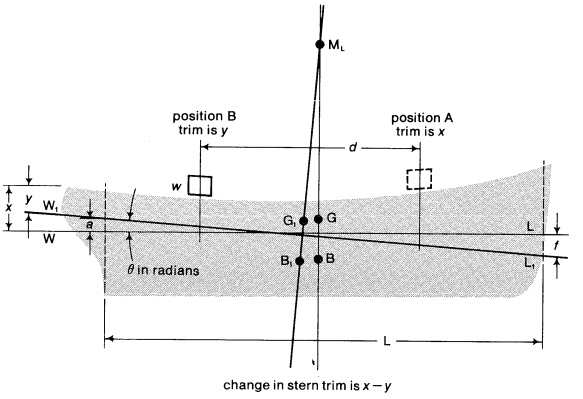


Figure 4: Longitudinal section of floating ship showing change in stern trim as deck load W is shifted from position A to position B (see text).

position A is moved a distance *d*, to position B, the new waterline W_1L_1 intersects the original waterline WL at the centre of flotation (the centre of gravity of the water plane area WL), the new centre of buoyancy is B_1 and the new centre of gravity is G_1.

For a small angle of trim, signified by the Greek letter theta (θ),

$$\theta = (a+f)/L$$
$$wd = \Delta GM_L\,(a+f)/L$$

Thus if $(a+f) = 1$ inch $= 1/12$ foot, $wd = \Delta GM_L/12L$ and this represents the moment to change trim one inch.

The inclining experiment. A simple test called the inclining experiment provides a direct method of determining GM, the metacentric height, in any particular condition of loading, from which the designer can deduce the position of G, the ship's centre of gravity. If a weight *w* (tons) is transferred a distance *d* (feet) from one side of the ship to the other and thereby causes an angle of heel θ (theta) degrees, measured by means of a pendulum or otherwise, then $GM = wd/\Delta\tan\theta$ (see Figure 3).

For any particular condition, KB and BM can be calculated, GM is found by the inclining experiment, whence $KG = KM - GM$. It is simple to calculate the position of G for any other condition of loading.

Ship motions. Ship motions are defined by the movements from the equilibrium position of the ship's centre of gravity along the three axes shown in Figure 5 and by rotations about axes approximately parallel to these. The linear displacements along the horizontal (*x*), lateral (*y*), and vertical (*z*) axes are termed surge, sway, and heave, respectively. The rotations about the corresponding body axes are respectively termed roll, pitch, and yaw (veering off course). Roll, pitch, and heave are oscillatory because hydrodynamic forces and moments oppose them.

Ship motions are important for many reasons. A ship should be able to survive any sea that may be encountered

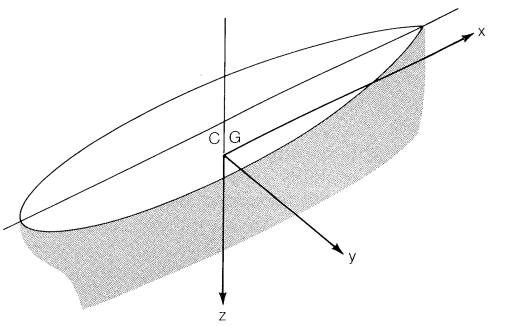


Figure 5: Coordinate axes of ship motions (see text).

and, in addition, to behave well and to respond to control. In brief, a ship should respond to the action of the sea in such a manner that the amplitudes of its motions and its position never become dangerous, and so that the accelerations it undergoes are kept within reasonable limits. Propulsive performance always deteriorates when a vessel is rolling, pitching, or heaving. Hence these motions are made as small as possible.

Rolling, pitching, heaving

Ship motions are excited by waves, whose growth is governed by the wind velocity at the sea surface, the area of water, or distance, over which the wind blows (the "fetch"), and the length of time during which the wind has been blowing (the "duration"). Any seaway is always a complex mixture of waves of different lengths, as wind itself is a complex mixture of gusts. All wave components do not travel in the same direction, but the directions of most of them in a single storm lie within 30° of each other. Regular trains of waves of uniform height and length are rarely, if ever, encountered. Most seas are confused and can be considered as made up of many separate component waves that differ in height and length. Pitching, rolling, and heaving are all excited by the changing pattern of surface waves in relation to the speed and course of the ship. In practice, it is possible to damp one motion only—that of rolling. The fitting of bilge keels (finlike longitudinal projections along the part of the underwater body of a ship between the flat of the bottom and the vertical topsides) has this effect, and still more effective means are the activated fin stabilizer (a device along the side of a ship activated by a gyroscope and used to keep the ship steady) and the passive or flume stabilizing tank, filled with water inside the ship.

Ship vibration. Vibration is undesirable in any ship, and especially in passenger vessels. The main causes of vibration are: (1) unbalanced forces and couples (or moments) in main machinery, auxiliary machinery, and propellers; (2) propeller-excited forces and moments arising in the wake of the ship; (3) impacts from sea waves. Vibration may be local, occurring in structural parts or members only; or it may be general, arising when the hull vibrates like a girder. Though the local type is almost impossible to predict when the ship is being designed, it can usually be cured at little expense, often by the addition of local stiffening or pillar support. Though general vibration can be avoided in many instances by careful design, alleviation presents formidable difficulties in a completed vessel.

Location of vibration sources

The hull of a ship is a girder that varies in cross section sometimes with large openings in the decks, and is subjected to nonuniform loads arising from the distributions of weight and buoyancy. Additionally, the hull is subjected to acceleration forces arising from the ship's motions in a seaway and from wave impacts. Forces and moments from the main and auxiliary machinery that cause vibration can be eliminated almost entirely, but forces and moments from the propellers and from waves are inevitable. Furthermore, this circumstance holds for each condition of loading—intermediate between-load draft, when the ship is fully loaded with cargo, and ballast draft, when the ship is empty of cargo, but the ballast tanks (which are filled with water to give the ship stability when it is empty of cargo) are full. Though it is almost impossible to calculate all of the possible modes and frequencies of vibrations to which a ship will be subjected, it is possible to compute

with some reasonable degree of accuracy the fundamental modes of vertical and horizontal vibrations. Though the accuracy of calculated results for higher modes decreases as the order of vibration increases, reasonable approximations can be obtained. The results of such calculations are sometimes presented to indicate the ranges of engine speed, in revolutions per minute, at which excessive vibration may occur.

Resistance and propulsion. The motion of a surface vessel, whether towed or propelled, is characterized by three effects. They are: resistance to motion, formation of a wave pattern, and change of both draft and trim, as compared with draft and trim at rest.

When a vessel is in motion, friction exists between the hull surface and the water touching it; the vessel's motion also causes pressure between the hull surface and the adjacent fluid. These two forces cause waves. The relation of these forces to various ship designs is often studied on small models in experimental tanks.

Any propelling device operates by accelerating, and so increasing the momentum of, the fluid that passes through it. For the great majority of surface vessels, the screw propeller provides an efficient and convenient method of propulsion. The propeller may be described as a device for obtaining a propulsive thrust from a rotary motion.

Ships are usually propelled by one or more propellers situated at the stern. Thus the propeller works in a region where water flow is disturbed by the hull—*i.e.,* in the "wake," where the water has already been set in motion by the passage of the ship.

The action of the propeller increases flow velocities over the hull near the stern and thus reduces fluid pressure in this region. Consequently, there is in effect an addition to the resistance of the hull, commonly termed resistance augment or thrust deduction.

Efficiency of propulsion is generally defined at any speed as the ratio between the power required to tow the vessel and the power absorbed by the propeller(s). The propulsive qualities of a vessel are ascertained by means of: (1) resistance tests with the naked hull; (2) resistance tests with the hull, plus any appendages that affect propulsion (*e.g.,* the struts that support the propeller); (3) self-propulsion tests. The optimum hull has the minimum increase of resistance in waves; hence further tests may be made in waves.

Propellers for high-speed vessels present special difficulties, notably cavitation, which may be described as the formation and collapse of cavities in a stream of flowing liquid as a result of pressure changes within the stream caused by changes in the velocity of flow. Cavitation tunnels are employed for the appropriate propeller investigations. Cavitation is overcome by increasing the propeller blade area and by modification of the blade section shape.

Manoeuvrability. Increases in the size and speed of ships bring problems of safe operation in congested waters and control at high speed in waves. Therefore, designs necessarily represent a compromise between manoeuvrability and course-keeping ability. Ship operators desire maximum manoeuvrability in port to minimize the need for assistance from tugs and to reduce delays in docking. They also desire a ship that can hold a steady course at sea with the minimum use of helm. These aims, however, are mutually conflicting.

Rudders and steering

A ship is steered by means of one or more rudders arranged at the stern or, in rare cases, at the bow. There are many types and shapes of rudders, depending upon the type of ship, design of stern, and number of propellers. When a yaw—that is, a change of angle about a vertical axis through the centre of gravity—is started, a turning moment is set up and the ship swings off course unless the swing is corrected by rudder action. This turning effect arises because the hull's centre of lateral resistance is much nearer the bow than the ship's centre of gravity.

Good course keeping demands directional stability. This is aided by design features that bring the centre of lateral resistance nearer to the ship's centre of gravity. These measures, however, increase the diameter of the ship's turning circle, requiring a design compromise. In warships, in vessels operating in confined waters, and in tugs, a small turning circle is essential. In merchant ships, rapid manoeuvring is required only in port; accordingly, the everyday function of the rudder is to ensure the maintenance of a steady course with the minimum use of helm. In this sense, turning circle properties are of less practical significance than the effect of small rudder angles.

Structural strength. In still water, the loads acting on a hull include weight, fluid pressure, and buoyancy. In a seaway, additional forces arise from the wave-induced alteration of buoyancy distribution, from the ship's motions, from vibration, and from static and dynamic fluid pressures, including wave impact and slamming. The resultant loading on the main hull girder can be resolved into distributions along the length of shearing force, the bending moment in both vertical and horizontal planes, and the twisting moment.

Stresses acting upon the ship

The maximum stresses usually arise from the longitudinal vertical bending moment when the vessel is moving at right angles to a train of waves. As indicated above, this bending moment is considered as the sum of two components: first, the still water bending moment, due to the unequal distribution of weight and buoyancy along the ship length; and, second, the wave bending moment, which arises from differences in buoyancy distribution between the still-water and wave surfaces. The first of these components is governed by the disposition of weights within the ship and can often be minimized by skill and ingenuity at the design stage. The second is governed by the shape of the hull and the length and height of waves encountered. The length of a wave is the horizontal distance from the lowest point of one wave (the trough) to the lowest point of the next, or from the highest point of one wave (crest) to the highest point of the next. The height of a wave is measured vertically from the trough to the crest of any single wave. It is customary to assume that the worst conditions occur when the length of the wave equals the ship length or when the midpoint of the ship is either at the crest or in the trough. Height of wave was formerly taken to be $1/20$ of its length, but a common present-day practice is to assume that height h is the square root of the length l times the factor 1.1 ($h = 1.1\sqrt{l}$).

With the wave crests at the ends of the ship, the supporting buoyancy is mainly at the ends of the ship, while with the wave crest at amidships the support is mainly at amidships. In the former case the greatest bending moment acting on the structure will occur when there is a concentration of weight amidships. Similarly, with a wave crest at amidships, the greatest bending moment will occur when there is concentration of weight toward the ends. The former case is known as sagging and the latter case as hogging.

In addition to stresses arising from the vertical forces which tend to bend the hull and which have their origin in heavy seaways and in engine pulsations, shear stresses are often investigated, notably in the longitudinal bulkheads of oil tankers. In container ships, where there is little actual deck area because of the large hatchways, special structural reinforcement is required to obtain the necessary strength. High superstructures, extending over a large fraction of the length, are a characteristic feature of passenger liners; and the design of such superstructures presents special problems. Transverse strength also poses difficult problems, notably in oil tankers, container ships, and bulk carriers.

The design of all structures is based upon experience, and experience is nowhere more requisite than in the design of ships. The great majority of ships are built according to the requirements of the classification societies of their countries, which publish rules and regulations for the construction of ships. To build a vessel of the classification specified by the owner, the shipbuilder must conform to these rules and regulations; he is thereby spared much difficult and tedious calculation work.

Steel is the main structural material in shipbuilding. The permitted types, composition, and qualities of mild and high-tensile steel, together with the approved manufacturing and the requisite tests, are specified in the classification society rules. The rules also cover the composition, manufacture, and tests appropriate to castings, forgings, and other manufactured components. Aluminum alloys

Materials in shipbuilding

that are strong and highly resistant to corrosion are widely employed in the construction of above-deck components, particularly in the superstructures of passenger vessels, in which their relative lightness contributes to the stability of the ship. High-tensile steels have been used increasingly in the construction of very large tankers, bulk carriers, and container ships. Wood and wood substitutes are often employed for the sheathing of weather decks. Various synthetic materials that are highly resistant to fire can be used in accommodation spaces.

Spacing of watertight bulkheads. Typical ship hazards are stranding—when the bottom may be holed—and collision, when spaces within the hull may become flooded. Dry-cargo ships are protected from rupture upon stranding by their double bottom construction. All ships must have a collision bulkhead, bulkheads to enclose the machinery space, and others—the number of which depends, in cargo and passenger ships, on their length. Oil tankers are well subdivided by both transverse and longitudinal bulkheads because a large number of separate tanks is required. In all vessels, but especially in passenger ships, the number and spacing of transverse watertight bulkheads must be adequate for safety under specified conditions of flooding. The extent to which a compartment open to the sea will flood depends on the permeability of its contents, this being defined as the ratio of the volume that can be occupied by water to the total volume. Formulas for the permeability of different compartments are laid down by governments, according to international agreement. Calculations to determine the floodable length—that is, the length at any point in the ship that may become flooded, without resulting in the immersion of any part of the margin line, which is three inches (76 millimetres) below the deck to which the bulkheads extend—are made on the basis of an assumed permeability. The rule book also specifies formulas which ensure that one, two, or three adjacent compartments may be flooded before the margin line is immersed. Ships are thus described as having one-, two-, or three-compartment standards. The product of the floodable length and the factor of subdivision gives the permissible length of compartment. In the case of passenger ships, calculations must be made to discover whether their stability remains adequate under flooding.

Safety. In most maritime countries, a government department assumes responsibility for specifying and enforcing standards of safety in merchant vessels relating to maximum allowable draft, lifesaving appliances, stability, bulkhead subdivision, carriage of dangerous cargoes, lights and signals, and so on. These departments employ surveyors who supervise the testing, fitting, and inspection of all lifesaving appliances and other equipment essential for the safety of passengers and crew. Official inquiries are made into accidents, strandings, collisions, or wrecks, when loss of life is incurred.

Classification societies

The classification societies, such as Lloyd's Register of Shipping in England, the American Bureau of Shipping, Det Norske Veritas, Bureau Veritas, Registro Italiano Navale, Germanischer Lloyd, and Nippon Kaiji Kyokai, undertake the task of determining appropriate technical standards in ship construction and classification. Standards are laid down in the rules and regulations published by these societies, together with detailed specifications for the composition, quality, and testing of all materials of construction. The societies' surveyors are responsible for the testing of materials and for supervising the construction and repair of ships and their machinery. An owner specifies that a ship built to his order obtain a certain classification-society class because, for insurance purposes, this ensures a definite standard of strength and workmanship and a minimum freeboard. In the U.K., Lloyd's Register of Shipping is empowered by the Board of Trade to assign freeboards. Lloyd's also publishes annually its register book, which gives technical information for all merchant ships in service throughout the world and is kept up-to-date by the regular issue of supplements.

The naval architect. A naval architect asked to design a ship may receive his instructions in a form ranging from such simple requirements as "an oil tanker to carry 100,000 tons deadweight at 15 knots" to a fully detailed specification of precisely planned requirements. He is usually required to prepare a design for a vessel that must carry a certain weight of cargo (or number of passengers) at a specified speed with particular reference to trade requirements; high-density cargoes, such as machinery, require little hold capacity, while the reverse is true for low-density cargoes, such as grain.

Deadweight is defined as weight of cargo plus fuel and consumable stores, and lightweight as the weight of the hull, including machinery and equipment. The designer must choose dimensions such that the displacement of the vessel is equal to the sum of the deadweight and the lightweight tonnages. The fineness of the hull must be appropriate to the speed. The draft—which is governed by freeboard rules—enables the depth to be determined to a first approximation.

Major design factors

After selecting tentative values of length, breadth, depth, draft, and displacement, the designer must achieve a weight balance. He must also select a moment balance because centres of gravity in both longitudinal and vertical directions must provide satisfactory trim and stability. Additionally, he must estimate the shaft horsepower required for the specified speed; this determines the weight of machinery. The strength of the hull must be adequate for the service intended; detailed scantlings (frame dimensions and plate thicknesses) can be obtained from the rules of the classification society. These scantlings determine the requisite weight of hull steel.

The vessel should possess satisfactory steering characteristics, freedom from troublesome vibration, and should comply with the many varied requirements of international regulations. Possessing an attractive appearance, the ship should have the minimum net register tonnage, the factor on which harbour and other dues are based. (The gross tonnage represents the volume of all closed-in spaces above the inner bottom. The net tonnage is the gross tonnage minus certain deductible spaces that do not produce revenue. Net tonnage can therefore be regarded as a measure of the earning capacity of the ship, hence its use as a basis for harbour and docking charges.) Passenger vessels must satisfy a standard of bulkhead subdivision that will ensure adequate stability under specified conditions if the hull is pierced accidentally, as through collision.

Compromise plays a considerable part in producing a satisfactory design. A naval architect must be a master of approximations. If the required design closely resembles that of a ship already built for which full information is available, the designer can calculate the effects of differences between this ship and the projected ship. If, however, this information is not available, he must first produce coefficients based upon experience and, after refining them, check the results by calculation.

Training. There are four major requirements for a good naval architect. The first is a clear understanding of the fundamental principles of applied science, particularly those aspects of science that have direct application to ships—mathematics, physics, mechanics, fluid mechanics, materials, structural strength, stability, resistance, and propulsion. The second is a detailed knowledge of past and present practice in shipbuilding. The third is personal experience of accepted methods in the design, construction, and operation of ships; and the fourth, and perhaps most important, is an aptitude for tackling new technical problems and of devising practical solutions.

The professional training of naval architects differs widely in the various maritime countries. University degrees in naval architecture are offered by many universities and polytechnic schools; such academic training must be supplemented by practical experience in a shipyard.

Trends in design. The introduction of calculating machines and computers has facilitated the complex calculations required in naval architecture and has also introduced new concepts in design. There are many combinations of length, breadth, and draft that will give a required displacement. Electronic computers make it possible to prepare a series of designs for a vessel to operate in a particular service and to assess the economic returns to the shipowner for each separate design. Such a procedure is best carried out as a joint exercise by owner and

builder. As ships increase in size and cost, such combined technical and economic studies can be expected to become more common.

SHIPBUILDING

The shipyard. The wooden ship was constructed on a building berth, around which timbers and planking were cut and shaped, and then fitted together on the berth to form the hull. A similar practice was followed with iron vessels and, later, with the earlier steel ships, as these tended to be replicas of wooden hulls. Gradually iron came to be used more effectively in its own right, rather than as a substitute for timber. The berth or slipway from which the vessel is launched is an assembly area, rather than a ship construction site. In many shipyards the number of launching berths has been reduced to increase the ground area available for prefabrication sheds. Greater ease of fabrication means that, despite the reduction in the number of berths, more ships can be built and construction costs lowered.

The building berth

Organization. A shipbuilder undertakes to deliver to the client by a certain date and for a stated sum a vessel with specific dimensions, capabilities, and qualities, a vessel that has been tested on trial and is ready for service. The function of a shipyard is the production of completed ships in accordance with the shipbuilder's undertakings. The raw materials for construction and finished items to be installed on board are delivered there. The labour force in the yard consists of various workmen—steelworkers, welders, shipwrights, blacksmiths, joiners, plumbers, turners, engine fitters, electricians, riggers, and painters.

Management is headed by a chairman and a board of directors, consisting usually of about 6 to 12 members from the technical, commercial, and secretarial departments, with one or more representing outside interests. The chief departments are the design, drawing, and estimating offices, planning and production control, the shipyard department—responsible for construction up to launching—and the outside finishing department, which is responsible for all work on board after launching. Other departments are responsible for buying and storekeeping and the yard maintenance.

The construction of the hull is only one of a shipbuilder's responsibilities. As soon as a contract is placed, he must negotiate with subcontractors for the supply of items that shipyards do not produce—the electric power plant, propulsion machinery, shafting and propellers, engine-room auxiliaries, deck machinery, anchors, cables, and furniture and furnishings. Production planning and control is therefore a complex undertaking, covering subcontracts, assembly, and installation, in which costs must be kept as low as possible.

Layout. In general, a shipyard has few building berths and uses extensive areas around them for the construction of large components of the steel hull. Building berths slope downward toward the waterway, to facilitate launching. Building basins, or dry docks, are sometimes used for the construction of very large vessels, because it is convenient to lower, rather than to lift, large assemblies, and this method also eliminates problems associated with launching. Extensive water frontage for the building berths is unnecessary. The main requirement is a site of considerable depth, rather than width, with a large area extending inland from the berths. Steel plates and sections are delivered to the shipyard at the end of the area farthest from the berths. There they are stored in a stockyard and removed, as needed, for cleaning, straightening, shaping, and cutting. Separate streams of plates and rolled sections converge toward the prefabrication shop, where they are used to build structural components or subassemblies. The subassemblies are transported to an area nearer the berths, where they are welded together to form large prefabricated units, which are then carried by cranes to the berth, to be welded into position on the ship.

Prefabrication

In practice, there are many variations in this general procedure. At the Götaverken shipyard at Göteborg, Sweden, for example, partly fitted-out sections of the ship are fabricated in sheds and welded to sections already completed. As the vessel becomes built up with component sections, it is moved out of the covered area into a building dock and further sections are attached until the ship is complete. Water is then admitted to the dock and the ship is floated out for completion in the fitting-out basin. An example of an outstanding shipbuilding facility is the Harland and Wolff building dock in Belfast, Northern Ireland, completed in 1970, then the largest of its kind in the world: 1,825 feet (556 metres) long, 305 feet (93 metres) wide, and 38 feet (12 metres) deep. It is large enough for the construction of a 1,000,000-ton dead-weight tanker. The dock is spanned by a travelling crane capable of lifting prefabricated structures of 840 tons maximum weight.

Planning. Delivery of a completed ship by a specified date requires careful planning. Following the introduction in the United States of the critical path method of planning and control by the E.I. du Pont de Nemours and Company about 1959, new techniques were adopted in many shipyards.

Critical path method

The critical path method is the basis of network analysis, which is used in planning complex production projects. The network, and information derived from it, is used for overall planning of a project and also for detailed planning with production progress control. The network gives a logical, graphical representation of the project, showing the individual elements of work and their interrelation in the planned order of execution. Each element of work is represented by an arrow, the tail of which is the starting point of activity and its head the completion. The arrows are drawn to any suitable scale and may be straight or curved. An event, which represents the completion of one activity and the beginning of another, is usually indicated by a circle and described further by a number within the circle. But each activity need not be completed before the next activity is begun. The logical order of steelwork in a hull, for example, is: (1) detailed drawings of steelwork; (2) ordering of steel; (3) manufacture and delivery of steel; (4) storing of steel material in stockyard; (5) shotblasting, cleaning, and forming operations; (6) subassembly work; and (7) erection of structure on berth. These operations can be represented on a ladder type of diagram. Many such diagrams—ladder and other types—go toward making up the complete aggregate operation of building a ship. When the proper sequence of operations is decided upon, times must be allocated to each operation to ensure that the workers in charge understand their obligations. Planning, based on realistic estimates of times and costs, must begin at the precontract stage, so that, throughout the building program, a clear plan, with scheduled dates for each major section, is available. Detailed networks must be prepared for each of the major sections, showing dates for completion. The earliest and latest permissible starting and finishing times are indicated for each activity.

The critical path of a project is a series of activities whose duration cannot be increased without delaying the completion of the project as a whole. In large networks there may be more than one critical path. Up to about 100 activities can be dealt with manually but, for more complex cases, the numerical work is done by computer. The spare time available for a series of activities; *i.e.,* the maximum time these activities can be delayed without retarding the total project, is aggregated into a "total float." This is regarded as a factor of safety to cover breakdowns, mishaps, and labour troubles. Intelligent and experienced use of critical path methods can provide information of great value. Savings in production costs depend upon the use that management makes of this information.

General-arrangement drawing

From contract to working plans. Before an order is placed, the main technical qualities of the ship are decided upon and a general-arrangement drawing of the vessel, showing the disposition of cargo, fuel, and ballast, and crew and passenger accommodation is prepared. This plan provides a complete picture of the finished vessel. It is accompanied by detailed specifications of hull and machinery. This general-arrangement plan and the specifications form the basis of the contract between shipowner and shipbuilder.

As soon as an order is confirmed, drawing offices and planning departments produce working plans and instructions. Since ships are usually constructed according to the

rules of a classification society, the stipulated structural plans are normally submitted to the society for approval. The spacing of bulkheads in passenger ships, for example, must be approved by the appropriate authority. For all ships, passenger and cargo, the approval of the maximum permissible draft must be sought from the classification society. Necessary working drawings include the lines plan and detailed plans of the steel structure—shell plating, decks, erections, bulkheads, and framing—as well as accommodation spaces, plumbing, piping, and electrical installations, and main and auxiliary machinery layout. The planning and production department prepares a detailed progress schedule, fixing dates for the completion of various stages in the construction. A berth in the yard is allocated to the ship, arrangements for the requisite materials, labour, personnel, and machines are made, and precautions are taken to ensure that the many interrelated operations will progress according to the timetable.

The lines plan; fairing. A lines plan, usually a 1/48 life-size scale drawing of a ship, is used by designers to calculate required hydrostatic, stability, and capacity conditions. Full-scale drawings formerly were obtained from the lines plan by redrawing it full size and preparing a platform of boards called a "scrive board" showing the length and shape of all frames and beams. Wood templates were then prepared from the scrive board and steel plates marked off and cut to size.

An alternative to the full-scale scrive board is a photographic method of marking off, introduced about 1950 and widely adopted. The lines plan is drawn and faired (mathematically delineated to produce a smooth hull free from bumps or discontinuities) to a scale of one-tenth full-size by draftsmen using special equipment and magnifying spectacles. The formerly used wood templates are thus replaced by specially prepared drawings, generally on one-tenth scale. Photographic transparencies of these drawings are then projected full size from a point overhead onto the actual steel plate. The plate is then marked off to show the details of construction, such as position of stiffening members, brackets, and so on. This optical marking-off system is much more economical in terms of space and skilled labour than the older method.

Numerical method of fairing

By the 1960s, digital computers were being used to fair the preliminary lines plan by a numerical method. Faired surfaces can be produced to a specified degree of accuracy and the lines can be drawn by a numerically controlled drawing machine, bringing the process under continuous scrutiny. Tapes can be produced for use in numerically controlled plate-burning machines, which cut plates to shape, and for the automatic cold bending of frames and curved girders. Fairing calculations produce data that can be fed back into a computer, and programmed to generate hydrostatic and stability data and other information.

Fabrication and assembly. Since 1930, rivetting has been progressively supplanted by welding. This has proved more than a mere alteration in the method of connecting structural components because welding facilitates prefabrication of large component parts of the main hull structure.

Before welding came into wide-scale use every ship was constructed on the building berth. The keel was laid, floors laid in place, frames or ribs erected, beams hung from the frames, and this skeleton, framed structure was held together by long pieces of wood called ribbands. Plating was then added and all the parts of the structure were rivetted together. In other words, the ship was built from the keel upward.

The modern method is to construct large parts of the hull, for example, the complete bow and stern. Each of these parts is built up from subassemblies or component parts, which are then welded together to form the complete bow or stern. These sections of the ship are manufactured under cover in large sheds, generally at some distance from the building berth, before being transported to the berth and there fitted into place and welded to the adjacent section. The advantages of this procedure are that work can proceed under cover, unhampered by bad weather, and the units or component parts can be built up in sequences to suit the welding operations—not always possible at the building berth itself.

A number of techniques can be used to weld together two pieces of the same metallic material. The ideal weld is a continuity of homogeneous material, with the same composition and the same physical properties as the parts being joined. In steel shipbuilding, metal arc welding is produced by an electric arc formed between the parts to be joined; the fusion material is supplied by a coated electrode. The welding electrode consists of a core rod that is deposited as weld metal; it is flux coated to protect the molten metal from the atmosphere during deposition and to supply certain metallurgical properties to the weld. A great deal of research has gone into the production of the best possible coated electrodes for specific duties. The main advantages of welding over rivetting are: (1) a lighter structure (because overlaps are eliminated), (2) improved watertightness and oiltightness, (3) smoother surfaces, and (4) reduced hull upkeep. Certain precautions, however, are necessary. The design of the structure must be adapted for welding because structural details which can be rivetted are seldom suitable for welding. The joints must be carefully prepared beforehand for welding. Incomplete penetration, lack of fusion, porosity, and cracking are typical weld defects that must be avoided. Hard spots must be avoided and gradual tapering off of stiffness is necessary if defects in service are to be minimized.

The welding electrode

Launching, outfitting, and trials. *Launching.* Apart from certain small craft built on inland waterways, which are launched sideways, the great majority of ships are launched stern first from the building berth. Standing structures called ways, constructed of concrete and wooden blocks, spaced about one-third of the vessel's beam apart, support the ship under construction. The slope of the standing ways—which are often cambered (slightly curved upward toward the middle or slightly curved downward toward the ends) in the fore and aft direction—ranges from one-half to three-quarters of an inch per foot of length (from 42 to 62 millimetres per metre of length); ways extend from a position near the bow to past the stern and for a certain distance into the water. Over these standing ways is built the launching cradle, which consists of sliding ways on which are built poppets, or supporting structures, of timber to provide support for the hull. Between standing ways and launching ways is a layer of lubricant.

During construction the ship is supported by at least one line of blocks under the keel, with side supports and shores as necessary. As the vessel nears completion, the standing ways are built under it, the sliding ways are superimposed, and the cradle is built up. The weight of the vessel is transferred gradually to the standing ways. The full weight must not be supported by the ways for too long because the thickness of lubricant would be reduced by squeezing and its properties would be adversely affected. It is common to fit launching triggers which, when released at the moment of launching, permit the sliding ways to move over the standing ways.

As a vessel moves down the ways, the forces operating are: its weight acting down through the centre of gravity, the upward support from the standing ways, and the buoyancy of the water. As it travels further, the buoyancy increases and the upthrust of the ways decreases, with the weight remaining constant. As the centre of gravity passes the after end of the standing ways, the moment of the weight about the end of the ways tends to tip the ship stern first. At this position and for some time later, it is essential that the moment of buoyancy be greater than the moment of weight about the after end of the ways, thus giving a moment to keep the forward end of the sliding ways on the standing ways; otherwise there would be concentration of weight at the end of the ways, causing excessive local pressure. Calculations are made to determine the most important factors in launching, namely, the moment at which the stern lifts, the difference between weight and buoyancy when the stern lifts, the existence of a moment against tipping, and the equality of weight and buoyancy before the vessel reaches the after end of the ways to ensure that the cradle will not drop off the end of the standing ways.

The launching of a vessel into a restricted waterway requires the application of a retarding force. Usually piles

Launching into a restricted waterway

of chains are laid alongside the sides of the ship to act as drags, and these are secured to chain plates by wire cables, fixed temporarily to the hull. As the vessel slides down the launching ways, the drags come serially into operation after, or sometimes before, the bow has cleared the after end of the ways. Launching can be a hazardous operation. If the lubricant is ineffective, the vessel will not move. If the stern does not lift as the vessel slides down the ways, the ship may tip about the way ends. The bow may sustain damage when it drops into the water at the end of the ways and may damage the slipway when the stern lifts. Excessive loads on the poppets may cause their collapse.

Outfitting. After launching, the ship is berthed in a fitting-out basin for completion. The main machinery, together with auxiliaries, piping systems, deck gear, lifeboats, accommodation equipment, plumbing systems, and rigging are installed on board, along with whatever insulation and deck coverings are necessary. Fitting out may be a relatively minor undertaking, as with a tanker or a bulk carrier, but in the case of a passenger vessel, the work will be extensive. Although fitting-out operations are diverse and complex, as with hull construction there are four main divisions: (1) collection and grouping of the specified components, (2) installation of components according to schedule, (3) connection of components to appropriate piping and/or wiring systems, and (4) testing of completed systems.

The tendency in planning has been to divide the ship into sections, listing the quantities of components required and times of delivery. Drawings necessary for each section are prepared and these specify the quantities of components required. A master schedule is compiled, specifying the sequences and target dates for completion and testing of each component system. This schedule is used to marshal and synchronize fitting work in the different sections and compartments.

Trials. As the vessel nears completion a number of tests are made. The naval architect makes a careful assessment of the weight of the finished ship and checks its stability and loading particulars by reference to data for the ship's lightweight and centre of gravity, obtained from a simple inclining experiment. The inclining test also provides a check on calculations.

Before the official sea trials, dockside trials are held for the preliminary testing of main and auxiliary machinery. Formal speed trials, necessary to fulfill contract terms, are often preceded by a builder's trial. Contract terms usually require the speed to be achieved under specified conditions of draft and deadweight, a requirement met by runs made over a measured course.

Speed trials

It is usual to conduct a series of progressive speed trials, when the vessel's performance over a range of speeds is measured. The essential requirements for a satisfactory measured course are: adequate depth of water; freedom from sea traffic; sheltered, rather than exposed, waters; and clear marking posts to show the distance. Whenever possible, good weather conditions are sought. With a hull recently docked, cleaned, and painted, sea-trial performance can provide a valuable yardstick for assessing performance in service. Ideally, the ship should be run on trial in the fully loaded condition; but this is difficult to achieve with most dry-cargo ships. It is, however, comparatively simple to arrange in oil tankers, by filling the cargo tanks with seawater. Large vessels with a low displacement–power ratio must cover a considerable distance before steady speed is attained; hence they need to make a long run before entering upon the measured distance.

Modern trends. The rapid expansion of world trade, combined with the closure of the Suez Canal from 1967 to 1975, created a great demand for shipping of all types, but particularly for large vessels to carry bulk raw materials, such as oil, ore, and grain. Ship passenger transport has declined dramatically in favour of air transport, but passenger vessels continue to find employment in the tourist and cruise fields.

The effect of these changes on the shipbuilding industry has been profound. Shipyards have had to be reorganized and, in many cases, rebuilt to cope with the demand for very large ships. This has necessitated vast capital expenditure, and in a period of fierce world competition there have been amalgamations of small firms into large shipbuilding groups.

PROPULSION

Marine engineering is a branch of mechanical engineering. Its distinctiveness lies in the special features of engine design and operation that are necessary for ship propulsion. Otherwise, production processes, organization, and other matters have a similar basis.

A shipowner considers the following factors before ordering a ship: (1) capacity and deadweight on suitable draft, (2) speed, (3) first cost, (4) running and maintenance charges, and (5) general suitability. Items (1) and (2) indicate the purchaser's requirements. Items (3), (4), and (5) are computed by the engine builder.

Factors affecting choice of engine

The choice of propelling machinery is influenced by the engine power necessary for the required ship speed; the effects of the weight and space of engines plus fuel bunkers on the dimensions of the ship; fuel costs; upkeep costs—including overhaul, replacements, and consumable stores; and special considerations, such as personnel required. The choice of alternatives in marine propulsion may be grouped into two basic classes: diesel engines and geared steam turbines. The diesel engine is available in double-acting two-stroke; single-acting two-stroke or single-acting four-stroke direct drive; and in indirect drive systems, such as geared diesel engines and diesel-electric systems. It is usually possible to determine, without elaborate analysis, which type of machinery is most suitable for a particular vessel engaged in a particular trade. Occasionally, however, detailed investigation is necessary.

Steam installations. *Steam generators.* Immediately after World War II general mercantile practice favoured boiler pressures of 450–600 pounds per square inch (psi) with steam temperatures of 750–825° F (400–440° C). Occasionally, installations were built with levels as high as 850 psi and 900° F (480° C). In the early 1970s, pressures of 600–800 psi and temperatures of 800–900° F (425–480° C) became general; but higher ranges (*e.g.*, 900–1200 psi and 1,000° F [540° C]) were accepted.

The following design data apply to a typical steam generator: normal evaporation, 180,000 pounds (82,000 kilograms) of water per hour; steam pressure/temperature, 880 psi, 990° F (530° C); feed temperature, economizer inlet, 280° F (140° C); temperature air heater outlet, 250° F (120° C); carbon dioxide in funnel gases, 15 percent; boiler efficiency, 86.8 percent. Such a generator normally would be provided with four fuel injectors, directed downward, arranged in the combustion chamber roof. Combustion air from the atmosphere is delivered by motor-driven fans. The furnace is surrounded by a wall of water-filled tubes, to which the heat of combustion is transmitted. The steam thus generated rises to a horizontal steam drum and from there it flows through pipes to the superheater, to be further heated by combustion gases. Finally, it is conveyed to the turbine. Hot gases rising from the superheater flow to the economizer, through which feedwater is circulated on its way from the feed heating system to the boiler. The products of combustion finally rise to the funnel and thence into the atmosphere.

Steam turbines. A succession of evolutionary changes has brought steam-turbine practice to its present level. Two-turbine units are now standard for all except low-power installations, and these are single-turbine sets.

Gearing. Reduction gearing between the steam turbines and the propeller enables the prime mover to rotate at speeds sufficiently high to ensure maximum thermal efficiency, while propeller revolution is slow enough to realize the maximum propulsive efficiency. In this respect the geared turbine has an advantage, compared with the direct-driven diesel engine.

Double- and single-reduction gearing

Double-reduction gearing is the most widely applied arrangement, the ratio of reduction ranging between 30 to 1 and 60 to 1; the lower values are for low-pressure turbines and the higher values for high-pressure turbines. Single-reduction gearing, in which the ratio may be from 12 to 1 up to 20 to 1, has only limited applications.

For twin-screw ships, the propeller revolutions are usually

of the order of 110 to 120 per minute (rpm). For single-screw vessels, the propeller may revolve at 90 to 100 rpm. Where the ship form permits, propellers of large diameter may be installed, as in single-screw ships with a large cross section (*e.g.*, tankers); in such ships the revolution may be as slow as 80 per minute, to the great advantage of propulsive efficiency.

Internal-combustion engines. *Diesel engines.* Heavy-oil engines, commonly termed diesel engines, operate on a cycle comprising four phases: suction, compression, expansion, and exhaust. Fuel is injected at the end of the compression stroke. In the so-called four-stroke cycle, each of the phases occupies one stroke; that is, the complete cycle extends over two revolutions. The earliest marine engines were of this form and were heavy and costly. The next step was to complete the four-phase cycle in one revolution, hence the term two-stroke cycle. In this cycle, the revolution is divided into three approximately equal periods, for compression, combustion and expansion, and exhaust and scavenge. In all the early engines, combustion air was drawn from the atmosphere and the engines were termed atmospheric induction engines. Air density determines the maximum fuel weight that can be burned; it therefore determines the power that can be developed. If the charge-air density is increased by the interposition of a compressor between the atmosphere and the cylinder, then the weight of air is increased and a greater weight of fuel can be burned, with a proportionate increase in power output. The compressor, or blower, is driven by a gas turbine utilizing the heat energy of the exhaust gases. This is the principle of exhaust turbocharging. Single-acting, two-stroke, exhaust turbo-pressure-charged engines are almost standard in marine practice.

Geared engines. When engine powers were moderate, there was little or no need for geared engines. With the trend to larger direct-coupled engines, the scope for geared units, as an alternative method of propulsion, has widened. Moreover, there is greater opportunity for variation with geared engines than with direct-coupled engines. In general terms, the advantages of geared drive are: (1) greater reliability, by having more than one engine per screw; (2) when the ship is running light or partly loaded, one or more of the engines can be shut down and the others operated at their most efficient ratings; (3) maintenance is easier because the engines are of more manageable size; (4) engines can be overhauled one after the other at sea; (5) by modifying the number of engines per ship and cylinders per engine, the propelling machinery of a fleet of vessels of different sizes and powers can be standardized upon a single cylinder size, with advantage to initial cost, delivery time, and replacements.

Flexible couplings

Geared engines occasionally have been made with a heavy flywheel as the only damping agent between engine and gearing. More commonly, however, a flexible coupling of some kind is interposed between engine and gearing: to serve as a cushion, preventing the transmission of detrimental torque pulsations from engine to gearing; to function as a quick-disconnecting clutch for use when manoeuvring; and to limit the transmissible torque to, say, 1.5 to 2.5 times its normal value and thus obtain a measure of protection should one of the engines stop suddenly.

A flexible coupling can be mechanical, electrical, hydraulic, or pneumatic. A mechanical coupling cannot easily be made to serve as a clutch. Cost, weight, space, and reliability are the factors to be considered when comparing forms of coupling. The gearing is always single-reduction. The efficiency of the energy transmission, from engine to propeller, is approximately 92 percent for a geared installation and about 97 percent for a direct-coupled engine.

Diesel-electric drive. An alternative to geared engines is the diesel-electric drive, in which a diesel engine powers an electric generator and the propellers are powered by electric motors. Its advantages are similar to those of the geared arrangement. With the exception of small powers, however, direct-current machinery is too heavy and too costly to compete with geared systems; and alternating current and high voltages are accompanied by certain problems, especially those associated with manoeuvring and reversing.

Gas turbines. The gas turbine has not fulfilled the expectation at the beginning of the 1940s that it would become popular as a marine prime mover. It has been applied to certain types of naval vessels, but it has seldom been used for mercantile ship propulsion. One of its basic problems is the low efficiency of the compressor; moreover, the metallurgical problems associated with the very high temperatures required for efficient operation are difficult to surmount.

Nuclear propulsion. In principle, the difference between a conventional steam installation and a nuclear power plant is that the combustion chamber in the former is replaced by a reactor using enriched uranium, the process of nuclear fission being utilized as the source of heat. The steam produced is fed into a normal set of steam turbines. The essential part of the reactor system is enclosed within a sealed, shielded structure capable of containing the products of any possible rupture in the reactor when in service. The weight of the shielding structure is more than three times that of the reactor system proper. The thickness of the surrounding concrete wall may vary from 30 to 48 inches (76 to 120 centimetres).

Many types of reactor are in use or under investigation in land plants. There are about 25 possible types of ship reactor. Nuclear plants for merchant ship propulsion ultimately may take the form of a direct-cycle reaction gas turbine, thus eliminating steam and other forms of intermediate heat exchange. Such a cycle may be a pressurized gas reactor coupled directly to a gas turbine. In many cases pressurized water is preferred as the primary coolant-moderator medium in marine installations; it is a logical development from high-pressure marine boilers.

The first nuclear propelled merchant vessel, the 21-knot NS "Savannah," which was launched in 1959, made clear the great potentialities of the system. This vessel is propelled by a single-screw, two-cylinder geared turbine arrangement delivering 21,000 shaft horsepower (shp). Steam pressure is only 460 psi, 475° F (245° C)—considerably below the usual 650 psi, 875° F (470° C) level of conventional marine boilers.

A conventional steamer of the same size and power as the "Savannah" is limited in its steaming range to 13,000 nautical miles before refuelling. The nuclear-powered ship can travel 350,000 nautical miles before fuel renewal is necessary. Whereas fuel for a fuel-oil fired ship is paid for when purchased at routine intervals, in a nuclear ship the first fuel charge must be built-in. Thus the initial capital cost is high.

The engine room. *Engine-room auxiliary machines.* In addition to the main propulsion machinery, an engine room contains a number of auxiliary machines. These auxiliaries fall into two groups: (1) those needed for ship services, and (2) those attendant upon the propulsion machinery. In the first group are the bilge, ballast, sanitary, freshwater, and general service pumps. The bilge pump deals with leakage water, and the ballast pump is used to empty the double-bottom ballast tanks before cargo is shipped.

Pumps

In steam installations, pumps are required for main condenser circulation and water extraction, boiler feed, and fuel-burning equipment. There are also evaporators, feed heaters, forced- and induced-draft fans, soot collectors, an auxiliary condenser with circulating pump, forced lubricating pumps, and oil purifiers. Pumps needed for diesel propelling engines include water-circulating pumps for cylinder jackets, lubricating oil pumps, filters and purifiers, oil transfer and surcharging pumps, and an auxiliary boiler for fuel heating and other services. Water pumps are almost invariably of the vertical, motor-driven, centrifugal type. Cogwheel or screw pumps, as well as reciprocators, are commonly used for pumping fuel.

Electric power for all purposes is provided by turbogenerators or by diesel-driven generators. Some of the auxiliaries run continuously at full load; others run intermittently, often at reduced load. To determine the total power requirements of the electric generators, the rated horsepowers of the various units are added and multiplied by a diversity factor that may vary between say, 0.60 and 0.75. On deck, there are cargo winches, capstan machinery,

anchor winches, steering gear, and, perhaps, stabilizers. Although operated by the deck department, the upkeep of this equipment is undertaken by the engineering staff.

The piping systems for bilge, ballast, fuel oil, and other services, although out of sight under the floorplates, are elaborate and costly. In cargo liners and tankers, the total weight of piping is 20 to 27 percent that of the main engines; the cost of pipework can be 35 to 45 percent that of the main engines.

Engine-room automation. Automation, to a greater or lesser degree, is standard. The makers of proprietary engines have their own characteristic systems. There are differences of opinion regarding the extent to which propelling engines and auxiliary equipment should be automatically and remotely controlled, and the number of engineers that should be available at a central console. The American Bureau of Shipping, Lloyd's Register of Shipping, Bureau Veritas, and Det Norske Veritas have each published memoranda and guidance notes; but none has issued mandatory regulations.

(C.C.P./J.F.C.C.)

Clothing and footwear industry

The clothing and footwear industries, sometimes called the apparel and allied industries, the soft-goods industries, or the garment industries, include factories and mills producing outerwear, underwear, headwear, footwear, belts, purses, luggage, gloves, scarfs, ties, and household soft goods such as drapes, linens, and slipcovers. The same raw materials and equipment are used to fashion these different end products.

HISTORY

In the late Stone Age northern Europeans made garments of animal skins sewn together with leather thongs. Holes were made in the skin and a thong drawn through with an instrument like a crochet hook. In southern Europe fine bone needles from the same period indicate that woven garments were already sewn. Weaving and embroidery were developed in the ancient civilizations of the Middle East. The equipment used in the fabrication of clothes remained simple and has always lagged behind the development of techniques for spinning and weaving. An important advance took place in the Middle Ages, when iron needles were introduced in Europe.

Invention of the sewing machine

All operations continued to be performed by hand until factory production of cloth was made possible by the invention of foot- and water-powered machinery for spinning and weaving in the 18th century. This development in turn stimulated the invention of the sewing machine. After several attempts a practical machine was patented in 1830 by Barthélemy Thimonnier, of Paris, who produced 80 machines to manufacture army uniforms. Thimonnier's machines, however, were destroyed by a mob of tailors who feared unemployment. Thimonnier's design used one thread; an American, Elias Howe, improved on it significantly with a lock-stitch machine that used two threads, a needle, and a shuttle. Though patented, it was not accepted in the United States; Howe took it to England, where he sold part of his patent rights. The objections of the U.S. tailors and seamstresses were overcome by a machine designed in 1851 by Isaac M. Singer of Pittstown, New York. When the sewing machine was first introduced, it was used only for simple seams; the more complex sewing operations were still done with a hand needle. The machines before Singer's were hand powered, but Singer quickly popularized foot-powered machines.

Before the second half of the 19th century, the fabric or leather sections of clothing and footwear were cut by shears or by a short knife with a handle about five inches (13.5 centimetres) long and a three-inch tapered blade. All pressing, whether the finished press or underpressing (between sewing operations), continued to be done with the stove-heated hand flatiron. The flatiron and the iron (later steel) needle were for a long time the only major advances in making clothing and footwear since caveman days. Tailors and dressmakers used hand needles, shears, short knives, and flatirons. Footwear was made using hand needles, curved awls, curved needles, pincers, lap stone, and hammers.

Early cutting and spreading machines

For many years the sewing machine was the only machine used by the clothing industry. The next major development was the introduction in England in 1860 of the band-knife machine that cut several thicknesses of cloth at one time. It was invented by John Barran of Leeds, the founder of the Leeds clothing industry, who substituted a knife edge for the saw edge of a woodworking machine. The resulting increased cutting productivity motivated the development of spreading machines to spread fabric from long bolts in lays composed of hundreds of plies of fabrics. The height and count of the lay depended on the thickness and density of the fabric as well as the blade-cutting height and power of the cutting machine.

The first spreading machines in the late 1890s, often built of wood, carried fabrics in either bolt or book-fold form as the workers propelled the spreading machines manually and aligned the superposed plies vertically on the cutting table, thus making the cutting lay. Although most of the early machines operated with their supporting wheels rotating on the cutting table, on some machines the wheels rode on the floor.

The Reece Machinery Company of the U.S. pioneered buttonhole machines at the end of the 19th century; later the Singer Company developed its own buttonhole machines and machines for sewing on buttons. The introduction of the Hoffman press enabled pressing to be done more quickly than by hand, although hand pressing is still used at various stages for high-grade garments. All of these developments made the factory production of clothing economic in industrialized countries. Though the first manufactured garments were shoddy in both make and materials, they were welcomed by poorer people, who previously had had to make their own. As the industry developed it improved the quality of production and materials and catered more and more to the well-to-do.

Social aspects. Until the second half of the 19th century, practically all clothes and shoes were produced by individual tailors and cobblers working either alone or with one or two apprentices or journeymen. The goal of every apprentice tailor was to learn how to make an entire garment as soon as possible. The output of a tailor or seamstress was usually limited to specific women's, men's, or children's garments; the journeyman sought to learn as much as possible from a specialized master craftsman. The same apprentice-journeyman system prevailed in the footwear industry, where all cobbler craftsmen were male.

The advent of the sewing machine enlarged craftsmen's shops and converted them to factories. In many factories workers owned their machines and carried them from factory to factory whenever they changed jobs. Needleworkers lugging their machines on their backs were a common sight on the downtown East Side streets of New York City, the garment-manufacturing capital of the world at the turn of the 20th century. Taking advantage of the low capital investment per worker, many clothing entrepreneurs began to farm out their cut garments to be sewn at home. The bundle brigades—men, women, and children trudging through the streets lugging bundles of cut or finished garments to and from their flats in the East Side tenements—replaced the sewing-machine carriers of previous years.

Most apparel factories at this time were as crowded, poorly lit, airless, and unsanitary as the home workshops. The term sweatshop was coined for such factories and home workshops at the beginning of the 20th century, when workers in the apparel industries began forming unions to get better pay and working conditions. The International Ladies' Garment Workers' Union, organized in 1900, and the Amalgamated Clothing Workers of America, formed in 1914, became pioneer unions in mass-production industries in the U.S. as well as the largest garment unions in the world.

Unionization

Modern developments. Throughout the first half of the 20th century, the apparel industry remained largely concentrated in the United States and the United Kingdom, especially the United States, where the industry received

an enormous impetus from World War II. In most other countries, garment making remained a home or cottage industry. The U.S. industry was divided among six types of firms: contractors, who produced apparel from raw material for a jobber or manufacturer; jobbers, who purchased raw materials they supplied to contractors to make into garments; manufacturers, who bought materials and designed, made, and sold the products wholesale; manufacturer-distributors, who sold their products through their own retail outlets; vertical mills, which performed all operations from yarn to finished garment under one corporate roof, and usually one plant roof; and vertical-mill distributors, who marketed their products through their own retail outlets.

By the 1950s other countries were beginning to develop and expand their apparel industries. Besides the United Kingdom, which continued to specialize in high-quality goods, the Scandinavian countries, Belgium, The Netherlands, Canada, South Africa, Japan, and Australia expanded readymade clothing manufacture. Another development of the 1950s was the expansion of many firms inside the industry into other areas; for example, some manufacturers of men's clothing entered the women's-wear field.

During the 1960s the garment industry of the world underwent rapid expansion, with many of the newer producing countries showing spectacular increases. Most of the industrialized countries of Europe and North and South America, as well as Australia, New Zealand, South Africa, and Israel, had clothing and footwear industries capable of meeting virtually all of their own needs. The United Kingdom, France, Italy, Spain, Sweden, West Germany, South Korea, Japan, Taiwan, and Hong Kong all expanded their export trade throughout the decade. Great Britain, which more than doubled its exports, continued to concentrate largely on men's fashion items in clothing and footwear. France exported principally high-fashion women's wear, especially in the form of selected original designs sold to manufacturers abroad to be copied and mass-produced locally. Italy became a major producer of knitted outerwear and of footwear; Israel exported knitted outerwear and all types of women's wear, especially pantyhose; Spain produced leather goods, knitwear, and high-fashion clothes; while Sweden and West Germany concentrated on sport and spectator wear.

The tremendous increase in productivity and exports of clothing and footwear from the Orient resulted from well-engineered factories established there during the 1960s and 1970s. These plants are not sweatshops like the crowded, ill-lighted factory lofts in which garment workers of the United States, the United Kingdom, and western European countries once worked 12 and 14 hours a day. Many Asian factory workers have better working and living conditions than obtained in U.S. and European plants in the 1920s and '30s, and in some cases Asian plant facilities are superior in working conditions and productivity to contemporary U.S. and European factories.

There has been, however, a distinct difference between Asia and the West in working hours and pay, though pay and hours have been upgraded in Japan, Hong Kong, and Taiwan. Beginning in 1968, for example, legislation in Hong Kong progressively reduced the factory work week to 48 hours, which was the average work week in clothing factories in the United States in the 1930s. By 1979 the average week in U.S. apparel plants was 35 hours; in the United Kingdom and western Europe the average week ranged from 28 to 45 hours. Wage rates in Hong Kong also have been increased, but daily rates there are still only slightly more than hourly rates in the United States. Radical differences in living costs and various fringe benefits to Hong Kong workers (*e.g.*, subsidized meals or food allowances) have reduced this differential.

None of the Communist countries of eastern Europe or Asia is a major exporter of clothing, but many, notably the Soviet Union, have developed large-scale manufacturing. In several countries, especially the Soviet Union and East Germany, highly developed production methods are widely used.

MODERN MATERIALS AND DESIGN CONSIDERATIONS

Raw materials. Raw materials used for apparel and allied products may be classified according to construction. Strand construction converts yarns into woven, knitted, and braided fabrics. Matted construction converts fibres into felts, paper, and padding yardage. Molecular-mass construction produces plastic film, metal foil, and rubber sheetings, while cellular construction is the building block for skins, furs, hides, and synthetic foam.

All four constructions are used for all types of apparel, though only minute quantities of molecular-mass and cellular construction are used for underwear. Most outerwear is made from woven and knitted fabrics with some use of hides, skins, furs, plastics, rubber, foams, and metallics. Footwear that was originally made exclusively from leather (treated hides) may now be made from fabrics, plastics, rubber, foams, and metallics.

Textile fabrics. Woven fabrics are constructed by interlacing two or more yarns perpendicularly to each other. Braiding is an interlacing in which two or more yarns are interlaced diagonally to each other. In knitting, yarns are interlooped. Yarns are strands spun from either natural fibre such as cotton, linen, or wool or from synthetic fibres such as rayon and nylon. Practically all synthetic fibres are made originally in filament form and then cut into staples, or fibre lengths. A textile filament is a single hairlike strand of indeterminate length. The only natural filament is silk.

Weave construction

Basic weave constructions are plain, twill, satin, basket, jacquard, lappet, leno, and pile. The two basic knit constructions are warp, or flat, and weft, or circular knitting. Types of weft knitting are jersey, rib, purl, run resist, tuck stitch, and interlock. Types of warp knitting are tricot, milanese, and raschel simplex. The classifying is based on principles of linking the yarns in structuring the fabric. (See also INDUSTRIES, TEXTILE.)

Leathers and synthetics. Leathers are made from skins of sheep, goats, kids, calves, pigs, horses, cattle, lizards, snakes, alligators, elk, buffalo, ostrich, kangaroo, chamois, walrus, elephant, and seal (see also FURS, LEATHERS, AND HIDES). The most common leathers are from the first seven listed; the others are exotic leathers, used to a lesser degree. Suede and patent leather are types of finishes. Plastics, foams, felts, paper, rubber, and metallics are used in thicknesses ranging from gossamer, cobweb thinness to the thickest of hides.

Quality in apparel and allied products. Quality is measured in three characteristics—durability, utility, and emotional appeal, with respect to the raw materials used, the product design, and the construction of the product.

Durability factors are tensile strength, tear strength, abrasion resistance, colourfastness, and cracking and bursting strength. Utility factors are air permeability, water permeability, thermal conductivity, crease retention, wrinkle resistance, shrinkage, and soil resistance. Appeal factors are eye appeal of fabric face, tactile response to fabric surface, fabric hand (reaction to hand manipulation of the fabric), eye appeal of the garment's face, silhouette, design, and drape. The principles involved are the same whether the garment is made of leather, plastic, foam, or textiles such as woven, knit, or felt fabric.

The quality of textiles and apparel is usually governed by standards set up by the industry or the government or the two acting together, along with textile associations, testing societies, retail associations, consumer organizations, and service associations such as laundering and cleaning groups. Standards compiled in many places cover performance in cleaning and wearability for garments in several categories: tailored clothing, outerwear and rainwear, work clothes, slacks, shirts and blouses, men's underwear, women's underwear, sleepwear, and accessories.

Design in clothing and footwear. Clothing, headwear, footwear, and accessories businesses are the fashion industries par excellence. As such their goal is to give the wearer a sense of well-being based on being attractive to oneself and others. At the same time, an inescapable function of fashion in most countries is to serve as a status symbol, a consideration leading to the wardrobe concept in designing; that is, separate business attire, evening wear, casual

clothes, spectator-sportswear, active-sportswear, and other prescribed attire that clothing entrepreneurs and designers promote.

Five elements of design

Designers use five elements to create a design that will stimulate the potential consumer to buy: (1) colour, (2) silhouette, (3) drape, (4) texture, and (5) line balance on the product's surface. Design ideas are derived from configurations or colour combinations or both in historical, ethnic, national, natural, geographic, and current themes that can be advertised and promoted to boost sales. Besides choosing textiles and other raw materials with specific properties and characteristics of colour, drape, hand, and texture, the designer selects findings—buttons, zippers, snaps, grommets, thread, lace, tapes, braids, medallions, sequins, and a variety of ornaments and closures—as decorative devices to impart the desired design effect. Line balance is generally achieved by the sectional patterns of raw materials such as textiles and leather that the designer shapes to form the finished product. Fashion in clothing and footwear operates in cycles, but with respect to design construction, clothing and footwear may be lined, interlined, or unlined regardless of the cycle.

Designers use a variety of fitting forms for clothing, shoe lasts for footwear, and hat blocks for headwear, approximating human anatomical dimensions. The basic patterns that fit the form, last, or block with minimal seams and skintight precision form the foundation pattern. In the drafting method of designing style patterns, the designer manipulates the foundation pattern to develop the style pattern, which is used to cut the raw materials into the sections required for the garment. The drapery method is used by designers who prefer to drape the actual material for the garment on the form, block, or last; cutting patterns for cutting the raw materials are traced from the draped sections. Cutting patterns may be made of rigid or semirigid paper or plastics.

Pattern grading

Pattern grading, making sets of patterns to fit a range of sizes, is the next step in the design process. Anthropometric tables for sizing apparel have been compiled by various government agencies and other sources. Formerly pattern grading was a completely manual drafting process, but in the 1950s pattern-grade machines were invented to increase the speed of grade-drafting patterns.

These machines were manually paced; that is, they followed the lead of a draftsman. But in 1967 the computerized grader machine was invented to grade and cut patterns directly from an original set, performing automatically without manual contact with the drafting and cutting process. By 1970 various U.S., British, and Japanese computerized graders were available, which automatically take measurements from a set of style patterns, then grade, draft, and cut sets of patterns in the sizes desired. These computerized graders contain memory banks of the anthropometric specifications required. Some computerized grading machines can provide graded patterns in any of three media: acetate patterns for making photomarkers; rigid plastic patterns for conventional chalk, crayon, pen, pencil, or spray manual marking methods; or stack drafts, which put the entire desired size range from smallest to largest, concentrically, on one sheet. In conjunction with any of these, the computer can issue data sheets with costing and material usage per size or range. The actual processing of the raw material begins after the patterns are graded.

MODERN MANUFACTURING PROCESSES AND EQUIPMENT

Many different sequences of the three major processes, cutting, sewing, and pressing, are used. The exact sequence depends on the raw materials for the garment, the processing equipment, the garment's design, and quality specifications. Five other processes are used to assemble, decorate, and finish the components into the finished garment: baking or curing, cementing, fusing, molding, and riveting, including grommetting and nailing.

Cutting processes. Cutting involves three basic operations: making the marker, spreading the fabric, and chopping the spread fabric into the marked sections (see Figure 6). The marker, or cutting lay, is the arrangement of patterns on the spread fabrics. When hides are cut the lay

Spreading

length is the hide size; many hides are cut in single plies. Short lengths are spread by hand, but large lays, made from large bolts of material, range in length to over 100 feet (30 metres) and heights containing hundreds of plies and must be spread with travelling spreading machines. Stationary spreaders are used for small sample lots. Manual and semi-automatic spreading machines are propelled manually over the lay length as the machine feeds the fabric ply onto the cutting table. Some machines bookfold the successive plies as the fabric is spread; others have turntable devices permitting one-way spreads. Lays may be spread either with all plies of fabric facing one way or with successive plies facing each other in face-to-face spreads. Turntable spreaders were introduced in 1920, face-to-face spreaders in 1938, and in 1946, electric-powered spreading machines that spread a full bolt automatically without manual attention. In 1950, cutting blades were invented to cut the ply at each end of the lay as it is spread. These cut-off spreaders are automatic. Electric-eye edge controls for precise superposing of plies were available on automatic machines in 1962. In 1969 piggyback automatic spreaders were introduced, which carry a second bolt that is spread as soon as the first bolt is on the lay.

Making the marker and cutting

The marker is superposed on the completed lay. Makers are made of one of three materials: the fabric being cut, an inexpensive felt of muslin type cloth, or one of a variety of papers. When paper with a low coefficient of friction is used, the marker is fastened to the lay by stapling or two-sided adhesive stripping. Papers with an adhesive on one side can be heat-sealed to a fabric and are commonly used with woollens or soft fabrics. Photomarking machines are used for duplicating often-used paper markers. Many markers are first made in miniature, with precise, scaled-down patterns to determine the optimum layout for minimal yardage; the optimal miniature marker is then used as the guide for making the full-scale cutting marker. Some automated equipment is capable of both making the graded pattern and laying it out on the fabric to minimize waste. A sprayer machine, which sprays the entire length of the lay around the pattern, eliminates the need for manual marking-in.

Six types of machines are available to chop or cut a lay into the component parts of the marker: rotary blade machines; vertical reciprocal-blade machines; band-knives, similar to band-saws; die clickers, or beam presses; automatic computerized cutting systems with straight blades; and automated computerized laser-beam cutting machines.

Round-knife machines rotate a circular blade down into the lay whereas straight-knife machines oscillate a straight blade in and out of the lay in jigsaw fashion. Both machines are portable, manual-paced machines; that is, the machine is pushed through the lay as the blade cuts. Some models have dual speed controls and automatic blade sharpeners. In band-knife cutting, blocks cut from the lay with round or straight-knife machines are trimmed precisely to pattern specifications as the blocks are manipulated against the band-knife rotating in a fixed orbit. Though most band-knife machines are stationary, some are mounted on travelling platforms that carry machine and operator along the entire length of the cutting table, permitting band-knife cutting at any point of the lay.

Round-knife machines vary in diameter and rotary speed of the blade. Vertical and band blades come in circular, waved, or sawtooth perimeters. Vertical blade edges may be straight, waved, notched, serrated, or striated; band blades may be straight, waved, or saw-toothed. Straight-edge blades including the circular perimeter of rotary blades are used generally; the others are special-purpose blades.

Die clickers cut by pressing dies, superposed on the lay, through the depth of the lay. The cutting dies outline the patterns to be cut. Die presses are stationary or travelling; travelling die presses cover the entire width of the lay and move throughout the lay length and press dies into the fabric with intermittent strokes across the lay width until the entire lay is cut. In stationary clickers, the lay or section of the lay is pulled under the pressure beam for each die cutting stroke. Machine cutting of footwear, bags, pocketbooks, and similar items is done with die presses.

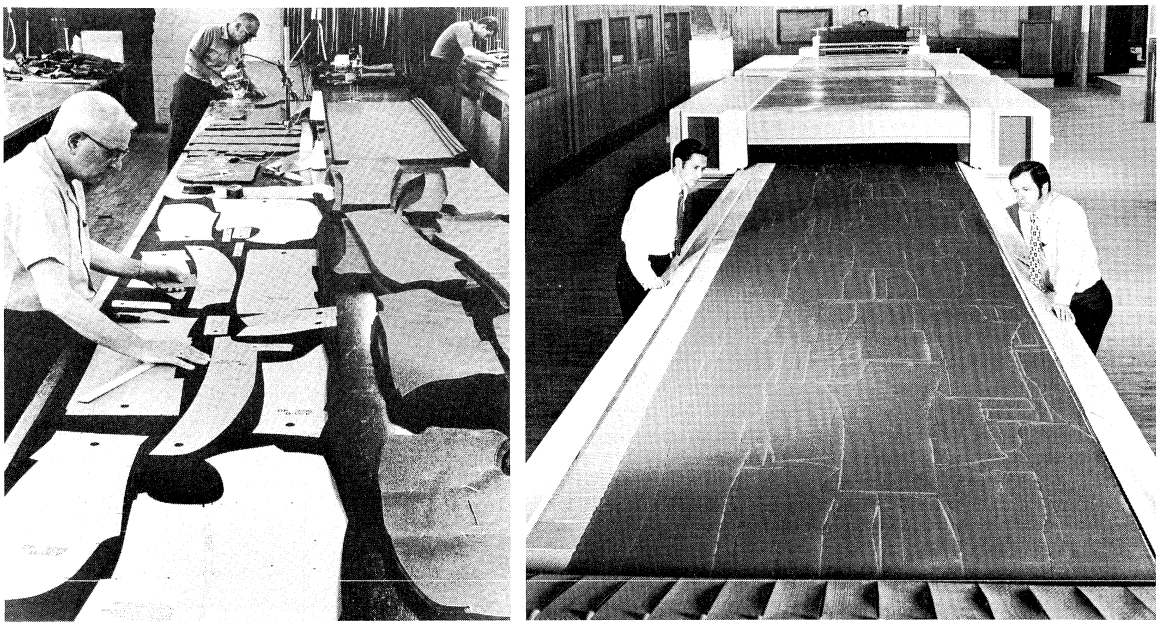


Figure 6: *Pattern cutting.*
(Left) Traditional manual operation: working with multiple plies of fabric, patterns are laid out and chalked (foreground), then cut with an electrically powered tool (background). (Right) Laser beam apparel cutter: single ply fabric is unrolled in 5-foot sections onto a conveyor belt (background); programmed by a separately housed computer for pattern specifications, the laser beam, directed by a pentaprism arrangement of mirrors, cuts the pattern (centre); excess material is then removed and pieces are rolled by conveyor belt (foreground).
By courtesy of (left) Hart Schaffner & Marx, (right) Genesco, Inc.

In automatic computerized cutting systems, introduced in 1967, the lay is covered with a thin plastic film drawn firmly to the lay by a vacuum operating through a porous cutting table and the porous fabric of the lay. The vacuum pulls the impermeable film firmly onto the lay preventing any movement during the cutting action. There are two types of cutting actions: the lay may be stationary and the knife move or the lay may move forward and the knife move horizontally.

A cutting system introduced in 1971 employed a computer-controlled laser beam to burn, or vaporize, the fabric rather than cut it. Unlike other methods requiring the accumulation of large orders before it becomes efficient to cut a specific style, the laser system, which provided for storage of programmed cutting instructions, allowed one complete garment to be cut at a time from a single layer of material (see Figure 6). Among the advantages claimed for the system are the elimination of variations within a specific size, improved cloth utilization, efficient production of smaller orders, lower inventory requirements, and faster delivery.

Two types of auxiliary cutting equipment are used: cutting drills to drill holes through the superposed plies in a lay, and notches to notch the perimeters of the cut sections. These holes and notches guide the sewing operations. Cut sections are ticketed to insure proper sizing and shading during the assembly of the garment.

Sewing production. Clothing, footwear, and allied industries have been known as the needle trades because sewing is the major assembly and decorative process used. Some items such as plastic raincoats and footwear are assembled and decorated by fusing, but only a tiny fraction of garments were produced completely by fusing, cementing, and mold casting.

Sewing machines

Over 10,000 different models of industrial sewing machines have been made. Most are produced in Great Britain, Germany, Italy, Japan, and the United States. Sewing machines are classified, according to stitch type and bed type (the shape of the machine's frame). The seven basic beds, or frames, are flatbed, raised-bed, post, cylinder, off-the-arm, closed-vertical, and open-vertical. The bed type is determined by the manner in which fabric passes through the machine as it sews. There are four categories with regard to operational control, all electrically powered: manual-paced, automatic cycle with manual loading and extraction, fully automatic, and automated.

The prime characteristic of a sewing machine is the stitch it makes. Until 1926, stitches were classified willy-nilly with trade terms that often varied from one place to another and even from shop to shop. In 1926 the U.S. government became the first government to issue a seam and stitch classification to specify its requirements. During the 1960s other countries began to adopt the latest versions of these specifications for sewing equipment and sewn products, and these specifications are now being adopted throughout the world for industrial sewing.

The first hand-powered sewing machines in the 19th century sewed 20 stitches per minute. At the turn of the century some electrically powered machines sewed 200 stitches, and by mid-20th century machine speeds reached 4,500. By 1970 most machines could sew 7,000, and some could sew 8,000 stitches per minute. The first integrated sewing machine was introduced in 1969 by the Singer Company. Before that, manual-paced sewing machines had a separate clutch motor with start, speed, and braking controlled by foot treadle action; a belt drive ran the machine via treadle action to the motor's clutch, which actuated or stopped the belt drive while the motor ran continuously. The integrated sewing machine eliminated the separate motor, its clutch, and the belt drive. The integrated machine frame contains the motor module, which is actuated, controlled, and stopped by a four-speed switch operated by a device resembling the former foot treadle. The motor in this machine rotates only when the treadle is actuated, and so electricity is used only when the machine is sewing.

Special-purpose machines sew automatic cycles for operations such as buttonholing, button sewing, contour seaming, profile stitching, seaming for patch pockets, dart stitching, tacking, welt pocketing, and padding cycles such as blindstitching interlinings to the outer shell. Semiautomatic special-purpose machines are manually reloaded after each task; in automatic machines reloading and extraction are both done by the machine. Contour seamers are sewing machines that sew curved seams automatically; most such machines were semi-automatic by 1970. Profile seamers and stitchers seam or stitch angular or curved designs in which a backtrack path, such as a U seam or angled seam (a square U), is sewed. These machines, too, were semi-automatic by 1970.

A sequential sewing machine introduced in the 1960s repeatedly sews an automatic cycle on the same garment with predetermined spacing between the operations. A sequential buttonhole machine, for example, sews five buttonholes on a shirt front automatically, one after the other, with given spacing. Sequential sewing machine modules are synchronized automatic systems of two or more sewing machines that sew the operation in series; the first machine completes its operation, the garment or section is fed into the second machine for the next operation, etc. For example, the first machine sews the centre front placket of a shirt front; the second machine sews a series of buttonholes on the placket. In tandem-machine arrangements, two machines sew simultaneously on the same unit. Gang-machine operation is an arrangement of three or more machines operating automatically under the care of one operator. The introduction in 1930 of stop-motion devices for stopping a sewing machine when the thread broke or ran out made gang-machine operation possible, as well as tandem machines in the 1940s and sequential machines and modules in the 1960s.

Before 1950, most industrial sewing machines had only the basic mechanical-linkage system of shafts, cams, gears, rods, belts, chains, and pulleys, with manual lubricating systems. Higher speeds, fully automatic cycling, and automatic sequential systems were developed later and were made possible by automatic lubrication systems with pumps and reservoirs, fluidic controls, and electronic controls.

The quality of manual-paced sewing depends on the integration of six variables: the needle, its size, shape, and finish; the type of feed system; the coordination of needle and feed; thread tension adjustments; the thread; and operator handling. Seam slippage, yard severance, puckering, elongation, gathering, and feed mark off are some of the quality areas affected. Machine manufacturers make needles in a variety of diameters, point shapes, and finishes as well as different types, shapes, and sizes of feeds and presser feet to improve quality and output.

Sewing machine attachments

Sewing machine attachments are jigs and fixtures used with sewing machines to decrease downtime (the time a machine is inoperative) and thus increase productivity by getting the fabric to the needle, aligning and repositioning fabric under the needle, or extracting and disposing of the sewed materials sooner. Trade terms for some of these sewing aids are needle positioners, stackers, programmers, guides, hemmers, binders, thread trimmers, stitching templets, seam folders, pipers, and shirrers. Needle positioners automatically set the needle in or out of the sewn materials as desired when the machine stops. Stackers extract and dispose the sewed sections with one of five actions: flipping, sliding, lifting, shuttle-drop, or conveyor cycle. Programmed sewing is an automatic sewing cycle induced by a module set to control the time sequence of initial positioning, sewing, repositioning if needed, thread trimming, extraction, and disposal. The times and sequence of these elements in the sewing operation may be changed for different sewing cycles. Automated sewing is a self-correcting system when the section sewed varies beyond given tolerance limits.

Fusing and cementing are two major processes for stitchless or decorative seaming in apparel and allied production. In fusing, the seam bond or decoration is formed by melting some fibre or finish content in the material in a manner that joins the sections or decorates in the desired area. In cementing, the bond, or decoration, is made by an adhesive, such as cement, glue, or plastic, which is applied to the materials during or immediately preceding the cementing process. Fusing is either by direct heat; by hot-head fusing presses, in which pressure surfaces are heated by electric heating grids or steam; or by electronic high-frequency or infrared systems. Cementing processes use mechanical-pressure systems with or without heat application, depending on the adhesive and materials used. Fusing, introduced in the 1950s, is replacing sewing in some operations such as joining interlining to collars, cuffs, and coat fronts as well as seaming clothing and footwear made from certain synthetic yarns or plastic films.

Pressing and molding processes. Molding is any process that changes the surface characteristics or topography of a garment or shoe or one of its sections by application of heat, moisture, or pressure. Pressing, pleating, blocking, mangling, steaming, creasing, curing, and casting are trade terms for various molding processes in producing clothing and footwear.

Buck pressing and iron pressing

Pressing has two major divisions: buck pressing and iron pressing. A buck press is a machine for pressing a garment or section between two contoured, heated, pressure surfaces that may have steam and vacuum systems in either or both surfaces. Before 1905 all garment pressing was done by hand irons heated directly by gas flame, stove plate heat, or electricity; the introduction of the steam buck press changed most press operations. The first pressing machines had no pressure, heat, or steam controls such as those built after 1940. Modern buck presses are made to fit certain garment sections, such as a jacket front, pant leg, pant top, or shoulder area for a specific style and size. These improved buck presses have gauges to measure and control steam pressure and temperature, mechanical pressure, vacuum, and the press cycle time. Cycle-time controls permit one operator to work a series of machines. For example, a presser handles four presses doing the same or different operations; by the time a worker finishes extracting and loading the fourth machine, the first machine is ready for extraction and reloading. Cycle-time controls apply and shut off steam and vacuum action and open the pressing machine automatically. Conveyor buck-presses, which may move intermittently or continuously, are buck-pressing systems in which sections or garments to be pressed are fed into a buck press and extracted from it by a conveyor belt.

In iron pressing, a hand iron functions as the top pressure surface. The two major types of hand irons are steam ejectors and dry irons. Electric hand irons are equipped with thermostats that regulate temperature. Steam-heated irons, whether ejection or dry, have fixed temperatures depending on the pressure of the steam supplied to the iron. Many hand irons are equipped with lift devices and gear drives to control stroke rate and minimize operator fatigue. Hand irons are made in a variety of sizes, weights, shapes, and surfaces; the specific usage determines the combination.

Pleating. Pleating is the process of putting a design of creases into fabric. Accordion, side, box, inverted, sunburst, air-tuck, Van Dyke, and crystal are trade terms for some pleat designs. Pleating is accomplished by machine or by the use of interlocking paper pleat patterns. Pleating machines have blades or rotary gearlike surfaces that crease the fabric as it passes between two heated rotary mangles, setting the creases. Machines may be used for pleating either specific cut garment sections or lengths of fabric that are then cut after pleating into garment sections. In pattern pleating, the garment section or fabric length is sandwiched between two complementarily creased plies of paper that shape the fabric into the desired pleat design. This creased trio is inserted in a steam chamber, or autoclave, for a given length of time, depending on fabric characteristics and pleat durability desired.

Creasing. Creasing machines differ from pleating machines in that they fold the edges of garment sections and set the fold crease as an aid for such operations as sewing the edges of collars, cuffs, and patch pockets. Creasing diminishes the time for positioning the creased section during sewing.

Mangling. Mangling is the process of pressing a garment or section between two heated cylindrical surfaces.

Blocking. Blocking consists of encompassing a form, block, or die with the garment with skintight precision. The item is blocked or pressed by superposing a complementary pressing form that sandwiches the shaped garment or section between the interlocked blocks. This process is used for such items as hats, collars, cuffs, and sleeves.

Curing. Curing consists of baking a garment or garment section in a heated chamber to either set creases in the fabric permanently or to decompose auxiliary media used as a sewing aid. For example, curing permanently sets previously pressed creases in certain permanent press, durable press, and wash and wear garments. Curing decomposes

the backing material used for facilitating the embroidering in certain embroidered garments.

Casting. Casting consists of making a garment or garment section by pouring a fluid or powder into a mold that forms the garment or section when the fluid or powder evaporates or solidifies.

Special footwear processes. Footwear may be classified according to the section of the foot it covers and how it is held on: sandals, slip-ons, oxfords, ankle-support shoes, and boots. The term shoe refers to footwear exclusive of sandals and boots. Sandals cover only the sole and are held onto the foot by strapping. Slip-ons cover the sole, instep, and may or may not cover the entire heel; styles include pumps and moccasins. Oxfords cover the sole, instep, and heel and have closures such as laces, straps, buckles, buttons, or elastic to secure the shoe to the foot. Ankle-support shoes cover sole, instep, heel, and ankle and secure the shoe to the foot with a closure device; the chukka is an ankle-support style. Boots cover the foot from the sole to various heights above the ankle: shin height, calf length, knee length, and hip length. Closures may or may not be used, depending on the degree of snugness desired.

Most footwear factories that produce dress, play, and work footwear in slip-on, oxford, ankle-support, and boot categories from leather or synthetics simulating leather have eight processing departments: (1) cutting; (2) stitching, which sews the upper section above the sole; (3) stock fitting, which prepares the sole section; (4) lasting, which attaches the upper and its lining to a wooden foot shape, the last, in order to assemble the sole section to the upper; (5) bottoming, which attaches the sole to the upper; (6) heeling, which attaches and shapes the heel bottom into its final form; (7) finishing, which includes polishing, extracting the lasts, stamping the shoe brand and name on the sole, inserting heel and sole pads, and inspection of the inner shoe; and (8) treeing, which includes attaching laces, bows, and buckles, and final cleaning and inspection.

There are three basic methods of attaching soles to uppers. The bottoming may be done by sewing, cementing, nailing, or a combination of these three joining techniques. Nailing includes the use of nails, screws, staples, or pegs. Sewing may be performed with or without the use of welt, insole, middle sole, and filler sections; the same applies to cementing soles to uppers. Sole sections vary in ply count; a three-ply sole has a middle sole sandwiched between outer sole and inner sole; the two-ply sole consists of outer and inner sole; the single sole has only one ply.

PRODUCTION CONTROL AND PLANT CONSIDERATIONS

Division of labour. In the first clothing and footwear factories, one worker assembled and finished an entire garment or shoe. But this whole-garment system rarely existed after 1940 for ready-to-wear apparel. Sectionalization came into being for three reasons: (1) to increase productivity per man-hour, (2) to improve product quality, and (3) to reduce inventory-in-process time. The main function of production control is to schedule the operations required to produce the garment or shoe in such a manner as to hold the total processing time, or calendar time, to a minimum. This scheduling is accomplished by determining the operations required per garment to yield the desired quality with minimal processing costs (labour, utilities, capitalization) and by arranging these operations so that most of them may be done simultaneously in the least number of successive steps, or time units. If 24 operations are required to make a given garment, a schedule in which these 24 operations are performed in six time units, with four operations made simultaneously in each unit, is superior to a schedule requiring seven or more successive time units with four or fewer simultaneous operations in each time unit.

Unit flow and multiple flow. There are two basic section systems in assembly-line production: unit flow; and multiple flow, or bundle or batch system. The unit-flow system is continuous; the unit moves to the next operation as soon as it is processed by the previous operation. The multiple-flow system is intermittent; a given group of units moves to the next operation simultaneously as a bundle, or batch, after the last unit in the batch is processed in the previous operation. Regardless of which flow is used, efficient production control depends on rapid, synchronized movement of the garment sections.

Plant layout and materials handling. To minimize the cost of transporting and storing processed garment sections from the first operation through all the prescribed production sequence until completion, two factors are considered: the arrangement of processing equipment to minimize distances between operations; and the utilization of labour, transport equipment, and utilities to minimize total cost per product moved.

Transportation equipment such as chutes, conveyor systems, and carts are utilized to transport the work through the successive operations. Some conveyor systems, such as Eton of Dalsjöfårs, Sweden, incorporate a stationary synchronized extractor on the conveyor that extracts and disposes a sewed section to a mobile conveyor hanger.

The two types of plant layout are process layout and product layout. In process layout, equivalent processing equipment is grouped in the same area regardless of whether the product can be passed through successive operations without backtracking. In production layout, processing equipment is arranged in the succession required for making the garment in the prescribed sequence without backtrack and with minimal transport between successive operations. Most apparel and allied production is with product layout, but in a few instances the process layout is more economical.

(J.So.)

Floor coverings

Included in this section is a treatment of both handmade and machine-made rugs and carpets and of smooth-surfaced floor coverings. Although the words carpet and rug are frequently used interchangeably in referring to textile floor coverings, in modern usage carpets are fastened to the floor and usually cover an entire floor area, and rugs are not fastened and rarely cover the entire floor. Carpets and rugs may be classified as handmade or machine-made.

Handmade carpets and rugs are usually made by knotting a number of pile tufts to a backing structure so that the loose knot ends form the pile. Mainly produced in Asia and the East, knotted types are often given the general name of Oriental carpets and may be classified according to the country of manufacture, such as Persian (originating in Iran) or Chinese. Well-known districts or towns may give their names to the carpets they produce; the Persian Kermān (Kirman) is an example. Other hand-knotted rugs include the Savonnerie rugs of France and the modern rya rugs of the Scandinavian countries. Handmade rugs woven by the tapestry method, often described as the Aubusson type, for the French town in which the method was perfected, have a flat or slightly ribbed surface instead of pile. Hooked rugs are made by drawing yarn or fabric strands through a basic material so as to form a pile of loops that may be clipped or remain uncut.

Machine-made carpets include such woven types as Axminster and Wilton, and also tufted, knitted, and flocked types. Axminsters resemble hand-knotted carpets, but their pile yarn is mechanically inserted and bound and not knotted. Wilton types may have looped (uncut) or cut pile, with designs formed by bringing yarns of the desired colour to the surface and burying the others beneath the surface. Velvet carpeting is made by looping strands that form the pile over wire strips that are removed as each row of loops is completed. Chenille rugs have soft, deep pile formed by long, furry strips. The pile of tufted carpets is formed by tufts inserted into a backing with needles. In knitted carpets, the backing, locking, and pile yarns are all looped together. Flocked types are produced by systems in which adhesives are used to bind fibres or yarns to the backing fabric.

Smooth-surfaced floor coverings include linoleum, rubber floor coverings, cork tile, asphalt tile, printed felt base, and the vinyl types. Most are available in varying degrees of thickness, usually from 1/16 to 3/16 inch, and may have some form of backing. Although flexible types

are available in rolls at least six feet wide, square tiles are increasingly popular in all types. The various materials differ in their ability to take colour or pattern, and mottled or spatter effects are often achieved by blending in one or more mixes of the same general composition as the basic material.

HISTORY

Early floor coverings. Prehistoric man may have happened upon a method of forming thread from twisted grass or hair. Evidence obtained from recent excavations near the Caspian Sea indicates that the shearing of sheep and goats, and the spinning and weaving of the fibres obtained, was practiced as early as 6000 BC. Before the development of weaving, fibres were probably interlaced to produce a simple form of plaited basket-work matting, replacing still earlier crude mats made of strands of dry stalks and tendrils.

Early matting

Findings in burial mounds at Pazyryk in southern Siberia, 2,400 years old, indicate that furs, leather, woven textiles, and felts were used, not as floor coverings, but as wall hangings. The first true carpets, characterized by pile surfaces, were probably rough cured skins that early hunters laid on the floors of their crude dwellings. Most carpets still retain the same tough flexible backings and upright pile, affording protection from cold and hard floors, agreeable to the touch, and serving a decorative function.

Smooth floorings also have ancient origins. In the Late Bronze Age (1600–1000 BC) water-worn pebbles were laid as flooring in Crete and also on the Greek mainland. The Greeks refined the technique between the 6th and the 4th centuries BC, and ancient decorative pebble mosaics have been found in Greece, Asia Minor, and Sicily. Such mosaics were also made of marble, serpentine alabaster, some forms of granite, and other stones suitable for polishing. Timber flooring, originally used in rough form for a strictly functional purpose, was eventually made into smooth boards, and was later used decoratively in parquetry designs.

Carpet and rug weaving. Although the exact origins of carpet weaving have not been determined, it is known that the Egyptians of the 3rd millennium BC wove carpets for the most part of linen ornamented by sewn on brightly coloured pieces of woollen cloth. Egyptian influence apparently spread throughout the Middle East and then to Mongolia and China. Some investigators credit Central Asia, Turkestan, and China with the origination of carpets, and in the early 1950s a rug dating back 2,400 years, made with Turkish knots, was found in Siberia.

Early Chinese carpets were made of knotted silk pile with backings of wool or cotton, but the pile of later carpets was made of wool. Wool pile was also used in Central Asia by early nomadic tribes who acquired it easily in their wanderings. Nomadic rugs were woven on simple horizontal frames that could be rolled up for travelling.

Early looms

Early looms consisted of two forked branches joined by a crosspiece holding the suspended warp, or lengthwise threads, through which the weft, or crosswise threads, were woven. A wooden bar was used to flatten the binding weft threads, allowing the loose warp ends to stand out to form the luxurious pile. The early weavers used wools in their natural gray, white, cream, fawn, brown, or black colours, but eventually learned to produce fast colours with dyes made from vegetable, flower, and insect materials.

During the Middle Ages, Italian merchants imported Oriental rugs to Europe, where they were usually hung on the walls; Europeans continued to cover their floors with rushes and straw. Moorish weavers were probably taken from Spain in the 13th century to set up the looms at Aubusson in France. Eleanor of Castile introduced Spanish rugs to England in 1255, and carpets imported from Turkey in the 15th century encouraged the development of an English rug-weaving industry.

By 1600 French carpet weavers had formed a strong guild, and in 1608 Henry IV set up looms in the Louvre. During the reign of Louis XIV, carpet manufacture was revived at Aubusson, where it had suffered from the religious wars of the 16th century, and was established at Beauvais, in Normandy. The revocation of the Edict of Nantes, that had guaranteed religious and civil freedom to French Protestants, drove French and Walloon Protestant artisans into England and Germany, where they contributed to the development of spinning and weaving techniques.

English carpet weavers were chartered at Wilton and Axminster in 1701, and in 1740 the Earl of Pembroke brought weavers from France to perform Brussels and Wilton weaving. At about the same time, carpet weaving was also established at Kidderminster, and the trade extended to northern England and Scotland. In 1830 a Parliamentary paper noted that carpet wool comprised one-twenty-eighth of the wool produced in the United Kingdom.

In the 18th century Richard Arkwright and others invented machinery that radically improved textile manufacture and together with the steam engine led to the development of the power loom, first applied to carpet making in 1839. The so-called Jacquard mechanism, which employed punched cards to control the warp yarns, gradually began to replace the complicated harness of the hand loom for the production of designs. The tapestry process of printing patterned carpets was evolved in Edinburgh in the 1830s, and in 1839 a chenille Axminster process, which was patented by James Templeton of Glasgow, gave increased colour range to carpet designs.

The U.S. carpet industry began by adapting the British system on a modest scale. Largely a cottage industry, it was organized by agents who marketed the small amount produced, until the first half of the 19th century. By 1830, the use of carpets had become popular throughout the eastern U.S., and factories were being established in New England, New York, and Pennsylvania. The continued dominance of the U.S. market by British carpets led U.S. manufacturers to encourage the development of power equipment, and a power loom first appeared in 1841. In 1876 an Axminster loom was invented. This development stimulated the carpet industry by permitting an unlimited range of colour and design with an economy of pile.

Broadloom

Loom widths increased from the formerly conventional 18, 27, or 36 inches (46, 69, or 91 centimetres), to the broadloom, usually 12, 15, or 18 feet (4, 5, or 6 metres) wide, resulting in large economies in weaving costs and producing larger and more convenient unseamed areas for laying. After World War II, needle tufting developed, employing a prewoven backing for the basic construction, and the major portion of carpeting manufactured in the U.S. was produced by this system. Some tufted carpet manufacturers even began to produce outdoor carpets and imitation lawns.

Smooth-surfaced floor coverings. In 1860 Frederick Walton of Great Britain patented a process for making linoleum, the first widely used smooth-surfaced floor covering. Plain linoleum, without design, was popular until the mid-1930s, when decorative linoleum was developed. In the 1920s, dark-coloured asphalt sheet and tile materials were developed in the U.S., made from mixtures of asbestos fibre, mineral fillers, and asphalt, and although light-coloured resins, not containing asphalt, became available within the next 10 years, the name asphalt tile persists in the U.S. for this type of flooring. In the U.K. the term asphalt tile was used for a different product, and the somewhat misleading term thermo-plastic tile was applied to a similar British product. Vinyl asbestos tiles, containing asbestos fibres, were developed next and introduced at the Chicago World's Fair in 1933, but resin shortages prevented quantity production until 1948. Vinyl, a newer composition material with a high content of polyvinyl chloride resins, was eventually perfected. The number and variety of smooth-surfaced floor coverings multiplied after World War II, and plastics had considerable impact. Although traditional linoleum was still in use, such materials as asphalt, cork, rubber, vinyl asbestos, and the various types of vinyl were achieving greater popularity. A new development in the 1960s was a type of flooring applied directly to the area to be covered and allowed to harden; epoxy resins have generally been used.

HANDMADE CARPETS AND RUGS

Orientals. Major classifications of Orientals, based on place of origin, include Persians, the largest and most

important group; Turkomans, popular, vividly coloured carpets including Turkoman, Afghan, and Baluchistan rugs made in Central Asia; bold, geometric patterned Caucasian carpets, from Caucasia and Transcaucasia; the Turkish Anatolian group, less intricately designed than other Orientals; and the Indian, Pakistani, and Chinese group, frequently less durable than the other types.

Materials. The availability of excellent materials is probably the factor most responsible for the origin of carpets in the East. The nomads had access to fibres from their camels, goats, and sheep; cotton was cultivated in Persia and China, and silk in China. Nomadic carpet makers often used wool for the warp and weft of a rug foundation fabric, as well as for the pile. Although a variety of materials may be used in making Oriental rugs, wool is the most important pile fibre, and cotton is most often used as the base and binder material.

Ghiordes and Sehna knots

Knots. The pile surface of knotted rugs is formed entirely by the ends of knotted tufts (see Figure 7). The Ghiordes, or Turkish, knot brings both tuft ends to the surface together between two warp yarns. It is common in the Near East, especially in Turkey and the Caucasus. The Sehna, or Persian, knot brings each end of the tuft to the surface separately. It predominates in Central Asia and the Far East, mainly in Afghanistan, India, Pakistan, Turkestan, and China. In Iran either knot is used, depending upon the origin or site of the tribe or town producing the rug.

From I. Wingate, *Textile Fibers and Their Selection*, 5th ed. Prentice-Hall Inc. © 1964

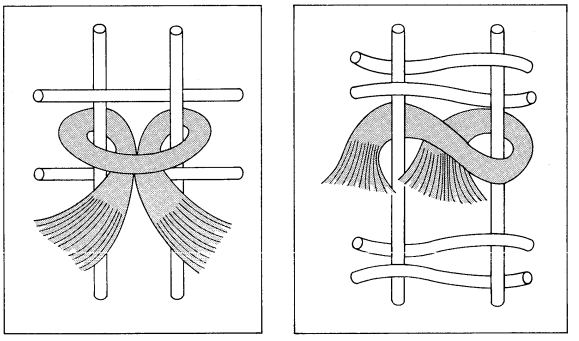


Figure 7: (Left) Ghiordes knot. (Right) Sehna knot.

Looms. The loom employed is upright, consisting of two strong beams connected by two vertical posts to make a steady frame. It is often adjustable for the weaving of different sized carpets and rugs. The weaver is positioned conveniently in relation to the row of knots being worked either by means of a seat that can be raised, moving him upwards, or by winding each completed row of knots onto a separate cloth beam. The warp, or lengthwise, threads stretched between the two beams are evenly spaced and regularly spun, assuring that the pile forming beneath the surface will also be even.

Weaving. The weaver ties his rows of knots forming the pattern, and when an entire row of pile is knotted, the two, three, or four weft, or crosswise, threads are forced down by a comb or knife, causing the pile to stand out. Density of pile is about 300 knots to the square inch and a weaver completes about 8,000 per day; several weeks' work is needed to produce an ordinary carpet, and possibly months for a more complex one. The weaving instructions required to produce the desired pattern may be chanted by a Salim or may be provided on a coloured chart of squared paper.

Dyestuffs. In olden days, craftsmen used natural dyestuffs, obtaining reds from the roots of the madder plant; carmine red from cochineal, the bodies of the female *Coccus cacti;* reddish-browns from ox blood; yellow from the reseda plant or from saffron crocus, vine leaves, and pomegranate skins; and blue from the indigo plant. Mixtures of certain blues and yellows produced greens; and natural wool shades produced greys and brown, although nutshells and bark were also used. Oak apples were often used to produce black, but if their iron oxide content was high the wool was likely to be damaged; some old carpets today show the most wear in the black portions. Modern synthetic dyestuffs allow greater flexibility than these traditional dyes.

Characteristics. Persian rugs have intricate all-over patterns, mainly floral, but sometimes including animal or human figures, often with a central medallion. Colours include soft pastels and muted reds, browns, and blues. The rugs are fringed at both ends.

Turkoman rugs are woven in geometric designs, employing vivid reds, browns, and greens, and usually have webbed fringes at the ends. Caucasian rugs have sharply outlined, bold, geometric patterns. Brilliant and strongly contrasting colours are employed, frequently including reds, yellows, and blues. Turkish rug patterns have precise, stylized geometric or floral designs, with bright, sharp, contrasting colours. Indian rugs are made with botanical designs in a naturalistic style and are brilliant in colour. Chinese rug designs include religious symbols. Designs are usually in blue, and background colours include dulled yellows, browns, and roses.

Other handmade carpets and rugs. *Tapestry weave.* Another kind of hand weaving is the tapestry method, wherein the coloured weft threads, wound upon wooden needles, are threaded around and between the warp ends, leaving a flat or slightly ribbed surface. Since a tapestry carpet lacks a tufted pile, it does not have a luxurious texture, even though a fine pitch, the number of warps per inch, can be employed, and the richest and most delicate effects of design and colour obtained. Carpets of this type have long been made at Les Gobelins, Paris, Aubusson, and Beauvais in France, and Tournai in Belgium. The work involved in producing tapestry carpet is slow and requires great skill; the product therefore is expensive.

Hooked rugs. The origin of the hooked rug is obscure. A rug of this type is made by pulling narrow strips of wool or cotton cloth or wool yarn, with a tool roughly resembling a buttonhook, up through a basic material of coarse linen or burlap. The loops, approximately ½ inch high and the width of from two to four of the mesh openings in the basic material, are often clipped. Frames of various designs hold the basic material taut. Frequently the rug is made by pushing, instead of pulling, the loop through the basic material, usually employing a large threaded needle.

Modern handweaving. It is possible to weave rugs, with or without a pile surface, on hand looms. Almost any material can be used for hand-weaving, including plastic strips and some forms of rope. Weaving speed can be increased by the use of a chain-and-pedal drive similar to that used on bicycles. In modern times, handweaving is mainly limited to design development and study in educational institutions and to the most costly floor coverings, frequently made to order and considered fine works of art.

MACHINE-MADE CARPETS AND RUGS

Early machine processes employed hand or water power to duplicate processes originally performed completely by hand. The invention of the power loom greatly reduced the amount of time and labour, and therefore the cost, of carpet manufacture.

Nomenclature and types. Machine-made rugs and carpets take their names from the looms employed—Wilton, for example—or the construction method, such as ingrain or Brussels.

Ingrain. After 200 years of use, the ingrain became almost obsolete during the 1930s because of its relatively poor wearing qualities. Also called Kidderminster or Scotch, it is flat ribbed, reversible, and completely without pile, and usually could be made with two- or three-ply warp and weft yarns, dyed before weaving.

Kidderminster

Design on the surface of the carpet was easily produced by bringing to the surface, at the chosen area, the desired colour of the weft; the latter almost completely formed the surface colouring. Although the warp intruded into the design, its colour was carefully chosen to blend satisfactorily. The term Venetian has been applied to lesser grades of ingrain carpet.

Wire-formed piles. Brussels carpet, with uncut looped pile formed by round wires, was first introduced about the mid-18th century. Shortly after, Wilton followed with the development of bladed wires producing a cut pile. Both

were capable of producing elaborate pile designs, and the Jacquard device was used with them from about the 1820s, when it replaced hand selection of coloured pile threads. Bigelow's power principles were applied to the loom from about 1849. Pile not used for surface design is woven longitudinally into the body of the backing fabric. Brussels carpets, woven on the Wilton loom, but with their pile remaining uncut, are no longer sold in large quantities.

Tapestry is a printed pile yarn carpet that may be cut or uncut but must, by its nature, normally be one pile thread per dent. In order to make a design by a method other than printing yarn, more than one pile yarn per dent (five or six usually) is used with a jacquard. The term velvet, more common in the United States, is used to mean cut pile single frame; that is, plain carpet for which tapestry technique is not used. A velvet carpet is basically one of Wilton type. The term frame is often used instead of pile thread per dent, and a single frame carpet is therefore a plain carpet with one thread per dent.

Tapestry carpets were invented in Scotland in 1830 but are now virtually obsolete. Few of them were produced in the main carpet centres (Kidderminster, West Riding of Yorkshire, and Scotland) after World War II.

Typical high quality Wilton and Brussels carpets contain about 120 tufts per square inch with up to 13 rows per inch. Tapestry, generally up to about 80 tufts per square inch, but as low as 50, is possible. Velvet may be similar in density to tapestry, but not necessarily so; it can be made with pile as dense as Wilton if desired.

Wilton carpet can be produced on the "face-to-face" principle. Two carpets are simultaneously woven sandwiched one above the other, and the two are then cut apart by a reciprocating blade traversing the loom width. The linear rate of production is thus almost doubled; Jacquard designs can be produced by this method.

Axminster carpets

Loom-formed pile. Axminster carpets, in which all of the pile yarn is effectively used for design (unlike Wilton and Brussels that waste some "dead" pile yarn by hiding it in the body of the carpet) include spool, gripper, and chenille.

Spool looms were invented in the U.S. in 1876, and the gripper Axminster loom was developed about 1890. The chenille two-stage process was invented in Glasgow about 1830.

With the first loom, each row of pile is drawn from an individual spool, and two blades cut away the tufts when woven. On the gripper loom, each tuft is held by its beak-like gripper and taken from its yarn carrier to the fell of the carpet, the point at which the warp and weft intersect, after being precisely cut away by a traversing knife blade. One type of spool-gripper Axminster loom employs spools instead of a jacquard; the tufts are taken from them and woven on the gripper principle. Chenille pile (from the French word for caterpillar) is formed on the carpet loom by weaving the "fur," or pile yarn, as a weft. The tufts are usually bound by cotton threads forming a long strand. The fur is woven in the first weaving process on normal cloth looms and cut longitudinally into the requisite patterned strips; pile yarn is woven as weft, and the warp is the cotton binding threads. A simple gauze or the lacelike leno weave is used to bind the weft pile yarn so that the pile does not fall away when the strips are cut, before they can be woven into the carpet.

Jute weft and cotton warp are common materials for Axminster carpets. All-wool pile is popular, although nylon–wool mixtures and various combinations of natural and man-made fibres are becoming common. Carpets made of 100 percent man-made fibres are increasing in popularity. Spool and gripper qualities average about 40 to 50 tufts per square inch, the two extremes being about 30 and 60.

Unconventional carpets: tufted, knitted, and bonded. In tufted carpets, the tufts are inserted by means of vertically reciprocating needles pushing through prewoven backing and are held below the carpet backing by loopers. The loop pile slips off the loopers, or is severed when cut pile is desired, as the formed carpet moves away from the looping elements. The rate of pile yarn feed from the creel governs pile height. The variation of this rate for groups of pile threads is the basis for certain patterned effects resulting from the inclusion of both high and low pile areas or achieved by alternating two different coloured threads across the full width, emphasizing the pattern in one colour by hiding the other low-pile colour with higher adjacent loops. Different rates of feed can be obtained by electrical or mechanical means. In the former, the pattern, in small scale, is placed on a pattern drum, and feelers touching the pattern areas change the delivery rate of the pile feed roller. Photoelectric means of influencing the delivery rate have also been devised. The mechanical device pushes down the required pile lengths by means of a castellated metal plate cut to the required pattern. Other effects can be obtained from fancy or printed yarns, and some tufting machines reciprocate the carpet widthwise to obtain a longitudinal zigzag effect. Designs including up to eight colours are possible, with quality almost as satisfactory as in traditional carpets having an equal number of colours.

Quality ranges now vary as much as for traditional carpets, since needle spacings may be as close as 12 to the inch, giving the same number of tufts per square inch as in Wilton. Traditional pile materials are employed, but polypropylene, a synthetic fibre, is commonly used on prewoven backing. There has also been some development of nonwoven backings.

Polypropylene backing

Production rates can be many times greater than that of Axminster looms; one machinery manufacturer has developed a yarn looping technique whereby the backing is pierced by a needle and the pile is then blown through the resulting opening. This method increases the rate of production up to about 1,200 rows of pile per minute. Patterned carpets have been produced mainly in the United States with either cut or loop pile, or a combination of both. In warp knitting, beams are used to feed in the yarn supply; in weft knitting, the yarns are fed from smaller packages.

A Czechoslovakian Arachne stitch-bonding machine achieves high production rates with low pile costs, employing a fibrous web stitched on the knitting principle with yarns drawn from beams. An East German Malipol machine uses knitting principles to bind pile to a backing fabric, although a later model uses unknitted weft threads instead of backing. Production rates for knitting are higher than for traditional looms, but do not reach the high speeds of modern tufting. The quality of the materials used and the finished carpet are much like that of the middle range of traditional carpets. No spectacular growth, however, in the sales of such carpets has occurred, since there are currently limitations in design, although development work is progressing. Several methods use the principle of bonding fibres, fibrous webs, or yarns with various adhesives. The technique imposes design limitations, but the less complicated pile forming method results in high production rates. The needle loom principle involves a fibrous lap attached to a base fabric and another method forms a lap into a loop pile sheet and sticks it to a base fabric. A special finishing operation can convert the loops into cut pile.

A German device projects positively charged fibres toward a negative pole; the fibres are then stuck to an adhesive-coated backing fabric. A Czechoslovakian technique makes double, or face-to-face, carpet by looping and bonding a fibrous web to two backing fabrics; the full layer is then cut into two separate carpets. A French machine cuts fibres from a sheet and then binds them to a fabric backing, with electrostatic means used to keep the fibres vertical.

Manufacturing methods. *Raw material preparation.* The warp and weft backing structures of carpets, and the pile yarn superimposed upon it, made originally from natural fibres, are now mostly synthetic. Backing yarns are frequently made from the lower priced cut-staple normal or high-tenacity viscose rayons. Selvages of synthetic filaments and pile yarns are produced from mixtures of cut-staple man-made and natural fibres, the latter primarily wool. Pile yarns composed of the more expensive synthetic materials are increasingly used, and viscose rayon has been used extensively for less costly carpets. As the use of man-made filament pile yarns has increased, tra-

ditional processing of natural and staple man-made fibres has decreased in importance.

Preparation of natural fibres

The preparation of natural fibres for backing and pile yarns involved many labour-consuming operations. Fibres had to be cleansed, often by mechanical means and a wet treatment. Pile-yarn wools had to be thoroughly scoured, and cotton fibres also needed special cleaning treatment. Straightening of fibres prior to spinning was a long-drawn-out process, with spinning often followed by additional single twisting operations to form ply or cable yarns. In contrast, man-made filament yarns are extruded in clean, continuous, and parallel form, and the twisting operation is uncomplicated. Backing fabrics are currently made from fibrilated yarns, consisting of small fibres twisted together, and sliced narrow tapes are made from sheets of synthetic film.

Selection and preparation of design. Design creation or selection involves consideration of the range or limitations of the various methods of carpet manufacture. The number of colours that can be used for Jacquard Wilton and gripper Axminster are limited; spool and chenille Axminster allow unlimited colour range. Density tends to be greatest for Wilton carpet, sometimes reaching as high as 130 per square inch.

Spool Axminster designs are made from a chain of spools, with one spool to each row of pile. There are usually seven threads to the inch across the width, with a pile yarn length per thread of about 35 feet (11 metres). Spool width varies according to the loom, the smallest is usually 18 inches (46 centimetres) wide. The person operating the machine creels all necessary colours for the tufts in a row and winds the full spool, repeating the process until the full set for the design is made.

Gripper Axminster and Wilton Jacquard looms utilize a perforated card system to select the required combination of different coloured threads per row of pile. There is one card per row, punched according to the predetermined design, and the possible colour choice for each tuft is usually one in five for Wilton and one in eight for gripper Axminster. "Planting" of several additional colours in the pile-yarn creel may increase the number of different colours possible in a design. High- and low-cut and uncut effects are also possible in Wilton carpets. Normally employed on tufted carpets, the process can be applied to all types. Different pile height effects are produced in tufted carpets by using different rates of yarn feed, and are sometimes emphasized by alternating the colours and hiding the colour of low pile under adjacent high pile. Mechanical and photoelectric devices are used to regulate the rate of yarn feed to the tufting needles.

Construction. The basic structure of traditionally manufactured carpets consists of the backing fabric and the pile, which is bound under and between the weft. Pile is formed on Axminster looms by inserting weft, with reciprocating needles, across the warp chain yarns. A central stuffer warp usually separates top from bottom weft. In Jacquard Wilton and Brussels carpet, the pile not used for surface design effect is concealed between top and bottom weft; Axminster carpet uses all pile yarn for surface effect. Wilton pile is formed by looping the yarn over wires that project it mechanically across the width of the loom and beneath the decorative pile yarns. When the wires are withdrawn, they either cut the pile, forming Wilton, or leave it uncut to form Brussels carpet.

In spool Axminster carpets each row of tufts is inserted and cut away from the spools. On gripper Axminster looms each tuft is inserted by its individual beaklike gripper, after being cut away from the carrier. Backing fabrics have warps held on flanged beams; in Wilton weaving, weft threads wound on cops are shuttled across the loom. Axminster weft is drawn from large stationary cones at the side of the loom.

Tufted carpets

Tufted carpets differ basically from traditional types because they have a prewoven backing into which the tufts are inserted. One unconventional method of making carpet involves the forming of pile on knitted structures. Another method involves the application of pile yarns or even undulating webs of fibres to backing fabric by means of adhesives. These methods often lack the means for controlling design.

Finishing. After weaving, carpeting may require machine brushing to remove loose fibre or yarn, before any correction of processing faults. Traditional carpets have a separate reparation process, but on tufted carpets, missing pile, the most common fault, is stitched in by a portable "gun" operating a reciprocating needle. The carpet pile is then sheared to uniform level with cutting blades similar to those of a wide lawn mower. Steaming of the pile causes it to expand or "burst" into an aesthetically enhanced state. Natural back-sizings were formerly applied to Axminster carpets, and Wilton and Brussels weft threads on their cops were soaked in sizing. Increasing use of such synthetic backing compounds as polyvinylacetate and different kinds of lattice backings now produces excellent tuft bind and stiffness.

Other modern finishing processes include mothproofing of wool, often performed in the dyeing sequence, and application of soil-resistant finishes to man-made pile, rot-proofing, and antistatic treatment.

Reinforcement or screening, particularly used for tufted backing, may employ adhesives on open structure woven materials, and the use of such cushioning materials as synthetic rubber foam is increasing.

SMOOTH-SURFACED FLOOR COVERINGS

Linoleum. Linoleum is produced by pressing a sheet containing oxidized linseed oil, gums and resins, ground cork or wood flour, and pigments on to a backing, and it is hung in huge "curing" stoves to toughen.

Two general types, plain and printed, or inlaid, linoleum, are produced; the latter has a constant pattern throughout its thickness. Different methods are employed to create various design effects. Thickness ranges from about 1.6 to 4.5 millimetres, depending upon the traffic expected; a standard width is six feet (two metres), and the weight in 3.2 millimetres gauge is about 7.6 pounds per square yard (3.5 kilograms per square metre).

Recently developed products are specially hardened to resist indentation. Certain alkalies and organic solvents, however, will attack linoleum if left in prolonged contact; staining and discoloration can arise from certain antioxidants in tires, rubber castors, or rubber heels. Linoleum is resilient, warm, unaffected by reasonable floor temperatures, and does not readily support combustion.

Printed felt base. Printed felt base is formed by applying a heavy film of paint to felt saturated with asphalt; the felt is sealed at both the top and bottom with one or more layers of coating before application of paint, preventing discoloration from the paint and leveling the surface. The paint used has low volatility and little flow, dries quickly in thick layers, and gives high gloss with good wearing properties. Wooden blocks are used to apply the paint on the base, with several colours being printed separately. The material is next dried in an oven, with the product hanging from racks or festooned, and this maturing process requires several days. Rugs or rolls are available in several thicknesses and sizes.

Asphalt tile. The tiles are made from asphalts (25 percent) or synthetic resins, asbestos fibres (25 percent), pigments, and mineral fillers (50 percent). If asphaltic binder is used, colour is restricted to black, brown, and dark reds. The plasticised resin-based tiles are much lighter in colour, including blues, reds, yellows, and greens, and some recently developed tiles have a small amount of vinyl binder, giving brighter, clearer colours and improved grease resistance. The ingredients are mixed at relatively high controlled temperatures to form a dough, that is then forced through successive calenders until the required thickness and finish is obtained. The sheet is die cut when cool, ensuring dimensional stability. Common sizes are nine inches (22 centimetres) by nine inches by $\frac{1}{8}$ inch or $\frac{1}{10}$ inch (three or two millimetres). The tiles are resistant to fungal attack, mild acids and alkalies, and oils and grease; and they are suitable for use where floor heat is less than 80° F (27° C). Sweeping and washing with warm water and soap or detergent keeps them clean; wax or resin emulsion polishes improve their appearance and prolong life.

Asbestos tiles

Vinyl floor coverings. Vinyl asbestos tiles were developed from asphalt tiles. Asbestos fibres, mineral fillers, and pigments are all bonded together with copolymers of vinyl chloride, or with vinyl acetate. Vinyl resin binders have greater flexibility without requiring heat treatment prior to installation.

Vinyl sheet coverings are made by calendering from various combinations of vinyl chloride resins, pigments, stabilizers, and fillers. Tiles 9 inches, 12 inches, or 18 inches (23, 30, or 46 centimetres) square can be cut from these sheets; widths vary from 3 to 6 feet (1 to 2 metres). Sheets can be unbacked or backed with felt, hessian, plastic foam, cork, or fabric. One method adds a coat of vinyl to a backing, and the surface appearance is influenced by the nature of the backing. Jointless seam welding can usually be applied to vinyl sheeting. A blowing agent can be added to vinyl paste, producing a foamed back and fully bonded vinyl floor covering. A thin layer of the paste and agent is spread on the back of a premanufactured vinyl sheet; these are passed through a heated tunnel causing the release of a gas from the agent, forming the foam.

An "all through" flushing in the mix can be used for secondary colour effects. Some laminated sheets have rotogravure pattern printed with vinyl inks and covered by a protective layer of transparent vinyl. Backed vinyl sheets can have added coloured chips, with a final clear coating producing a three-dimensional effect. Embossing allows the addition of a second colour into the resultant depressions, producing an inlaid effect.

Rubber floorings. Production figures for this material are comparatively small. Previously made with natural rubber, it is now produced with synthetic rubber mixed with pigments, fillers, resins and curing materials, and when sheeted is heat cured. Mottled effects are available, and thicknesses around two- and three-sixteenths of an inch (three to four millimetres) are available in tiles up to one foot square or in rolls. It is a comparatively expensive, excellent wearing flooring, although somewhat susceptible to grease damage, and is appropriate for use in public buildings where traffic is heavy.

Cork tiles and carpet. When ground cork is heated, either for long periods or by rapid high-frequency heating, the granules adhere, creating a resilient mixture that can be formed into tiles, usually two- to five-sixteenths of an inch thick, and six or nine inches square. This tile has a comfortable resilient feel and absorbs sound, but appearance and wear properties are only fair.

Cork carpet, although limited in use, can be produced from graded cork granules and polymerized linseed oil, homogeneously pigmented and calendered on a jute canvas backing.

Epoxy resins. Flooring compositions based on epoxy resins have developed steadily, giving a hard, chemical-resistant, seamless, and firmly adherent floor covering. The resin and curing agent must be blended immediately before use; colours and fillers can be added. The comparatively high cost of epoxy-resin systems restricts them principally to repairing or surfacing existing flooring substrates; *e.g.,* concrete. They frequently employ composition based on a cement aggregate and natural rubber latex. Polyvinyl acetate may be substituted for rubber latex, and such resins as acrylic, polystyrene, and styrene-butadiene have also been employed.

ECONOMIC ASPECTS

Modern research and development have resulted in improved equipment, processes, and materials contributing both to increased production and greater product durability. Although per capita use has increased, greater competition exists. Larger companies usually have larger facilities for development and expansion than smaller companies and are more likely to survive any large decrease in profit margins. The heavy capital expenditure required for increasingly sophisticated processes, and the equipment they require, may accelerate the trend in the developed countries toward larger groups of floor-covering manufacturers and their suppliers.

(P.E.)

Furniture industry

The modern manufacture of furniture, as distinct from its design, is a major mass-production industry in Europe, the U.S., and other advanced regions. It is very largely a 20th-century industry, its development having awaited the growth of a mass consumer market as well as the development of the mass-production technique. Earlier furniture making was a handicraft, going back to the most ancient civilizations.

HISTORY

Examples of ancient furniture are extremely rare, but there is considerable knowledge of the pieces made by craftsmen in China, India, Egypt, Mesopotamia, Greece, and Rome from pictorial representations. Beds, tables, chairs, boxes, stools, chests, and other pieces were nearly always made of natural wood, though veneering was known in Egypt, where it was used to produce coffin cases of great durability. The Romans too used veneers, though chiefly for decorative purposes. Bronze was also used in Roman tables, stools, and couch frames. Pompeian wall paintings show that plain, undecorated wooden tables and benches were standard in kitchens and workshops and that panelled cupboards were common. Chests for valuables were covered with plates or bound with iron.

The early Middle Ages were much poorer in household furnishings of every kind than the Roman world, but in the 14th and 15th century a growing affluence brought a major revival of furniture making, with many new types of cupboards, boxes with compartments, and various sorts of desks appearing. The religious houses in particular were well supplied with furniture. Framed panelling, reintroduced in the Burgundian Netherlands, quickly spread. The mortise and tenon and mitre provided greatly improved joints.

Introduction of veneering

The growing sophistication in technique brought a revolutionary change in the men who made furniture. Where previously carpenters and joiners had made furniture along with every kind of building construction in wood, several circumstances combined to create a new profession: that of cabinetmaker. The most important technical factor was the introduction, or reintroduction, of veneering, first in western Europe, then in Britain, North America, and elsewhere.

In the earlier system of framework and panel, the framing gave the required strength in both length and width, the panel being a mere filling held in grooves. Its attractive appearance was the result of highlights and shadows produced by the framing, moldings, and carving, which formed the chief means of decoration. The grain of the wood was incidental.

The introduction of veneering coincided with the use of walnut as a furniture wood. It was soon realized that the grain of such a wood could be of decorative value, especially as veneering made it possible to use such visually attractive parts of the wood as burrs, butts, and curls, unreliable if used as solid wood. It became the custom to have the grain of the veneer generally run crosswise because of its decorative appearance. Marquetry (a form of inlay in veneer) was another example of the decorative use of the grain and colour of wood in surfaces unbroken by panelling.

In addition to veneering and the new system of construction it involved, an impetus to the establishment of the trade of furniture making came from the increasing market demand provided by the growing affluence of the 17th and 18th centuries. In the new system of construction, plain, flat parts are dovetailed together and then veneered. It can be contrasted with the traditional framed method of rails and stiles put together with mortise and tenon joints, the panels fitting in grooves.

Coinciding with this change, or preceding it by a few years, was another breakaway: that of the chairmaker, who had become another specialized craftsman. At first chairmaking was closely associated with wood turning but by the 18th century turned legs were largely replaced by shaped legs of the cabriole type. Chairmaking has remained a separate branch of furniture making ever since.

This growth of cabinetmaking as a trade of its own eventually resulted in a considerable degree of standardization of methods of construction, particularly in the types of joints used and in the thicknesses of wood for the various parts. It also resulted in an increased division of labour. Turnery became a separate trade, while the cabinetmaker assembled the turned parts; veneer and marquetry cutting was not done by the cabinetmaker although he laid both; carving too called for the skill and experience and tools of a craftsman who did nothing else. Another specialist, the upholsterer, did his work after the chairmaker had made the frame; and it seems likely that finishing was seldom done by the cabinetmaker. This was certainly the case later in the 19th century when French polishing became the standard method of finishing furniture.

Chippendale, Hepplewhite, Sheraton

An important 19th-century change was the separation within the industry of those who made furniture from those who sold it. Previously the customer commissioned a cabinetmaker, perhaps after consulting a design book by Chippendale, Hepplewhite, or Sheraton. Or he might work out his requirements in consultation with the cabinetmaker or, if he were sufficiently wealthy, employ an architect or designer. After the midyears of the 19th century the showroom gained popularity. A large store often retained its own workshops where special items were made to customers' requirements, but for the greater part it became the practice to buy wholesale from furniture making firms.

MODERN FURNITURE MANUFACTURING

Materials. Modern methods of furniture construction are largely based on the availability of man-made materials such as reliable plywood, laminated board, chipboard, and hardboard as distinct from natural solid wood. It is not merely that manufacturers prefer the one to the other but rather that these substances are free from the great drawback fundamental to wood—movement. Natural wood shrinks as it dries or swells as it absorbs moisture from an atmosphere more humid than itself, and this movement must be allowed for in the method of construction. Unless this is done troubles may arise: splits along the grain or open joints on the one hand or jammed drawers or doors on the other. Over the years cabinetmakers have worked out ingenious systems to avoid these troubles in the use of solid wood, but today made-up materials may be regarded as inert if of good quality. To an extent solid wood has still to be used, notably for items that have to be turned, cut to shape, or molded, and for lippings to conceal the edges of manufactured boards; but virtually everything in the form of flat panels is made up.

Natural wood. The increase in the demand for reasonably priced furniture has placed a premium on the economical use of wood. Natural wood is extremely wasteful as a material. Hardly more than 25 percent of the natural substance of a tree actually goes into the furniture made of solid wood. When account is taken of the loss in sawdust in conversion from the tree trunk (taking off the outer slab portions and sapwood) and the further loss in bringing the lumber to usable size in the workshop (the offcuts, waste in sawing shapes, in turning, in planing, cutting joints, and final cleaning up), it becomes evident that much more wood is wasted than used.

Plywood. In making plywood, the veneers are peeled rotary fashion from the log by a long knife fitted to a lathelike machine. The resulting veneer can be of unlimited width to be cut up as required. There is no loss in sawdust, and the peeling is continued until only a polelike centre is left. Much the same applies to laminated board in which both the core material and the outer plies are peeled. In the case of chipboard the timber is merely regarded as raw material to be reduced to fine chips that are dried, compressed, and assembled into boards, with resin glue as an adhesive. Where a natural wood grain is desired, a veneer is flat sliced from a flitch (longitudinal section) selected for the beauty of its grain.

Certain materials, notably chipboard, must be machined, because trimming at the edges by hand almost always shows as a deterioration. It cannot be planed; the plane merely forms dust rather than taking shavings and, owing to the abrasive nature of the material, the edge of the cutter is quickly lost. Consequently, when a panel of a certain size is required, it needs to be machine sawed to size, no further trimming being needed. This is only practicable with a precision saw capable of fine adjustment. Furthermore it requires a saw blade having tungsten teeth to resist abrasion. The same applies to any plywood or laminated board assembled with resin glue.

Another influence on the construction of furniture is the introduction of new types of adhesives in place of the traditional animal glue. Many are highly water resistant, some waterproof. Some can be applied cold, avoiding the complication of heating joints before assembly. They can be cured by heat in a matter of minutes, leaving presses and other apparatus free for other work.

Other materials. Although wood has always been regarded as the traditional material for furniture making, several other materials are now used, either entirely replacing wood or combined with it. Plastic laminate, widely used for table and other tops, is obtainable in various colours and designs and in photographically reproduced natural wood grain. Its advantages are that it resists all liquid stains, is largely heat proof against burn marks, is mark free, and is easily wiped clean. It is laid as a form of veneer on any of the man-made materials—multiply, laminated board, or chipboard, usually with a contact adhesive. As a plastic edging is needed that must be applied before the main top is put down, an essential machine tool is the portable router with veneer-trimming unit. It trims the overlapping edges of the main plastic panel without cutting into the edging.

Plastic laminate

Metal is also used to some extent, particularly for the stands and legs of furniture. Iron is generally preferred, the parts joined by welding.

Finishes too have been revolutionized. French polish, the traditional finish of the Victorian period, and indeed up to the 1930s, has been largely replaced by gloss or eggshell lacquers, which are sprayed on and are heat and water resistant and are so hard as to be practically mark free.

Storage and transport. Two technologies important to furniture making are storage and transport. The space taken up by furniture in relation to the actual material used in its construction is disproportionately large; when furniture is mass-produced an enormous amount of storage space is required. This applies equally to its transport, especially when it has to be shipped abroad. Consequently a great deal of furniture is made of the "knockdown" type; that is, it can be taken to pieces and stacked flat. A wardrobe made in this way may occupy only a quarter of its assembled space when disassembled. Originally, parts were joined by screw fastenings, but a whole range of fittings has been devised to achieve the same result more easily and with more precision. Most such fittings require little more than recessing or the boring of holes, operations easily machined. Most work on cam, screw, or wedge action.

Woodworking machinery. The decline of the direct link between customer and maker, due to the rapid development of retail trade, was largely made possible by the invention of several woodworking machines, mostly steam powered. Much handwork remained, however, and only large manufacturers could afford major machinery installation. In the early 20th century it was still possible for a cabinetmaker in Britain or Europe to earn a living, though in most cases he installed a basic machine such as a circular saw or worked in a district in which machine shops were available. Thus in Shoreditch, London, whole streets of houses were occupied by cabinetmakers, often several in one house, who made pieces that varied from the finest individual items to the cheapest, turned out in pairs or perhaps six at a time. These men had their machining done in the trade machine shops that abounded in the district. The shops produced nothing themselves but performed any machining that was brought to them: sawing, spindle molding, fretting, turning, planing, and so on. These practices continued up to the beginning of World War I and for a time afterward, although most of the large stores also had their workshops where they made

The Shoreditch shop system

not only individual items for customers but also furniture in quantity to pattern.

Grand Rapids industry

But in the U.S., the development of mass-production furniture manufacture was already well advanced, with the principal manufacturing centres at Grand Rapids, Michigan; Jamestown, New York; and High Point, North Carolina. "Grand Rapids" became a byword for inexpensive furniture of reliable quality. Furniture factories have never become large in comparison with the huge production units in such industries as automobiles and steel—few today employ more than 100 persons—because of the continuing need for some hand operations. But their machines for many purposes and the volume in which they operated gave them insuperable advantages in cost over the old-fashioned craftsman. Mass-produced furniture began to have a serious impact in Britain and Europe between the wars.

The shortage of timber during and after World War II made conditions extremely difficult for the furniture maker; but in the 1950s there was a gradual return to more decorative furniture, marked by the introduction of new materials, new machines, adhesives, and finishes.

Modern commercial furniture production may be roughly divided into groups: general furniture—bookcases, wardrobes, tables, etc.; chairs and upholstered suites; and specialized items. Each of these may be further subdivided according to quality and type. In addition to this commercial furniture there are the specialized items made by a few hand craftsmen to special commission. Such goods are necessarily expensive, partly because they are individual pieces made singly to design and also because the best selected materials are used. Furthermore, hand methods are largely used that are costly because they are time-consuming. Even in this field, however, the machine has encroached to an extent. Thus a circular saw is invariably installed because its advantages are so obvious. There is no merit in laboriously ripping boards to size when a machine will do the work as well or better.

Modern hand craftsmen

Though furniture produced by modern hand craftsmen is beautifully made from the best materials, it often requires considerable discernment to detect the difference between it and the best commercial furniture.

Period furniture. In a class by itself is the manufacture of reproduction period furniture. The best work in this field is of an extremely high standard; and, although it often has to make concession to modern materials in using veneered plywood or laminated board for parts, it usually follows traditional methods of construction, at least where visible machine work would be obvious. On the other hand, all veneers are put down in a press, moldings worked on the spindle molder, and shapes cut on the bandsaw or jigsaw.

It is in this work that wood carving is chiefly used. Because of its high cost, carving has largely disappeared from modern commercial furniture, but to the manufacturer of reproduction furniture it is an obvious necessity. From early times and certainly from the 17th century, wood carving has been a separate trade. A highly skilled calling, it demands artistic sense as well as manual dexterity. It has become divided into classes of which furniture and indoor decoration represent only one branch, with further subdivision within the branch.

In the commercial grade of furniture there is wide variation in quality, from the lowest priced goods to high-grade items in which individual hand workmanship is used for processes where the quality would suffer if the machine were used. Thus drawer dovetails are cut by hand, and sometimes even hand-cut joints are used.

Modern factory layout. Most modern factories are laid out on mass-production lines. The earlier factories often had a cabinet shop, which had its rows of benches for individual work; a cabinetmaker needing machining done carried his wood to the machinist. Today the timber is cut to usable sizes in a main conversion shop and brought to the required moisture content in a kilning section. In the kiln, air is forced through stripped stacks by fans that periodically change the direction of the air flow. In recent years radio-frequency heating has been widely used to dry both natural wood and plywood. The applied radio frequency produces molecular activity in wood and resins (such as those in plywood glue); part of the molecular energy is converted into heat that greatly reduces the time required to dry the wood thoroughly and evenly and to set the glue. The wood is placed in a press between two metal plates to which the power is applied; great thicknesses of wood can be dried evenly by this method.

From the drying section the wood proceeds to the planing and jointing shop, in which it is reduced to the required section and any tenoning, dowelling, or dovetailing carried out. There is also a veneer department, and in many respects this has become one of the most important departments. In it veneers are jointed in width where necessary, and a remarkable recent invention is a machine that sews veneers together with fine fibreglass, the stitching passing through half the thickness of the veneer only. It does this with amazing speed and accuracy. Where required, veneers are matched, giving a balanced appearance; and any small defects are repaired by placing a waste veneer beneath, cutting through both simultaneously, and interchanging the cutout pieces. Veneer pressing follows, and, although multiplate presses are still used to an extent, the tendency in large-production work is toward the progressive presser. At one end of this the resin glue is applied with a spreader, the veneers placed in position on the groundwork, and the whole passed in batches beneath the presser where it is heat cured in about a minute and ejected at the other end ready for further operations.

It is in the assembly shop that the line or conveyor-belt system begins. This is not usually in continuous movement but takes the form of a series of loose rollers over which the work can easily be pushed by one man. Special cramping jigs are set up so that, for example, a wardrobe can be glued up in one operation by power-driven rams. The jig ensures squareness, and the resin glue is cured in a matter of seconds by radio-frequency heating. In fact, by the time the operator has applied glue to the joints of one set of parts, the previous assembly has hardened and can be removed to the conveyor, leaving the jig free for another cramping operation.

Sanding, finishing, and fitting

From this point onward the work remains on the conveyor belt, passing to a sanding shop where joints are levelled and finally to the finishing shop where it is stained, spray polished, and fine sanded and waxed. Lastly there is a fitters' shop, where doors are hinged, handles put on, mirrors fitted, and so on.

Before passing to stock or to the packers' department every piece has to be passed by an examiner who chalks any defect or attaches a small, coloured label indicating that there is a fault in either the woodwork or the finish.

Only a constant flow of orders in large quantities justifies such a setup. Smaller firms contract their veneering to outside specialists, have their turning done outside, and also any other work for which they are not equipped. The aim is to maintain a constant flow of production, which involves a balance in personnel in the various departments to avoid a holdup in any stage.

The production process. A basic preliminary in all furniture production is the provision of working drawings. In a firm of any size there is invariably a special department where full-size drawings are prepared from small-scale drawings provided by the designer. In some cases the designer may make his own full-size, detailed drawings; but in a large firm it is more usual for a draftsman to work out the practical details, though usually in consultation with the designer, who advises on proportions and decorative details. The hand craftsman, in contrast, usually does the whole thing himself. In the small-scale drawing the general form and essential requirements are worked out; the full-size drawing shows proportions and constructional details. A sample piece is made to check the design and cutting problems. Cutting lists are prepared; the cost of materials, fittings, finish, etc. figured; and an estimate of machining and assembly time worked out. When the work is to be produced in quantity, costs are lowered considerably because only one setting of the machine and only one set of cutters are needed for the whole run of any particular part.

Selection of timber, already passed through the seasoning kiln and converted to standard thicknesses, follows.

The wood passes to the machine shop, where it is sawed to size, planed, molded, grooved, or rebated as required. When a number of parts must be cut exactly alike, they are clamped in forms having the proper contour and are then brought in contact with high-speed rotating knives that shape the part to proper size as the form rides against a guide on hand or automatic shapers and routers. Intricately carved pieces such as legs are roughly carved on multiple-spindle carving machines. These duplicate a master leg by means of a follower point that is guided along the surface of the model and imparts the same motions to as many as 32 high-speed rotating knives as they whittle the leg blanks. After the rough carving, the pieces are machine sanded and finished by a hand carver.

If veneering is required, this is now done. Jointing follows—tenoning, dowelling, dovetailing, etc. Automatic machines often combine several operations. Exposed parts are sanded on edge belt sanders, three-drum travelling-bed sanders, or belt sanders. Rounded parts are sanded on soft pneumatic drums, and carved parts are sanded on a buffer, a machine in which shredded sandpaper is supported by brushes on a revolving wheel.

Furniture assembly

Finally the work passes to the assembly shop where door frames are put together, drawers glued up, and carcasses assembled. After the glue has set, the parts may be returned to the machine department for machining that could not be performed before assembly, such as sanding the joints and shaping the edges. Then it returns to the assembly department for final assembly. Air-driven clamps are used when the design permits; otherwise the piece is pressed by hand clamps. Unless electronically cured glues are used, clamps must be applied long enough to ensure a good bond. The completed article is cleaned to remove excess glue, inspected, and hand sanded. Finally, staining and spray polishing is done and fittings added.

In individually crafted work there is always a great deal of fitting to be done—doors trimmed and drawers made to run easily without slackness. In mass-production work this problem would be serious. It is almost entirely avoided by making drawers an easy rather than snug fit and by sanding the edges of doors to templet size so that they automatically fit the carcasses, which in their turn are made to standard size.

The art of chairmaking. Chairmaking has been a separate branch of furniture making since the mid-17th century. One of the most intricate branches of woodwork, it involves odd angles, compound shapes, and awkward joints and at the same time calls for maximum strength, chairs being subjected to more strain than most other furniture. There are three main types of chairs: the Windsor chair, made largely from turned parts, with solid wood seat; the framed type of dining chair with either loose or stuff-over seat; and the upholstered chair.

In Britain the Windsor chair belongs traditionally to the High Wycombe District of Buckinghamshire where beech trees abound. Until relatively recent times men worked in huts in the beech woods making turned parts for chairs. They felled the trees, cut the trunks and larger branches into suitable lengths, and split them into pieces of a section large enough to permit chair legs and uprights to be turned and also to provide lighter members for rails, etc. They turned the parts on a primitive pole lathe in which a cord was attached to a treadle, taken around the wood to be turned and up to a springy sapling anchored at the lower end to pegs outside the hut. The power was supplied by treadle, the cord revolving the wood; then as the foot was raised the spring of the sapling lifted the treadle and at the same time turned the work backward. The turning gouge or chisel could be used on the downward stroke of the foot only, but the economy of effort was amazing. A complete leg could be rounded, the curves and beads formed, and the ends brought to the required diameter in a matter of seconds. Of course, working in green timber enabled the turning to be done much more easily and quickly than if the wood were dry.

These bodgers, as they were called, made only the turned parts and delivered them to chairmaking firms for assembling. They had no overhead expenses, no power costs, and the only lighting they needed in winter was an oil lamp or candles. They were long able to compete with powered workshops.

The Windsor chair

The manufacture of the Windsor chair of Victorian and Edwardian times was a specialized trade. The seat, invariably of elm, was hollowed out (bottomed) with a form of adze, and the holes for the legs were bored with a brace fitted with a spoon bit held at the required angle solely by judgment. The better chairs had a hooped back of yew. Today this hand work has been replaced by boring machines that are fitted with a jig to maintain the correct angle. The hollowing of the seat is machined to an extent, but the depth is only slight, compared with the early hand work. Furthermore, traditional timbers—elm, beech, and yew—are frequently replaced by imported timbers.

The quality of framed chairs of the dining type varies widely, but perhaps the outstanding general feature of modern dining chairs is the wide use of dowelled joints rather than mortise and tenon. In the late 19th century this had already occurred to a large extent, the chairmaker's kit of tools invariably including a dowel plate with a series of holes through which the craftsman hammered roughly squared pegs to form the dowels. Today machine-made dowels are universal, with a glue-escape slot cut in. Dowelling is a far quicker and consequently cheaper process than mortising and tenoning, especially in shaped work where the curved part frequently must be joined at odd angles.

When a chair has compound curvature it becomes difficult and expensive to make. A chair back may be shaped in both front and side elevation (and often in plan as well). Taste and experience are indispensable in providing a continuous curve that will be aesthetically satisfying from every angle. Over the years, experience has been built up, especially on traditional models following period lines; a chairmaker's workshop invariably carries bundles of templets in plywood for the various parts of chairs, with the fullness provided (where necessary) for a good line.

Dining chairs may be made in sets of half-dozens or dozens, or more cheaply in batches of 50 or 100, depending upon the capacity of the factory. In some cases parts are standardized and interchangeable in different designs of chairs.

The upholstering of dining chairs is a separate trade, though carried out in the same factory, and may be of the loose seat, stuff-over, or plywood-covered type. Traditional stuffing materials such as horsehair have largely been replaced by foam rubber and synthetics.

Fully upholstered furniture. The manufacture of fully upholstered pieces is often a separate branch, though large manufacturers often make both upholstered and nonupholstered types. It remains to a great extent a handicraft, for the skill of trained and experienced craftsmen is needed in turning out quality pieces. The standard upholstery foundation consists of a system of coiled steel springs resting on a webbing of burlap and tied to the furniture frame, which may be of wood, fibreglass, or plywood. The springs are embedded in a filling material, such as rubberized hair, foam rubber, palm fibre, or Spanish moss; and the spring system is topped by a thick padding of cotton or foam rubber in sheet form. Muslin is frequently employed as an inner covering for this assembly, while a durable upholstery fabric is used as an outer and finishing cover. Loose cushions are filled with special spring units in a bedding of foam rubber or down; the springs are covered with layers of cotton, and the entire assembly is encased in upholstery fabric.

Only a few mechanical aids have proved satisfactory in upholstering. Multiple layers of fabric can be cut efficiently and economically by machines in production-line operations, and staples are used instead of tacks on less expensive pieces. But the mass production of such components as the basic coiled-spring system and the mechanical handling of such materials as the bedding, or filler, have not proved efficient and economical. Some useful substitutes have been found, however, for the coiled-spring supporting unit; they include the modern nonsagging springs that may be clipped to the frame and the steel bands that are held to the frame by helical springs. Sponge rubber may be molded to constitute a complete seat that is firm and

comfortable. Webbed seat frames also are used, and the natural resiliency of wood is utilized in building springy plywood supporting structures. (C.H.H.)

BIBLIOGRAPHY

Automotive industry. *General histories:* Informative general accounts of the U.S. automotive industry include: R.M. CLEVELAND and S.T. WILLIAMSON, *The Road Is Yours* (1951); M. DENISON, *The Power to Go* (1956); and J.B. RAE, *American Automobile Manufacturers: The First Forty Years* (1959), and *The American Automobile* (1965). Comparable accounts for the United Kingdom are H.G. CASTLE, *Britain's Motor Industry* (1950); and KENNETH RICHARDSON, *The British Motor Industry 1896–1939* (1977).

Economic and social studies: Two comprehensive works on the impact of the automobile are C.D. BUCHANAN, *Mixed Blessing: The Motor in Britain* (1958); and J.B. RAE, *The Road and the Car in American Life* (1971). J.J. FLINK, *America Adopts the Automobile, 1895–1910* (1970), is a unique description of the acceptance of the motor vehicle.

Biographical works: A brief biography of Henry Ford is R. BURLINGAME, *Henry Ford: A Great Life in Brief* (1954, reissued 1964). A.P. SLOAN, JR., has a good autobiography, *My Years with General Motors* (1963). The best biography of a leading British automotive figure is P.W.S. ANDREWS and E. BRUNNER, *The Life of Lord Nuffield* (1955). A unique perspective on the German automotive industry is FERRY PORSCHE, with JOHN BENTLEY, *We at Porsche* (1976).

Modern industry: Excellent analyses of the American automobile industry in the 1950s and early 1960s are C.E. EDWARDS, *The Dynamics of the United States Automobile Industry* (1965); and L.J. WHITE, *The Automobile Industry since 1945* (1971). A comparable study for the United Kingdom is G. MAXCY and A. SILBERTSON, *The Motor Industry* (1959). A study of the industry and its management can be found in GERALD BLOOMFIELD, *The World Automotive Industry* (1978). An authoritative work on the automotive industry in developing countries is J. BARANSON, *Automotive Industries in Developing Countries* (1969).

Aerospace industry. *The Illustrated Encyclopaedia of Aviation and Space,* 14 vol. (1971), a general source written in layman's language; DAVID MONDEY, *The Complete Illustrated Encyclopedia of the World's Aircraft* (1978), includes an international guide to aircraft manufacturers; H. GUYFORD STEVER and JAMES J. HAGGERTY, *Flight* (1973), a general treatise on aviation for the layman; CHARLES H. GIBBS-SMITH, *The Aeroplane* (1960), the best source for historical aerospace data; F. ALEXANDER MAGOUN and ERIE HODGINS, *A History of Aircraft* (1931), an excellent source for early aviation history; WALTER T. BONNEY, *The American Heritage History of Flight,* ed. by ALVIN M. JOSEPHY (1962), a general historical work; CHARLES D. BRIGHT, *The Jet Makers: The Aerospace Industry from 1945–1972* (1978), an account of the aerospace industry in the United States since World War II; *Jane's All the World's Aircraft* (annual), the best reference work for aircraft production data on types produced, including space vehicles and missiles; *Aerospace Yearbook* (annual, 1919–70), the standard reference work on the aerospace industry in the United States.

Ship design and construction. D.G.M. WATSON, "Estimating Preliminary Dimensions in Ship Design," *Trans. Instn. Engrs. Shipbldrs. Scotl.,* 105:110–174 (1961–62), a comprehensive treatment of merchant ship design, rich in factual information and providing a series of useful graphs—an ingenious method of designing by volume is explained; T. LAMB, "A Ship Design Procedure," *Mar. Technology,* 6:362–405 (1969), a useful treatment of the subject, covering all types of merchant ships, with an extensive bibliography; R.V. TURNER, M. HARPER, and D.I. MOOR, "Some Aspects of Passenger Liner Design," *Trans. R. Instn. Nav. Archit.,* 105:379–414 (1963), a thoroughly practical paper dealing with the resistance, machinery, performance, stability, strength, and general arrangement of large passenger vessels; F.G. FASSETT, JR. (ed.), *The Shipbuilding Business in the United States of America,* 2 vol. (1948, reprinted 1970); D.J. EYRES, *Ship Construction,* 2nd ed. (1978).

Marine engineering: H.L. SEWARD (ed.), *Marine Engineering,* 2 vol. (1942–44), on all forms of prime movers, thermodynamics, and metals; J.H. MILTON, *Marine Steam Boilers,* 4th ed. (1980), on natural and forced circulated boilers, superheaters, and classification society rules; A. STODOLA and L.C. LOEWENSTEIN, *Steam and Gas Turbines,* 6th ed., 2 vol. (1927); H. ROXBEE COX (ed.), *Gas Turbine Principles and Practice* (1955), on thermodynamics, compressor and turbine design, materials, and fuels; C.C. POUNDER *et al., Marine Diesel Engines,* 5th ed. (1972), a study of all well-known types, design, construction, and maintenance; L.R. FORD, *Practical Marine Diesel Engineering,* 4th ed. (1943), on operation and maintenance, with questions and answers; S. GLASSTONE and A. SESONSKE, *Nuclear Reactor Engineering,* 2nd ed. (1980), originally prepared and compiled under auspices of the U.S. Atomic Energy Commission in 1963.

Clothing and footwear industry. J. SOLINGER, *Apparel Manufacturing Analysis* (1961) and *Apparel Manufacturing Handbook* (1980), detailed analyses of all subjects in the manufacture of apparel and allied items; H.C. CARR (ed.), *The Clothing Factory* (1972), provides information and analysis on garment technology and garment production management; J. KORN (ed.), *Boot and Shoe Production* (1953), and W.E. COHN, *Modern Footwear Materials and Processes* (1968), detailed technology of footwear production; UNITED SHOE MACHINERY CORP., *How Modern Shoes are Made* (1936), a concise booklet outlining the elements in constructing basic shoe styles; M.D. DANISH and L. STEIN (eds.), *ILGWU News History, 1900–1950* (1950), a history of the first 50 years of the ILGWU; SINGER SEWING MACHINE CO., *Invention of the Sewing Machine* (1955), a pamphlet that gives a brief history of the invention of the sewing machine; *Manufacturing Clothier* (1946–), a monthly magazine that reports current changes and innovations in the design, production, and marketing of women's wear in Europe, with particular emphasis on the British Isles; *FemmeLines* (1956–), a bi-monthly magazine dealing with trends and innovation in the design, production, and marketing of women's wear in the United States; *MascuLines* (1956–), a bi-monthly magazine dealing with trends and innovation in the design, production, and marketing of men's wear in the United States; *Daily News Record* (1892–), a daily paper reporting trends and innovations in design, production, and marketing of men's and boys' wear; *Women's Wear Daily* (1910–), a daily paper that reports trends and innovations in the design, production, and marketing of women's and children's wear; *The Bobbin* (1958–), a monthly magazine for management in the needle trades.

Floor coverings. R.S. BRINTON, *Carpets,* 2nd ed. (1932), traditional carpet structures and processes described at a basic technical level, including some that are obsolete or obsolescent; A. CROSSLAND, *Modern Carpet Manufacture* (1958), a description of carpet processes and structures, including nonwoven types, with helpful diagrams; B. JACOBS, *The Story of British Carpets* (1968), a history of the British carpet industry; P. LIEBETRAU, *Orientalske Taepper i farver* (1962; Eng. trans., *Oriental Rugs in Colour,* 1963), an illustrated treatment of the making and nature of Oriental rugs; G. ROBINSON, *Carpets and Other Textile Floorcoverings,* 2nd rev. ed. (1972); short sections on nonwoven carpets; F. WALTON, *The Infancy and Development of Linoleum Floorcloth* (1925), a history of linoleum by its inventor; D.T. WARD, *Tufting: An Introduction* (1969), on the development of the tufting process; HEDLEY POINTON (ed.), *Carpet Surfaces* (1975); E.W. ALLEN, "Adhesives for Flooring," R.F. BRIGGS, "Contract Flooring: Its Use, Misuse, and Abuse," W.J. WARLOW, "Plastics for Flooring, Their Use and Assessment," and J.G. BRAMLEY and W.V. WILKINSON, "Epoxy Floor Coatings," in *Floorcovering Symposium* (1965), papers on major developments and problems in hard floorcovering industry; R. BEAUMONT, *Carpets and Rugs* (1924), on the design and manufacture of carpets and rugs, not new but has some relevancy; S.R. COCKETT, *An Introduction to Man-Made Fibres* (1966), a general treatment of man-made fibres and their production; R.B. HOLT, *Rugs, Oriental and Occidental, Antique and Modern* (1901), rug making outlined in a brief and straightforward manner; E. KORNREICH, *Introduction to Fibres and Fabrics,* 2nd ed. (1966), a quick introduction to several important basic ideas in textiles; F.J. MAYERS, *Carpet Design and Designing* (1934), on various aspects of carpet designing; HENRI MUCHERIE, *Du Tapis de Roi au Tapis pour Tous* (1966), an illustrated descriptive outline of carpet structures (in French); H. OSSWALD, *Die Teppichindustrie,* 3rd ed. (1965), detailed technical coverage of main fields of carpet manufacture (in German).

Furniture industry. Information on the furniture industry and trade may be found in W. SKINNER *et al., Manufacturing Policy in the Furniture Industry,* 3rd ed. (1968); D.E. BOND and R.J. WONNACOTT, *Trade Liberalization and the Canadian Furniture Industry* (1968); S.H. SLOM, *Profitable Furniture Retailing* (1967); and J.L. OLIVER, *The Development and Structure of the Furniture Industry* (1966). Descriptions and illustrations of types of furniture, mostly antique, are given in J.E. GLOAG, *A Short Dictionary of Furniture,* rev. and enl. ed. (1969); and ERNEST JOYCE, *The Technique of Furniture Making* (1970). Works by C.H. HAYWARD include: *English Period Furniture Designs* (1969), a history of styles from the 16th through the mid-19th centuries; *Period Furniture Designs,* rev. ed. (1968), measured drawings of period pieces and their construction; and *Antique or Fake?* (1970), on how old furniture was made. A handbook on woodworking techniques and on the design and construction of modern furniture is J. and R. HOOPER, *Modern Furniture and Fittings,* 2nd ed. rev. (1955). For descriptions and illustrations of craftsman-made furniture, see A.E. BRADSHAW, *Handmade Woodwork of the Twentieth Century* (1962); and L.J. MAYES, *History of Chair Making in High Wycombe* (1960).

Textile Industries

The term textile, derived from the Latin *texere* ("to weave"), originally applied only to woven fabrics, has become a general term for fibres, yarns, and other materials that can be made into fabrics and for fabrics produced by interlacing or any other construction method. Thus threads, cords, ropes, braids, lace, embroidery, nets, and fabrics made by weaving, knitting, bonding, felting, or tufting are textiles. Some definitions of the term textile would also include those products obtained by the papermaking principle that have many of the properties associated with conventional fabrics.

This article surveys the development of textiles and the history and development of the textile industry. It treats in some detail the processes involved in the conversion of fibres to yarn, fabric construction, finishing operations applied to textiles, uses of textile materials, and the relationship between the producer and the consumer. Information about specific natural and synthetic textile fibres such as wool, mohair, nylon, and polyester are treated in separate articles in the *Micropædia*.

The article is divided into the following sections.

Development of textiles and the textile industry 508
- From prehistoric times to the 19th century 508
  - Early textile production
  - Effects of the Industrial Revolution
- From the 19th century to the present 509
  - Application of scientific methods
  - The modern textile industry

Production of yarn 510
- Textile fibres 510
  - Raw materials
  - Factors affecting cost
- Conversion to yarn 510
  - Treatment of raw fibre
  - Spinning
  - Yarn packages
- Types of yarn 511
  - Classification based on number of strands
  - Novelty yarns
  - Classification based on use
  - Measurement systems

Production of fabric 513
- Woven fabrics 513
  - The weaving process
  - Development of the loom
  - Basic weaves
  - Complex weaves
- Knitted fabrics 518
  - Knitting machines
  - Weft knitting
  - Warp knitting
- Other interlaced fabrics 520
  - Net and lace making
  - Braiding or plaiting
- Noninterlaced fabrics 521
  - Felt
  - Bonding
  - Laminating

Textile finishing processes 521
- Basic methods and processes 521
  - Preparatory treatments
  - Finishes enhancing appearance
  - Finishes enhancing tactile qualities
  - Finishes improving performance
- Dyeing and printing 523
  - Dyeing
  - Printing

Textile consumption 526
- Changing uses of fabric in apparel
- Household textiles
- Industrial fabrics
- Fabrics for protective clothing

Development of textiles and the textile industry

FROM PREHISTORIC TIMES TO THE 19TH CENTURY

Early textile production. Textile structures derive from two sources, ancient handicrafts and modern scientific invention. The earliest were nets, produced from one thread and employing a single repeated movement to form loops, and basketry, the interlacing of flexible reeds, cane, or other suitable materials. The production of net, also called limited thread work, has been practiced by many peoples, particularly in Africa and Peru. Examples of prehistoric textiles are extremely rare because of the perishability of fabrics. The earliest evidence of weaving, closely related to basketry, dates from Neolithic cultures of about 5000 BC. Weaving apparently preceded spinning of yarn; woven fabrics probably originated from basket weaving. Cotton, silk, wool, and flax fibres were used as textile materials in ancient Egypt; cotton was used in India by 3000 BC; and silk production is mentioned in Chinese chronicles dating to about the same period. The history of spinning technology will be touched on below in the section *Production of yarn: Spinning* and that of weaving technology in the section *Production of fabric*.

Early fabrics. Many fabrics produced by the simple early weaving procedures are of striking beauty and sophistication. Design and art forms are of great interest, and the range of patterns and colours is wide, with patterns produced in different parts of the world showing distinctive local features.

Yarns and cloth were dyed and printed from very early times. Specimens of dyed fabrics have been found in Roman ruins of the 2nd century BC; tie-and-dye effects decorated the silks of China in the T'ang dynasty (AD 618–907); and there is evidence of production of printed textiles in India during the 4th century BC. Textiles found in Egypt also indicate a highly developed weaving craft by the 4th century AD, with many tapestries made from linen and wool. Persian textiles of very ancient origin include materials ranging from simple fabrics to luxurious carpets and tapestries.

Textiles in the Middle Ages. By the early Middle Ages certain Turkish tribes were skilled in the manufacture of carpets, felted cloths, towels, and rugs. In Mughal India (16th–18th century), and perhaps earlier, the fine muslins produced at Dacca in Bengal were sometimes printed or painted. Despite the Muslim prohibition against representation of living things, richly patterned fabrics were made in Islāmic lands.

In Sicily, after the Arab conquest in AD 827, beautiful fabrics were produced in the palace workshops at Palermo. In about 1130, skilled weavers who came to Palermo from Greece and Turkey produced elaborate fabrics of silk interlaced with gold.

Growth of the textile industry in Italy

Following the conquest of Sicily in 1266 by the French, the weavers fled to Italy; many settled in Lucca, which soon became well-known for silk fabrics with patterns employing imaginative floral forms. In 1315, the Florentines captured Lucca, taking the Sicilian weavers to Florence, a centre for fine woven woollens from about 1100 AD, and also believed to be producing velvet at this time. A high degree of artistic and technical skill was developed, with 16,000 workers employed in the silk industry and 30,000 in the wool industry at the close of the 15th century. By the middle of the 16th century a prosperous industry in velvets and brocades (see below *Production of fabric*) was also established in Genoa and Venice.

Textile industries of France and Germany. French manufacture of woven silks began in 1480, and in 1520 Francis I brought Italian and Flemish weavers to Fontainebleau to produce tapestry under the direction of the King's weaver.

Others were brought to weave silk in Lyon, eventually the centre of European silk manufacture. Until 1589, most of the elaborate fabrics in France were of Italian origin, but in that year Henry IV founded the royal carpet and tapestry factory at Savonnières. Flemish weavers were brought to France to produce tapestries in workshops set up by Jean Gobelin in the 16th century. By the time of Louis XIII (1610–43), French patterned fabrics showed a distinctive style based on symmetrical ornamental forms, lacelike in effect, perhaps derived from the highly regarded early Italian laces. In 1662, the French government, under Louis XIV, purchased the Gobelin factory in Paris. Rouen also became known for its textiles, with designs influenced by the work of Rouen potters. French textiles continued to advance in style and technique, and under Louis XVI (1774–93) design was refined, with classical elements intermingled with the earlier floral patterns. The outbreak of the French Revolution in the 1790s interrupted the work of the weavers of Lyon, but the industry soon recovered.

Flanders and its neighbour Artois were early centres of production for luxurious textiles: Arras for silks and velvets; Ghent, Ypres, and Courtrai for linen damasks; and Arras and Brussels for tapestries. The damasks, characterized by heraldic motifs, were especially well known, and linen damasks of very high quality were produced in the 18th century. In Germany, Cologne was an important medieval cloth centre, renowned for orphrey webs (narrow cloths of gold bearing richly embroidered woven inscriptions and figures of saints).

Textile manufacture in England. English textiles of the 13th and 14th centuries were mainly of linen and wool, and the trade was influenced by Flemish fullers (finishers) and dyers. Silk was being woven in London and Norwich in 1455, and in 1564 Queen Elizabeth I granted a charter to Dutch and Flemish settlers in Norwich for production of damasks and flowered silks. The revocation of the Edict of Nantes in 1685, renewing persecution of French Protestants, caused many weavers to move to England, settling in Norwich, Braintree, and London. The most important group of refugees, some 3,500, lived in Spitalfields, a London settlement that became the chief centre for fine silk damasks and brocades. These weavers produced silk fabrics of high quality and were known for their subtle use of fancy weaves and textures. Norwich was also famous for figured shawls of silk or wool.

Textiles in the New World. Weaving and dyeing were established in the New World before arrival of the Europeans. Weaving was in an advanced state in North and South America during prehistoric times; both the Peruvians and Mexicans had fine woven fabrics. The Peruvian fabrics were much like those of ancient Egypt, although contact between the two civilizations is generally considered unlikely. Inca cotton and wool fabrics were brilliantly coloured, with patterns based on geometric and conventionalized human forms. Fabrics, especially blankets, made by the Navajos of Arizona and New Mexico had exceptionally close texture and brilliant colour.

American Indian textiles

English settlers established a cloth mill in Massachusetts in 1638. There Yorkshire weavers produced heavy cotton fustians; cotton-twill jeans; and linsey-woolsey, a coarse, loosely woven fabric of linen and wool. Fulling mills were operating in Massachusetts by 1654, freeing the community from dependence on England for fine linen and worsted. The industry developed steadily, and received a major impetus from Eli Whitney's invention of the cotton gin in 1793.

Effects of the Industrial Revolution. The textile industry, although highly developed as a craft, remained essentially a cottage industry until the 18th century. The advantages of cooperative operations were realized much earlier, and numbers of workers occasionally operated together under one roof, with one such group operating a mill in Zurich in 1568 and another in Derby, England, in 1717. Factory organization became most advanced in the north of England, and the Industrial Revolution, at its height between 1760 and 1815, greatly accelerated the growth of the mill system.

John Kay's flying shuttle, invented in 1733, increased the speed of the weaving operation, and its success created pressure for more rapid spinning of yarn to feed the faster looms. Mechanical spinners produced in 1769 and 1779 by Sir Richard Arkwright and Samuel Crompton encouraged development of mechanized processes of carding and combing wool for the spinning machines. Soon after the turn of the century the first power loom was developed. The replacement of water power by steam power increased the speed of power-driven machinery, and the factory system became firmly established, first in England, later in Europe and the United States.

FROM THE 19TH CENTURY TO THE PRESENT

Throughout the 19th century a succession of improvements in textile machinery steadily increased the volume of production, lowering prices of finished cloth and garments. The trend continued in the 20th century, with emphasis on fully automatic or nearly fully automatic systems of machinery.

Application of scientific methods. The mechanical developments in textile production associated with the Industrial Revolution resulted from the application of comparatively simple engineering and physical principles. Further progress required a clear understanding of the scientific principles of textile processing. Lack of basic information on the structure and properties of fibres limited understanding, delaying a scientific approach to textile processing. In the late 19th century, however, increasing knowledge of the physical and chemical properties of fibres led to application of scientific methods. Applications of chemistry originally attracted the greatest attention, largely because of the production of new compounds and the realization that fibres could be considered the result of chemical activity. In the 20th century, with the development of electronics and computers, new physical and engineering concepts were employed in textile research and development. An outstanding application of science to the textile industry was the development of man-made fibres, providing new textile materials and leading to the application of new processes to traditional fibres, thereby providing faster processing methods and introducing a wider variety of new techniques. The man-made fibre industry originally employed textile expertise developed through years of experience with natural fibres, but the excellent results obtained by the scientific approach encouraged increased industrial use of applied science, and information was soon accumulated on the behaviour of fibres under a variety of conditions.

Effects of development of man-made fibres

The modern textile industry. Both industrialized and developing countries now have modern installations capable of highly efficient fabric production. In addition to mechanical improvements in yarn and fabric manufacture, there have been rapid advances in development of new fibres, processes to improve textile characteristics, and testing methods allowing greater quality control.

The modern textile industry is still closely related to the apparel industry, but production of fabrics for industrial use has gained in importance. The resulting wide range of end uses demands a high degree of specialization. In the most technically advanced communities, the industry employs technicians, engineers, and artists; and a high degree of consumer orientation leads to emphasis on marketing operations. Some manufacturing operations, usually serving specialized or local markets and dependent on a limited number of firms for product consumption, still employ many hand operations, however.

Modern fabrics. The many types of modern textile fabrics, produced from both traditional and man-made materials, are often classified according to structure. Fabrics made by interlacing include woven and knitted types, lace, nets, and braid; fabrics produced from fibre masses include bonded types, wool felt, and needle-woven types; composite fabrics are produced by uniting layers of various types. Conventional weaving and knitting methods are currently the major textile manufacturing techniques, but newer construction methods are achieving acceptance, and may replace certain long-established products as costs of conventional textiles continue to rise and rapid technological advances continually develop new materials.

Quality control. Textile fabrics are judged by many cri-

teria. Flexibility and sufficient strength for the intended use are generally major requirements, and industrial fabrics must meet rigid specifications of width, weight per unit area, weave and yarn structure, strength and elongation, acidity or alkalinity, thickness, and porosity. In apparel fabrics design and colour are major considerations, and certain physical properties may be of secondary importance. In addition, the various tactile properties of a fabric, described as its "hand," "handle," or "feel," influence consumer acceptance.

Criteria for evaluating textile fabrics

The textile industry increasingly employs research and development in the area of quality control. Medieval craft guilds were concerned with maintaining high quality standards, and later textile mills established rigid systems of inspection, realizing that a reputation for supplying fault-free goods encouraged repeat orders. Modern quality control has been assisted by development of techniques and machines for assessing fibre, yarn, and fabric properties; by the introduction of legislation regarding misrepresentation in many industrialized countries; and by the establishment of rigid specifications by a growing number of buyers. Specifications have been established for the purchase of industrial fabrics, for textiles used by the military and other branches of governments, and for similar purchasing methods adopted by some retailers and other large buyers. In consumer-oriented areas, the public is becoming aware of product testing and is beginning to require proof that products have met certain test standards.

Many modern textile organizations test product quality at every major stage of processing. Yarns are tested for uniform thickness and other characteristics; fabric pieces are checked for defects; and the fastness of finishes and colours to various conditions is determined. Although it would not be feasible to test each yarn or fabric piece produced, statistical techniques allow maintenance of quality within previously specified limits, and the introduction of automatic testing devices has greatly reduced testing time and cost. Methods for assessing such properties as dimensions, strength, and porosity have been established, and their validity is generally accepted within the industry. Standards are available for colour fastness, although such important properties as water-repellency, resistance to creasing, and flame resistance are presently more difficult to define, and various organizations have adopted their own test procedures. It is important, for example, that a fabric described as flame resistant should conform to some specification in which the meaning of flame resistance is clearly defined.

Some manufacturers attach trademarks and quality labels to tested goods, and licensed trademarks are often associated with particular processes for which the manufacturer has been granted a license. The terms of the license require the manufacturer to ensure that his products meet the standards laid down by the proprietors of the particular process.

Production of yarn

Yarn is a strand composed of fibres, filaments (individual fibres of extreme length), or other materials, either natural or man-made, suitable for use in the construction of interlaced fabrics, such as woven or knitted types. The strand may consist of a number of fibres twisted together; a number of filaments grouped together but not twisted; a number of filaments twisted together; a single filament, called a monofilament, either with or without twist; or one or more strips made by dividing a sheet of material, such as paper or metal foil, and either twisted or untwisted. The properties of the yarn employed greatly influence the appearance, texture, and performance of the completed fabric.

TEXTILE FIBRES

Raw materials. Fibres are units of matter having length at least 100 times their diameter or width. Fibres suitable for textile use possess adequate length, fineness, strength, and flexibility for yarn formation and fabric construction, and for withstanding the intended use of the completed fabric. Other properties affecting textile fibre performance include elasticity, crimp (waviness), moisture absorption, reaction to heat and sunlight, reaction to the various chemicals applied during processing and in the dry cleaning or laundering of the completed fabric, and resistance to insects and microorganisms. The wide variation of such properties among textile fibres determines their suitability for various uses.

Qualities required in textile fibres

The first fibres available for textile use were obtained from plant and animal sources. Over a long period of experimentation with the many natural fibres available, cotton, wool, jute, flax, and silk have become recognized as the most satisfactory. The commercial development of man-made fibres began late in the 19th century, experienced much growth during the 1940s, expanded rapidly after World War II, and in the 1970s was still the subject of extensive research and development. This group includes regenerated fibres, such as rayon, made from fibre-forming materials already existing in nature and manipulated into fibrous form, and synthetic fibres, with the fibre-forming substance produced from chemicals derived from such sources as coal and oil, and then made into such fibres as nylon and polyester.

Factors affecting cost. The cost of fibres is determined by availability, the kind and amount of processing required, and their versatility. Natural fibres usually require extensive land area for their production, are affected by climatic conditions, and must frequently be transported long distances to the point of manufacture. Because quantity and quality are not easily controlled, prices tend to fluctuate. Research has been directed toward improving various properties during the manufacturing processes.

Man-made fibres can usually be produced near the point of use; their production does not require large land areas; they can be manufactured quickly, in desired quantities, with specific built-in properties; and they require little advance preparation for conversion to yarn. Initial costs are high because of the production equipment employed, but prices tend to be stable and may be reduced as production expands. Research has been directed toward improving the properties of man-made fibres and developing types suitable for specific purposes.

Although the major natural fibres continue to dominate the textile industry, production and consumption of synthetic fibres are growing.

CONVERSION TO YARN

Because filaments, such as silk and the man-made fibres, have extreme length, they can be made into yarn without the spinning operation necessary for the shorter staple fibres. When grouped together in a loose, continuous rope without twist, man-made filaments are called tow. Filaments may be loosely twisted together to form yarns of a specified thickness. Staple fibres, such as cotton, only a few inches long, must be tightly twisted together to produce satisfactory length.

Tow

Filament yarns are usually thin, smooth, and lustrous; staple yarns are usually thicker, fibrous, and without lustre. Man-made filaments cut to a predetermined short length become staple fibres, usually described by combining the fibre name with the term staple, as in rayon staple.

Treatment of raw fibre. In modern mills, most fibre-processing operations are performed by mechanical means. Such natural fibres as cotton, arriving in bales, and wool, arriving as fleece, are treated at the mill to remove various foreign materials, such as twigs and burrs. Wool must also be treated to remove suint, or wool grease; silk must be treated to remove sericin, a gum from the cocoon, and the very short silk fibres, or waste silk. Raw linen, the fibre of flax, is separated from most impurities before delivery. Man-made fibres, since they are produced by factory operations, rarely contain foreign materials. Blending, frequently employed for natural fibres, involves mixing fibres taken from different lots to obtain uniform length, diameter, density, and moisture content, thus assuring production of a uniform yarn. Blending is also employed when different fibres are combined to produce yarn. Man-made fibres, which can be cut into uniform tow, do not require blending unless they are to be mixed with other fibres.

Cotton, wool, waste silk, and man-made staple are sub-

jected to carding, a process of separating individual fibres and causing many of them to lie parallel, and also removing most of the remaining impurities. Carding produces a thin sheet of uniform thickness that is then condensed to form a thick, continuous, untwisted strand called sliver.

Carding

When very fine yarns are desired, carding is followed by combing, a process that removes short fibres, leaving a sliver composed entirely of long fibres, all laid parallel, and both smoother and more lustrous than uncombed types. Slivers may be loosely twisted together, forming roving. Hackling, a process applied to straighten and separate flax, is similar to combing.

Spinning. *Early spinning methods.* Spinning is the process of drawing out and twisting fibres to join them firmly together in a continuous thread or yarn. Spinning is an indispensable preliminary to weaving cloth from those fibres that do not have extreme length. From early times through the Middle Ages spinning was accomplished with the use of two implements, the distaff and the spindle. The distaff was a stick on which the mass of fibres was held. The drawn-out length of fibre was fastened to the weighted spindle, which hung free. The spinner whirled the spindle, causing it to twist the fibre as it was drawn from the distaff. As a length was drawn out the operation was halted, the new yarn wound on the spindle and secured by a notch, and the operation repeated. The spinning wheel, invented in India and introduced to Europe in the Middle Ages, mechanized the process; the spinning of the wheel supplanted the whirl of the weighted spindle, and after each operation the spinner wound the new yarn on the spindle. This was accomplished simply and speedily by holding the yarn outstretched with the left hand and feeding it as the wheel was spun in the reverse direction.

An important advantage conferred by the spinning wheel was the fact that it tended to add more twist at thin places in the forming yarn and to draw out the thicker places, giving a more uniform yarn.

Improvements on the spinning wheel

The spinning wheel continued in use into the 19th century, receiving an important improvement in the 16th century in the form of the Saxony wheel, which made possible continuous spinning of coarse wool and cotton yarn. With this improvement in speed, three to five spinning wheels could supply one loom with yarn, but Kay's flying shuttle (described below under *Woven fabrics*) greatly increased the output of the loom and created a demand for spinning machinery. James Hargreaves' spinning jenny (patented 1770) operated a number of spindles simultaneously, but was suitable only for making yarn used as filling. Sir Richard Arkwright, making use of earlier inventions, produced a better machine, capable of making stronger yarn than Hargreaves' jenny. Still a third machine, Samuel Crompton's "mule" (1779), vastly increased productivity, making it possible for a single operator to work more than 1,000 spindles simultaneously; and it was capable of spinning fine as well as coarse yarn. Several further modifications were introduced in Britain and the United States, but the Crompton mule effectively put yarn spinning on a mass production basis.

Modern spinning. In modern spinning, slivers or rovings are fed into machines with rollers that draw out the strands, making them longer and thinner, and spindles that insert the amount of twist necessary to hold the fibres together. Tightness of the twist determines the strength of the yarn, although too much twist may eventually cause weakening and breakage. When the spirals formed by twisted yarns are similar in slope to the central portion of the letter Z, the yarns are described as Z-twist; when the spirals conform in direction to the central portion of the letter S, the yarns are described as S-twist. Crepe yarns, producing a crinkled effect in fabrics, are made with a very high degree of twist, producing a kink. Shadow effects can be produced in finished fabrics by the use of yarns combining opposing twists, producing differing light reflections. The spinning process is completed by winding the yarn on spools or bobbins.

Reeling and throwing. Reeling is the process of unwinding raw silk filament from the cocoon directly onto a holder. When several filament strands, either raw silk or man-made, are combined and twisted together, producing yarn of a specified thickness, the process is called throwing.

Yarn packages. The intended use of a yarn usually determines the packaging method employed. Bobbins are wood, cardboard, or plastic cores on which yarns are wound as they are spun, and have holes in their centres allowing them to fit on spindles or other holding devices. Spools are cylindrical, with end flanges. Cones, having a conical-shaped core, produce a package of conical shape; tubes, with cylindrical-shaped cores, produce cylindrical packages. Cheeses are cylindrical yarn packages wound on a tube, and, unlike most other packages, they have greater diameter than height. Skeins are coils of yarn wound with no supporting core.

Pirns are large barrel-shaped packages used to hold the weft, or filling, yarn supply for the shuttle in weaving; quills are small tapered tubes holding the weft yarns for weaving. Beams are wood or metal cylinders, about five feet long and up to 10 inches in diameter, on which yarns used as warp in weaving are wound.

TYPES OF YARN

Classification based on number of strands. Yarns can be described as single, or one-ply; ply, plied, or folded; or as cord, including cable and hawser types.

Single yarns. Single, or one-ply, yarns are single strands composed of fibres held together by at least a small amount of twist; or of filaments grouped together either with or without twist; or of narrow strips of material; or of single man-made filaments extruded in sufficient thickness for use alone as yarn (monofilaments). Single yarns of the spun type, composed of many short fibres, require twist to hold them together and may be made with either S-twist or Z-twist (see Figure 1). Single yarns are used to make the greatest variety of fabrics.

From Evelyn E. Stout, *Introduction to Textiles* (© 1970); John Wiley & Sons, Inc.

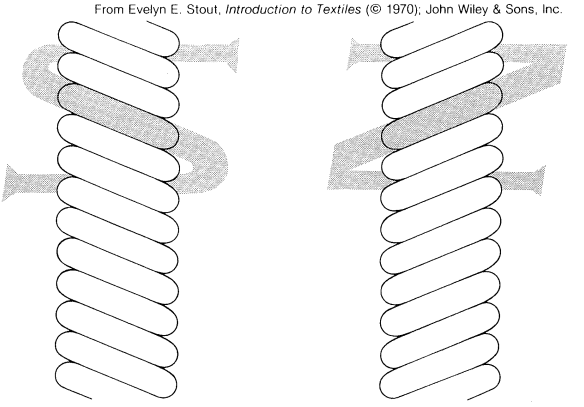


Figure 1: (Left) S- and (right) Z-twist yarns.

Ply yarns. Ply, plied, or folded, yarns are composed of two or more single yarns twisted together. Two-ply yarn, for example, is composed of two single strands; three-ply yarn is composed of three single strands. In making ply yarns from spun strands, the individual strands are usually each twisted in one direction and are then combined and twisted in the opposite direction. When both the single strands and the final ply yarns are twisted in the same direction, the fibre is firmer, producing harder texture and reducing flexibility. Ply yarns provide strength for heavy industrial fabrics and are also used for delicate-looking sheer fabrics.

Cord yarns. Cord yarns are produced by twisting ply yarns together, with the final twist usually applied in the opposite direction of the ply twist (see Figure 2). Cable cords may follow an SZS form, with S-twisted singles made into Z-twisted plies that are then combined with an S-twist, or may follow a ZSZ form. Hawser cord may follow an SSZ or a ZZS pattern. Cord yarns may be used as rope or twine, may be made into very heavy industrial fabrics, or may be composed of extremely fine fibres that are made up into sheer dress fabrics.

Novelty yarns. Novelty yarns include a wide variety of yarns made with such special effects as slubs, produced by intentionally including small lumps in the yarn structure,

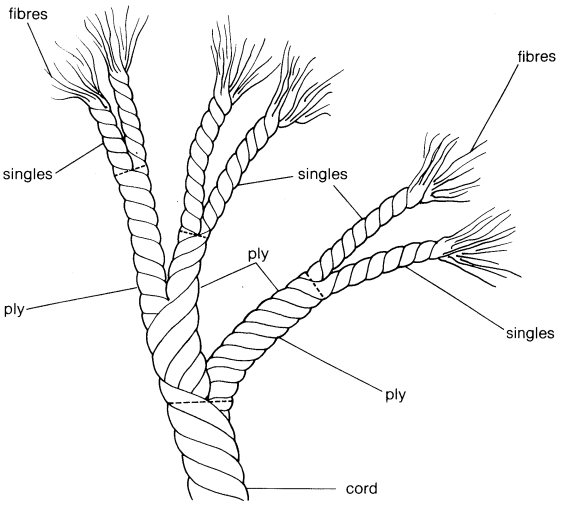


Figure 2: Single, ply, and cord yarns.

From E. Stout, *Introduction to Textiles*, 3rd ed., Copyright © 1970; John Wiley & Sons, Inc. By permission of John Wiley & Sons, Inc.

and man-made yarns with varying thickness introduced during production. Natural fibres, including some linens, wools to be woven into tweed, and the uneven filaments of some types of silk cloth are allowed to retain their normal irregularities, producing the characteristic uneven surface of the finished fabric. Man-made fibres, which can be modified during production, are especially adaptable for special effects such as crimping and texturizing.

Textured yarns. Texturizing processes were originally applied to man-made fibres to reduce such characteristics as transparency, slipperiness, and the possibility of pilling (formation of small fibre tangles on a fabric surface). Texturizing processes make yarns more opaque, improve appearance and texture, and increase warmth and absorbency. Textured yarns are man-made continuous filaments, modified to impart special texture and appearance (see Figure 3). In the production of abraded yarns, the surfaces are roughened or cut at various intervals and given added twist, producing a hairy effect.

From G.E. Linton, *Applied Basic Textiles*, 2nd ed. (1973)

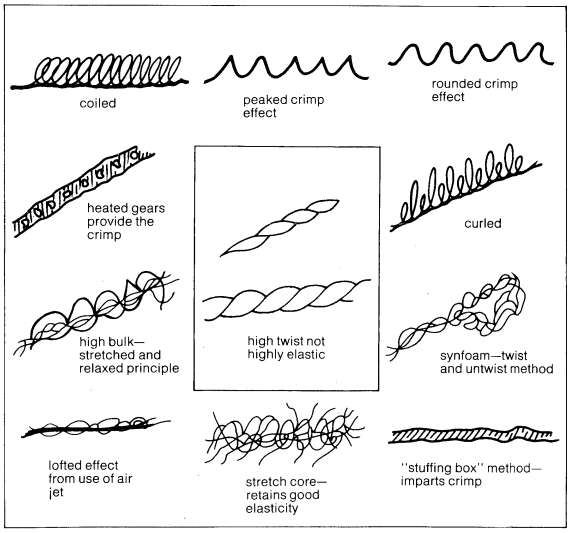


Figure 3: Examples of textured yarns.

Bulking

Bulking creates air spaces in the yarns, imparting absorbency and improving ventilation. Bulk is frequently introduced by crimping, imparting waviness similar to the natural crimp of wool fibre; by curling, producing curls or loops at various intervals; or by coiling, imparting stretch. Such changes are usually set by heat application, although chemical treatments are sometimes employed. In the early 1970s bulky yarns were most frequently produced by the "false twist" method, a continuous process in which the filament yarn is twisted and set, and then untwisted and heated again to either stabilize or destroy the twist. The "stuffing box" method is often applied to nylon, a process in which the filament yarn is compressed in a heated tube, imparting a zigzag crimp, then slowly withdrawn. In the knit-de-knit process, a synthetic yarn is knitted, heat is applied to set the loops formed by knitting, and the yarn is then unravelled and lightly twisted, thus producing the desired texture in the completed fabric.

Bulk may be introduced chemically by combining filaments of both high and low shrinkage potential in the same yarn, then subjecting the yarn to washing or steaming, causing the high shrinkage filaments to react, producing a bulked yarn without stretch. A yarn may be air bulked by enclosing it in a chamber where it is subjected to a high-pressure jet of air, blowing the individual filaments into random loops that separate, increasing the bulk of the material.

Stretch yarns. Stretch yarns are frequently continuous-filament, man-made yarns that are very tightly twisted, heat-set, then untwisted, producing a spiral crimp giving a springy character. Although bulk is imparted in the process, a very high amount of twist is required to produce yarn that has not only bulk, but also stretch.

Elastomeric fibres

Spandex is the generic term for a highly elastic synthetic fibre composed mainly of segmented polyurethane. Uncovered fibres may be used alone to produce fabrics, but they impart a rubbery feel. For this reason, elastomeric fibre is frequently used as the core of a yarn and is covered with a nonstretch fibre of either natural or man-made origin. Although stretch may be imparted to natural fibres, other properties may be impaired by the process, and the use of an elastic yarn for the core eliminates the need to process the covering fibre.

Metallic yarns. Metallic yarns are usually made from strips of a synthetic film, such as polyester, coated with metallic particles. In another method, aluminum foil strips are sandwiched between layers of film. Metallic yarns may also be made by twisting a strip of metal around a natural or man-made core yarn, producing a metal surface.

Additional information about the production, characteristics, and uses of modern man-made novelty yarns is provided in INDUSTRIES, CHEMICAL PROCESS: *Man-made fibres.*

Classification based on use. *Fabric construction yarns.* Almost any textile yarn can be used to produce such interlaced fabrics as woven and knitted types. In weaving, the warp, or lengthwise, yarns are subjected to greater stress and are usually stronger, smoother, and more even and have tighter twist than the weft, or crosswise, yarns. A sizing (stiffening) material such as starch may be applied to warp yarns, increasing their strength to withstand the stresses of fabric construction operations. Weft yarns, subjected to little stress during weaving, may be quite fragile.

Warp and weft threads used in the same fabric may be of differing diameter, producing such special effects as ribbing or cording in the fabric. Special effects may also be obtained by combining warp and weft yarns of fibre from differing origin, or with different degrees of twist, or by introducing metallic threads into weaves composed of other fibres.

Yarns for machine knitting are usually loosely twisted because softness is desired in knit fabrics.

Yarns used in handwork. Yarns used in hand knitting are generally of two or more ply. They include such types as fingering yarns, usually of two or three plys, light to medium in weight and with even diameter, used for various types of apparel; Germantown yarns, soft and thick, usually four-ply and of medium weight, frequently used for sweaters and blankets; Shetland yarns, fine, soft, fluffy, and lightweight, frequently two-ply, used for infants' and children's sweaters and for shawls; worsted knitting yarn, highly twisted and heavy, differing from worsted fabric by being soft instead of crisp, and suitable for sweaters; and zephyr yarns, either all wool, or wool blended with other fibres, very fine and soft, with low twist, and used for lightweight garments.

Embroidery floss, used in hand embroidery, generally has low twist, is of the ply or cord type, and is made of such smooth filaments as silk and rayon. Yarn used

for crocheting is frequently a loose cotton cord type; and darning yarns are usually loosely spun.

Sewing thread. Sewing threads are tightly twisted ply yarns made with strands having equally balanced twist, producing a circular cross section. Thread for use in commercial or home sewing machines and for hand sewing should allow easy movement when tension is applied and ease in needle threading; should be smooth, to resist friction during sewing; should have sufficient elasticity to avoid the breaking of stitches or puckering of seams; and should have sufficient strength to hold seams during laundering or dry cleaning and in use.

Threads for special uses may require appropriate treatment. Garments made of water-repellent fabrics, for example, may be sewn with thread that has also been made water-repellent. Thread is usually subjected to special treatment after spinning, and is then wound on spools. Thread size is frequently indicated on the spool end, and systems for indicating degree of fineness vary according to the textile measurement system used locally.

Silk, cotton, and nylon threads

Silk thread has great elasticity and strength combined with fine diameter. It can be permanently stretched in sewing, and is suitable for silks and wools. Buttonhole twist is a strong, lustrous silk about three times the diameter of normal sewing silk, and is used for hand-worked buttonholes, for sewing on buttons, and for various decorative effects.

Cotton thread is compatible with fabric made from yarn of plant origin, such as cotton and linen, and for rayon (made from a plant substance), because it has similar shrinkage characteristics. It is not suitable for most synthetics, which do not shrink, or for fabrics treated to reduce shrinkage. Its low stretch is useful for woven fabrics, but not for knits, which require more stretch.

Nylon thread is strong, with great stretch and recovery, does not shrink, and is suitable for sheers and for very stretchy knits. Polyester thread has similar characteristics, and is appropriate for various synthetic and preshrunk fabrics, and for knits made of synthetic yarns.

Measurement systems. Yarn measurements are expressed as yarn number, count, or size, and describe the relationship of length and weight (or approximate diameter). Because methods of measurement were developed in various areas of the world, there has been a lack of uniformity in such systems.

Indirect systems. Indirect measuring systems are those employing higher number to describe finer yarns, and are based on length per unit weight. Most countries measure yarns made from staple fibres according to the weight of a length of yarn. If one pound is used as a standard unit, for example, a very fine yarn will have to be much longer than a coarser yarn to weigh a pound, so higher counts indicate finer yarns. The size number is an indication of the length of yarn needed to reach a weight of one pound.

In the United States, the system is based on the number of hanks per pound, with a hank of 840 yards for cotton and spun silk, 300 yards (a lea) for linen, 256 yards for woollen yarns, and 560 yards for worsted yarns. A widely used continental system is based on the number of hanks of 1,000 metres (one kilometre) required to reach a weight of one kilogram.

Denier system. The denier system is a direct-management type, employed internationally to measure the size of silk and man-made filaments and yarns, and derived from an earlier system for measuring silk filaments (based on the weight in drams of 1,000 yards). Denier number indicates the weight in grams of 9,000 metres of filament or filament yarn. For example, if 9,000 metres of a yarn weigh 15 grams, it is a 15-denier yarn; if 9,000 metres of a yarn weigh 100 grams, it is a 100-denier yarn, and much coarser than the 15-denier yarn. Thus a smaller number indicates a finer yarn. This system is not convenient for measurement of staple yarns because their greater weight would require the use of very large numbers.

Tex system. The tex system, originally devised in 1873, is a universal method developed for the measurement of staple fibre yarns and is also applicable to the measurement of filament yarns. It is based on the weight in grams of one kilometre (3,300 feet) of yarn.

Production of fabric

Fabric construction involves the conversion of yarns, and sometimes fibres, into a fabric having characteristics determined by the materials and methods employed. Most fabrics are presently produced by some method of interlacing, such as weaving or knitting. Weaving, currently the major method of fabric production, includes the basic weaves, plain or tabby, twill, and satin, and the fancy weaves, including pile, Jacquard, dobby, and gauze. Knitted fabrics are rapidly increasing in importance and include weft types and the warp types, raschel and tricot. Other interlaced fabrics include net, lace, and braid. Nonwoven fabrics are gaining importance and include materials produced by felting and bonding. Laminating processes are also increasing in importance, and fairly recent developments include needle weaving and the sewing-knitting process.

WOVEN FABRICS

Woven fabrics are made of yarns interlaced in a regular order called a binding system, or weave. Weaving is the process of combining warp and weft components to make a woven structure. The components need neither be parallel to each other nor cross each other at right angles, but most woven structures are composed of two sets of components, both flexible and crossing at right angles. Weaving is differentiated from warp and weft knitting, braiding, and net making in that these latter processes make use of only one set of elements. In addition, there are geometrical differences, one of the most significant being the small angles through which the components of a woven structure are, in general, bent, in contrast with the components of other structures.

Weaving is a widely used constructional method because it is cheap, basically simple, and adaptable. Woven fabrics have valuable characteristics resulting partly from the geometrical conformation of their components and partly from the fact that the components are held in position not by rigid bonding but by friction set up at the areas where they make contact. Woven fabrics are used in household, apparel, and industrial textiles.

Textile designers can produce a very large variety of cloths by their selection of yarns, finishing processes, and binding systems. Yarns vary in thickness, smoothness, fibre content, twist, and colour, all of which have a profound influence on the finished cloth. Finishing processes range from such simple treatment as brushing up the nap on a woven fabric to such a complicated chemical process as that employed to change opaque cotton fabric to transparent, permanently stiffened organdy.

The binding system, or weave, however, is the basic factor in determining the character of a woven fabric. The three basic systems are plain or tabby, twill, and satin. In complex binding systems, the basic weaves are combined or enriched by hand manipulation or mechanical loom attachments; these include multiple-plane, pile, inlaid, and gauze weaves. Regardless of the binding system, other devices—manipulation of warp spacing, beating in, or tension—can be used to alter the appearance of any weave, to make it looser or more compact, to make it more or less regular.

Weave drafts

As musical notation conveys a composer's ideas, so weave drafts or point paper plans communicate a textile designer's directions for constructing woven fabrics. The draft is a plan on graph paper showing at least one repeat or weave unit of the fabric to be woven. This information enables the weaver or mill specialist to plot the drawing in of the warp, tie up of harnesses to shedding mechanism, and shedding order.

The weaving process. Woven cloth is normally much longer in one direction than the other. The lengthwise threads are called the warp; the other threads, which are combined with the warp and lie widthwise, are called the weft (synonyms are "filling," "woof," and "shoot," or "shute"). An individual thread from the warp, of indefinite length, is called an end; each individual length of weft, extending from one edge of the cloth to the other, is called a pick, or shot. Consecutive picks are usually consecutive

lengths of one piece of weft yarn that is repeatedly folded back on itself.

In all methods of weaving cloth (except the rudimentary form of darning), before a length of weft is inserted in the warp, the warp is separated, over a short length extending from the cloth already formed, into two sheets. The process is called shedding and the space between the sheets the shed (Figure 4A). A pick of weft is then laid between the two sheets of warp, in the operation known as picking (Figure 4B). A new shed is then formed in accordance with the desired weave structure, with some or all of the ends in each sheet moving over to the position previously occupied by the other sheet. In this way the weft is clasped between two layers of warp.

Shedding, picking, and beating in

Since it is not possible to lay the weft close to the junction of the warp and the cloth already woven, a further operation called beating in, or beating up (Figure 4C), is necessary to push the pick to the desired distance away from the last one inserted previously. Although beating in usually takes place while the shed is changing, it is normally completed before the new shed is fully formed.

The sequence of primary operations in one weaving cycle is thus shedding, picking, and beating in. At the end of the cycle the geometrical relation of the pick to the warp is the same as it would have been if the pick had been threaded through the spaces between alternate ends, first from one side of the cloth and then from the other, as in darning. This is the reason the weaving process is considered an interlacing method.

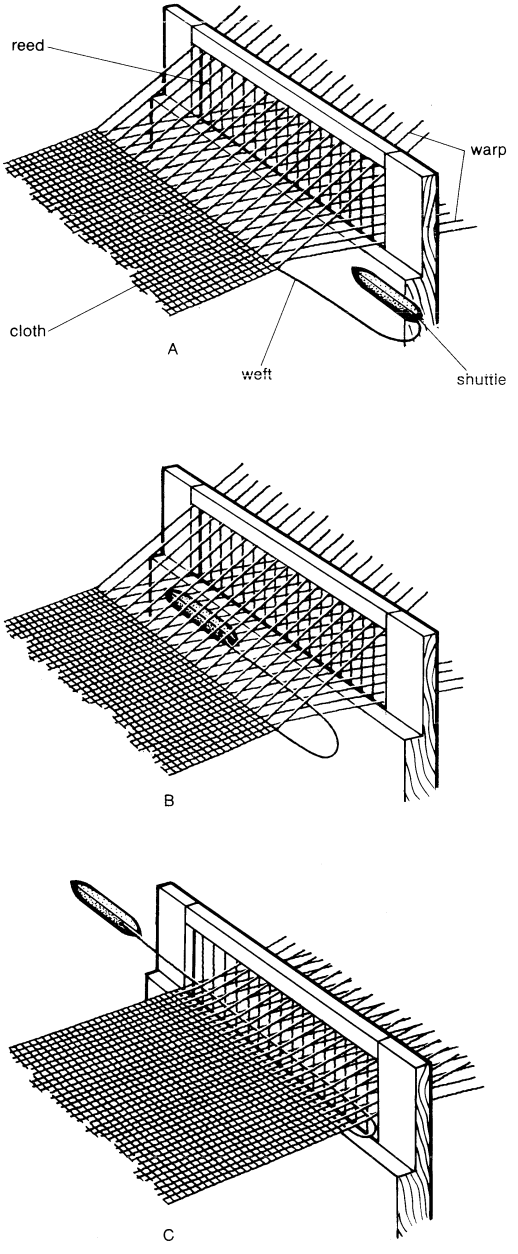


Figure 4: *The three basic motions of weaving:* (A) shedding, (B) picking, and (C) beating in.

Development of the loom. The word loom (from Middle English *lome,* "tool") is applied to any set of devices permitting a warp to be tensioned and a shed to be formed. Looms exist in great variety, from the bundles of cords and rods of primitive peoples to enormous machines of steel and cast iron.

Except on certain experimental looms, the warp shed is formed with the aid of heddles (or healds). Usually one heddle is provided for each end, or multiple end, of warp thread, but on some primitive looms simple cloths are produced with heddles provided only for each alternate end. A heddle (Figure 5A) consists of a short length of cord, wire, or flat steel strip, supported (in its operative position) roughly perpendicular to the unseparated sheet of warp threads and provided, in modern looms, with an eyelet at its midpoint, through which the warp end is threaded. By pulling one end of the heddle or the other, the warp end can be deflected to one side or the other of the main sheet of ends. The frame holding the heddles is called a harness.

By courtesy of the University of Manchester Institute of Science and Technology, England

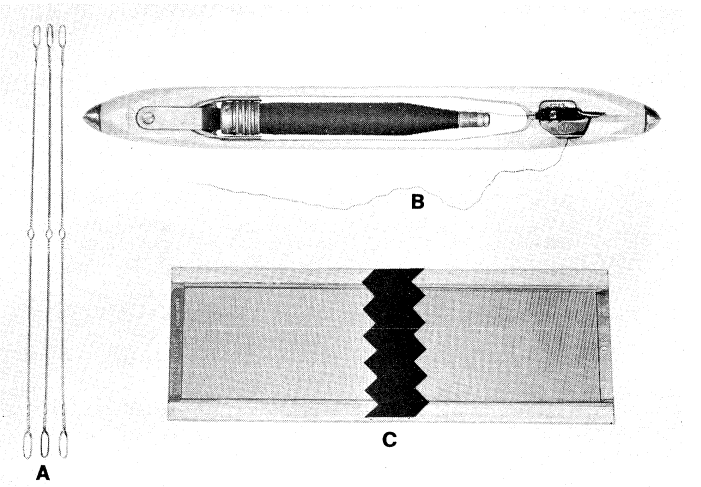


Figure 5: *Loom devices and their functions:* (A) heddles, used for shedding; (B) the shuttle, used for picking; (C) the reed, used for beating in (see text).

In most looms, the weft is supplied from a shuttle, (Figure 5B), a hollow projectile inside which a weft package is mounted in such a way that the weft can be freely unwound through an eyelet leading from the inside to the outside. The shuttle enters the shed and traverses the warp, leaving a trail of weft behind.

Beating in is generally effected by means of a grating of uniformly spaced fine parallel wires, originally made of natural reeds and thus called a reed (Figure 5C), which, mounted at right angles to the warp, oscillates between the heddles and the junction of the warp and the cloth. The ends pass, one or more at a time, through the spaces between consecutive reed wires, so that the reed, in addition to beating in, controls the spacing of the ends in the cloth.

Early looms. The earliest evidence of the use of the loom (4400 BC) is a representation of a horizontal two-bar (or two-beamed—*i.e.,* warp beam and cloth beam) loom pictured on a pottery dish found at al-Badārī, Egypt. The warp is stretched between two bars or beams, pegged to the ground at each of the four corners. Lease (or laze) rods are used to separate the warp yarns, forming a shed and aiding the hands in keeping the yarns separated and in order. Lease rods were found in some form on every later type of improved loom, and their use at this very early date indicates that the loom already had been in use long enough to have reached a stage of improvement by addition of devices to aid the hands.

Lease rods

Before lease rods were added, it would have been necessary for the fingers to separate each odd from each even warp thread to create the shed through which the weft yarn was passed. A third rod also seen in this early drawing may be a heddle rod. If so, this loom represents a still more advanced stage of development.

The heddle rod rests on top of the warps. To produce a plain weave, alternate warp yarns are tied to the rod, and when it is raised, the shed is formed quickly and accurately. Some authorities consider the heddle to be the most important step in the evolution of the loom. A shed stick is ordinarily used with the heddle, forming the second, or countershed, opening for the return of the weft.

In addition to the horizontal two-bar loom, there are two other primitive varieties, the warp-weighted and the vertical two-bar loom. The warp-weighted loom consists of a crossbar supported by two vertical posts. The warp threads hang from the crossbar and are held taut by weights of clay, ceramic, or chalk tied to their free ends. Loom weights have been found at archaeological sites dating from 3000 BC, but this type of loom may have originated even earlier. The earliest picture of a vertical two-bar loom is from the Egyptian 18th dynasty (1567–1320 BC). It coincides with the appearance of more intricate textile patterns, the earliest known tapestries (datable between 1483 and 1411 BC) having been found in the tomb of Thutmose IV at Thebes. (Even today the vertical loom is preferred for tapestry weaving.) In the vertical two-bar loom the ends of the warp yarns are attached to a second crossbar, thus combining features of both the horizontal two-bar and the warp-weighted looms.

The heddle rods and shed sticks are used in a similar way on all three types.

Counterparts of these very early looms have been used through the ages in many cultures. The Navajo Indians, probably the best known of the American Indian weavers, have used the simple two-bar vertical loom for several centuries to produce their beautiful rugs and blankets. A form of the horizontal two-bar loom was the back-strap loom, in which one bar was tied to a tree or other stationary device, the second being attached to the weaver's waist by a strap. The weaver could control the tension of the warp yarns by applying pressure as necessary. The back-strap loom was used in pre-Columbian Peru, in other cultures of Central and South America, in Asia, and elsewhere.

Horizontal frame looms. By about 2500 BC a more advanced loom was apparently evolving in the Far East. Fragments of silk fabrics found adhering to bronzes of the Shang (or Yin) period (18th–12th century) in China show traces of a twill damask pattern, suggesting an advanced weaving knowledge, since such fabrics could not practicably be woven on the looms described above. These fabrics were probably produced on a horizontal frame loom with treadles. The logical connecting link between the horizontal two-bar and the horizontal frame loom with treadles would have been a loom with a heddle rod that was controlled by one foot, for which no early illustrations have been found.

Early horizontal two-bar treadle looms

The earliest European pictorial record of the horizontal frame loom with a treadle dates from the 13th century, when it appears in a highly developed form, almost certainly introduced from the East. This two-bar loom was mounted in a frame; to this was connected a treadle operated by the feet, moving the heddles, an improvement of the heddle rod or cord controls now mounted between bars and called a shaft. The advantages of this type of loom were many. First, in the two-bar loom, though more than two heddle rods could be used, the number of groupings of warp threads was limited. Although highly complex patterns could be woven, it was not practical to do so in producing any but very small quantities of cloth. The shaft loom allowed as many as 24 shafts to be set up easily, enabling the weaver to produce comparatively intricate patterns. Second, the weaver's sword or comb formerly used to beat the weft into place was replaced by the batten, supported in a heavy wooden frame from the main frame of the loom; its weight and free-swinging motion improved the beating-in action and made it easier. Third, use of the foot treadle freed both hands to throw the shuttle and swing the batten. The loom remained virtually unchanged for many centuries thereafter.

Drawlooms. The shaft loom was adequate for plain and for simply patterned fabrics, but a more complex loom was needed for the weaving of intricately figured fabrics, which might require 100 or more shafts. This kind of weaving was accomplished on the drawloom. Its origin is unknown, but it probably was first used in the Far East for silk weaving and was introduced into the silk-working centres of Italy during the Middle Ages. The drawloom had two devices for shedding: in addition to the shafts, which the weaver operated by treadles, cords were also used to raise the warp threads, gathered into groups as required by the pattern. The cords were worked by an operator (called a drawboy) seated on top of the loom.

The drawloom was improved in Italy and France in the early 17th century by the addition of a type of mechanical drawboy, allowing the assistant to stand on the floor at the side of the loom and increasing the control of the cords. The continued inconvenience of employing an assistant, however, who might also make errors, led to a search for an automatic mechanism that would perform all work of the drawboy. Most of the later developments in automatic mechanisms to control the shedding operation originated in France, which had become one of the leading countries in the weaving of figured silks.

The mechanical drawboy

In 1725 Basile Bouchon added to the mechanical drawboy a mechanism that selected the cords to be drawn to form the pattern. Selection was controlled by a roll of paper, perforated according to the pattern, which passed around a cylinder. The cylinder was pushed toward the selecting box and met with needles carrying the warp-controlling cords; the needles that met unperforated paper slid along, and the others passed through the holes and remained stationary. The selected cords were drawn down by a foot-operated treadle.

The mechanical drawboy made the proper selection of warp threads, eliminating errors, but still required an operator. The mechanism was improved in 1728 by increasing the number of needles and using a rectangular perforated card for each individual shedding motion, the cards being strung together in an endless chain. In 1745 Jacques de Vaucanson constructed a loom incorporating a number of improvements. He mounted the selecting box above the loom, where it acted directly on hooks fastened to the cords that controlled the warp yarns. The hooks passed through needles and were raised by a strong metal bar. The needles were selected by perforated cards passing around a sliding cylinder, without the aid of a second operator or assistant. The cylinder was very complex, and the mechanism is not known to have been adopted, but it served as the foundation for the successful Jacquard attachment.

The Jacquard attachment. The French inventor Joseph-Marie Jacquard, commissioned to overhaul Vaucanson's loom, did so without any of the directions, which were missing. In 1801, at the Paris Industrial Exhibition, he demonstrated an improved drawloom, and in 1805 he introduced the invention that ever since has caused the loom to which it is attached to be called the Jacquard loom.

The Jacquard attachment is an automatic, selective shedding device, mounted on top of the loom and operated by a treadle controlled by the weaver. As in the drawloom, every warp yarn runs through a loop in a controlling cord, held taut by a weight. Each cord is suspended from a wire ("hook") that is bent at the bottom to hold the cord and bent at the top in order to hook around the blades or bars of the griff, the lifting mechanism. To allow only those warp threads that are needed to form the pattern to be raised, some hooks must be dislodged from the rising griff. This is accomplished by horizontally placed needles connected to the hooks. As the perforated pattern card moves into place on the cylinder (which is, in fact, a quadrangular block), the needles pass through the holes in the card, and the warps are raised; where there are no holes, the needles are pushed back (by a spring action on the opposite end of each), pulling the hooks away from the rising griff bar, and the warps are not raised.

Each card represents one throw of the shuttle, and the pattern is transferred to the cards from the designer's

weave draft. Although each Jacquard attachment is limited in the number of hooks it can control and, therefore, in the size of the repeat pattern, by adding several Jacquard attachments to one loom, the weaver not only can produce intricately figured fabrics but also can weave pictures of considerable size.

The flying shuttle. The first decisive step toward automation of the loom was the invention of the flying shuttle patented in 1733 by the Englishman John Kay. Kay was a weaver of broadloom fabrics, which because of their width required two weavers to sit side by side, one throwing the shuttle from the right to the centre, the other reaching between the warps and sending it on its way to the left and then returning it to the centre. The stopping of the shuttle and the reaching between the warps caused imperfections in the cloth. Kay devised a mechanical attachment controlled by a cord jerked by the weaver that sent the shuttle flying through the shed. Jerking the cord in the opposite direction sent the shuttle on its return trip. Using the flying shuttle, one weaver could weave fabrics of any width more quickly than two could before. A more important virtue of Kay's invention, however, lay in its adaptability to automatic weaving.

Power-driven looms. The first power-driven machine for weaving fabric-width goods, patented in 1785 by Edmund Cartwright, an English clergyman, was inadequate because it considered only three motions: shedding, picking, and winding the woven cloth onto the cloth beam. Cartwright's second patent (1786) proved too ambitious, but his concept of a weaving machine became the basis for the successful power loom.

Early difficulties of the power loom

One of the great obstacles to the success of the power loom was the necessity to stop the loom frequently in order to dress (*i.e.,* apply sizing to) the warp, an operation that, like many others, had been done in proportionately reasonable time when the weaving was done by hand. With the power loom a second man had to be employed continuously to do this work, so there was no saving of expense or time. In the early 19th century a dressing machine was developed that prepared the warp after it had been wound onto the warp beam and as it was passed to the cloth beam. Although later superseded by an improved sizing apparatus, this device made the power loom a practical tool.

Advances made by William Horrocks of Scotland between 1803 and 1813 included an improvement in the method of taking up the cloth (*i.e.,* winding the woven fabric onto the cloth beam) and making a more compact machine of iron, requiring little space as compared with wooden handlooms.

Francis Cabot Lowell, of Boston, experimented with the power loom, adding improvements to increase the weaving speed, and also improved the dressing machine.

A valuable improvement was that of the let-off and take-up motions, to maintain uniform warp tension automatically. The principle of holding at the beat (*i.e.,* not permitting the warp to be let off until the pick was beaten into place) first applied by Erastus Brigham Bigelow in the carpet loom, was successfully applied to all kinds of weaving. Another Bigelow invention, applicable to power looms in general, although first used on a carpet loom, was the friction-brake stop mechanism, allowing the loom to be stopped without a shock.

These developments were primarily concerned with the power loom used for weaving plain goods. William Crompton, an English machinist working in the machine shop attached to a cotton factory in Massachusetts, undertook the development of a loom that could weave fancy goods, patented in both the U.S. and England in 1837. The loom was later much improved by his son George Crompton. Such 19th-century inventions made possible the production of textile goods for every use in great volume and variety, and at low cost.

Modern looms. Modern looms still weave by repeating in sequence the operations of shedding, picking, and beating in, but within that framework there has been considerable development during the 20th century. Several new types of loom have come into industrial use, whereas older types have been refined and their scope extended. Two main influences have been the rising cost of labour and the increasing use of man-made, continuous-filament yarns. The first has led to an increase in automatic control, in automatic handling of yarn packages, and in the use of larger packages; the second, to greater precision and finish in loom construction, because deficiency in these qualities is readily reflected in the quality of the cloth made from these yarns.

Modern looms can be grouped into two classes according to whether they produce cloth in plane or tubular form (see Figure 6). Looms of the first kind, comprising all but a few, are called flat looms; the others are described as

From E.M. Pritchard, *A Short Dictionary of Weaving;* reproduced by permission of The Philosophical Library, Inc.

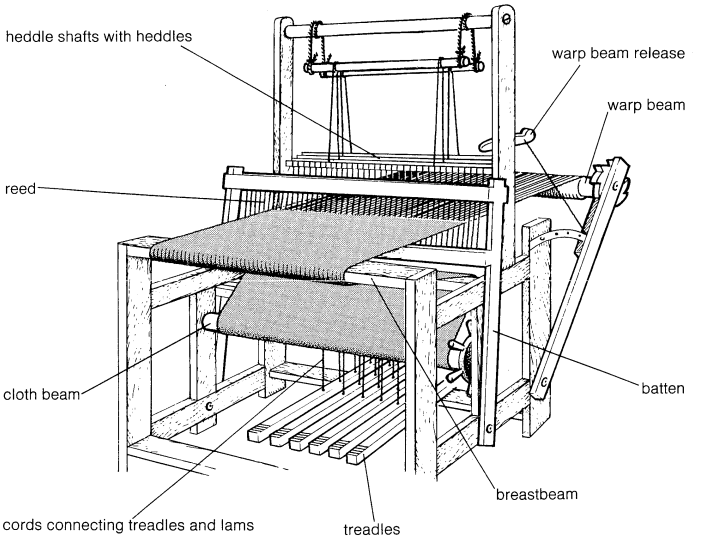


Figure 6: Principal parts of a modern handloom.

The two classes of modern looms

circular. Since the majority are flat looms, the adjective is used only when a distinction has to be drawn. Flat looms fall into two categories: those that employ a shuttle and those that draw the weft from a stationary supply, usually called shuttleless looms. (This term is not entirely satisfactory, as some primitive looms make no use of a shuttle, merely passing through the shed a stick with weft wound on it.) Shuttle looms fall into two groups according to whether the shuttle is replenished by hand or automatically. The second kind is often described as an automatic loom, but, except for shuttle replenishment, it is no more automatic in its operation than the hand-replenished or so-called nonautomatic loom, which, like all modern looms, is power-operated by electric motor. With both types of loom the actual weaving operation is entirely automatic and is performed in exactly the same manner.

Hand-replenished, or nonautomatic, looms are used only where particular circumstances—of yarns, fabrics, or use—make automatically replenished looms either technically unsuitable or uneconomic. Basically they differ little from the power looms of the latter half of the 19th century. They do not run appreciably faster but are better engineered, making use, for example, of machine-cut instead of cast gear wheels. Often there is no superstructure, which makes for cleanliness and improved illumination; frequently rigid heddle connectors are employed, leading to precise and stable setting of the shed; and usually the overpick mechanism has been replaced by the cleaner and safer underpick.

Automatically replenished flat, or automatic, looms are the most important class of modern loom, available for a very wide range of fabrics. In virtually all such looms, the shuttle is replenished by automatically replacing the exhausted bobbin with a full one. In principle they are thus the same as the automatic looms introduced at the end of the 19th century. Since that time, automatic shuttle-changing looms have also been introduced but have largely become obsolete, because bobbin-changing looms have been developed to a point where they can deal with most of the yarns for which it was once thought necessary to use shuttle-changing looms.

Apart from the general engineering refinements, automatic looms have advanced mainly in respect of the weft supply. Alternatives to the hand-replenished bobbin now exist in the form of the automatic bobbin loader, the loom being supplied with boxes of pirned (reeled) weft; and the automatic loom winder, the loom being fed with large cones of yarn, which is wound onto pirns at the loom. These alternatives are technically feasible and economic only with certain yarns. Therefore, all three types of weft supply continue to be used. An alternative to the rotary battery, when weft of more than one colour is used, is a series of vertical stacks.

The principle of shuttle replenishment is the same for all three systems. When the shuttle is stationary in the shuttle box, and the bar carrying the reed is farthest forward, a feeler enters the shuttle and senses whether the weft is on the point of exhaustion. Feelers may be mechanical or electrical, relying respectively on the change in friction or the change in electrical resistance brought about by the absence of weft. Alternatively, with delicate wefts, an optical feeler may be used that depends for its action on the change in the amount of light reflected when the bare pirn is revealed.

When the feeler has sensed that the bobbin is nearly empty, mechanical or electrical signals are transmitted to the transfer mechanism that, when the shuttle is appropriately positioned and momentarily at rest, both as regards warp-way and weft-way motion, hammers a new bobbin into position, simultaneously ejecting the empty one through the open base of the shuttle. The loom continues to run at its normal speed throughout.

In the course of this operation, there are created unwanted lengths of weft extending from the nearer selvage. These, if not controlled and disposed of, may find their way into the cloth and appear as defects. Modern looms supplement the earlier mechanical methods by pneumatic suction, with the result that the most delicate fabrics can be woven on automatic bobbin-changing looms without any loss of quality. To make certain of removal of the remnant of weft on the old bobbin, extending to the eye of the shuttle, a cutter moves forward into the shuttle box and cuts the weft close to the eye just before the bobbin is ejected.

High speed, often combined with the use of large and heavy shuttles, means that these modern looms are noisier than ever. The noise level in a typical textile mill is above the level at which deafness occurs following prolonged exposure.

Types of shuttleless looms

Shuttleless looms are of three kinds, of which the first predominates: dummy shuttle, rapier, and fluid jet. The dummy-shuttle type, the most successful of the shuttleless looms, makes use of a dummy shuttle, a projectile that contains no weft but that passes through the shed in the manner of a shuttle and leaves a trail of yarn behind it.

The rapier type conveys a pick of weft from a stationary package through the shed by means of either a single rapier or a pair of rapiers. Rapiers are either rigid rods or flexible steel tapes, which are straight when in the shed but on withdrawal are wound onto a wheel, in order to save floor space. Rapier looms are, on the whole, simpler and more versatile than dummy-shuttle looms, but they have failed to achieve such high rates of weft insertion, the maximum being not more than 400 yards (365 metres) per minute. They differ in respect of the number of rapiers employed and the type of selvage provided; some of them operate by gripping the free end of the weft and conveying that through the shed rather than by starting with a loop. Fluid-jet looms, most recently developed of the shuttleless types, are produced and used on a much smaller scale than the two other types described above. They are of two kinds, one employing a jet of air, the other a water jet, to propel a measured length of weft through the shed. The significance of this development is that for the first time nothing solid is passed into the shed other than the weft, which eliminates the difficulties normally associated with checking and warp protection, and reduces the noise to an acceptable level.

In addition to those looms that have established themselves industrially, there are looms still in the experimental stage. Loom development is always slow: some of the looms just gaining favour had their origins in inventions made 50 or even 100 years earlier. The most intense activity is in the field of shuttleless looms, because these offer the greatest prospect of achieving increased rates of weft insertion and of avoiding the drawbacks of noise, danger, vibration, high power consumption, and wear attendant on the use of a shuttle. The ultimate in direct projection is a method, still experimental, in which the weft is projected longitudinally at high speed and traverses the warp under its own momentum, nothing entering the shed but the pick of weft. The name inertial has been given to this method. Another experimental loom employs multiple rapiers for weft insertion and, in addition, eliminates the heddles and the reed.

Basic weaves. The basic weaves include plain (or tabby), twills, and satins (see Figure 7).

From Evelyn E. Stout, *Introduction to Textiles* (© 1970); John Wiley & Sons, Inc.

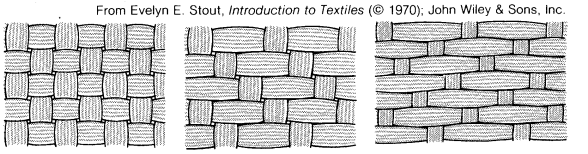


Figure 7: *Types of weaves:*
(Left) plain, (centre) twill, (right) satin.

Plain weave. Plain, or tabby, weave, the simplest and most common of all weaves, requires only two harnessses and has two warp and weft yarns in each weave unit. To produce it, the warp yarns are held parallel under tension while a crosswise weft yarn is shot over and under alternate warps across the width of the web. The weave unit is completed at the end of the second row, when the weft has been inserted over and under the opposite set of warps, thus locking the previous weft in place. Fabric length is increased with the insertion of each succeeding weft yarn. When warp and weft yarns are approximately equal in size and quantity, the finished fabric is balanced and potentially stronger than cloth made of the same kind and number of warp and weft yarns in any other basic weave. Tabby woven with different-sized warp and weft yarns results in such fabrics as taffeta and poplin, in which many fine warps are interlaced with proportionately fewer thick weft yarns to form cloths with crosswise ridges or ribs.

The term extended tabby describes any weave in which two or more warps or wefts, or both, are interlaced as a unit. The group includes fabrics with basketry effects and fabrics with ribs formed by groups of warps or wefts in each shed.

Tapestry weave

Tapestry weave is a tabby in which a variety of coloured weft yarns is interlaced with the warp to form patterns. It is usually an unbalanced weave, with wefts completely covering a proportionately low number of warps. These cloths are sturdy and compact. Although they are flat and generally do not drape well, they have been used for centuries to make ceremonial and decorative dress and costumes.

Twill weave. Twill weave is distinguished by diagonal lines. The simplest twill is that created by the weft crossing over two warp yarns, then under one, the sequence being repeated in each succeeding shot (pick), but stepped over, one warp either to the left or right. Twills with more warps than wefts floating on the fabric's face are called warp faced; those with wefts predominating, weft faced. The angle of the twill can also vary.

Twills can be varied by changing the relative number of warps and wefts in each repeat (2:1, 2:3, 3:1, 6:2, etc.); by stepping the repeat in one direction; by breaking the direction of the diagonals formed by the twill at regular intervals; by reversing the direction of the diagonal at regular intervals to form chevrons or lozenges; or by combining several twills or modifying them to create a pattern.

Twills drape better than plain weaves with the same yarn count because twills have fewer interlacings. Twill weaves have been used throughout history in many weights and textures, from wool serges mentioned in medieval French manuscripts to English diapered (diamond patterned) table linens, patterned bed coverlets, and Indian shawls.

Satin weave. Although satin-weave drafts superficially resemble those of twills, satin weave does not have the regular step in each successive weft that is characteristic of twills. Thus, there is no strong diagonal line, and the fabric is smooth faced, with an unbroken surface made up of long floating warp yarns. A true satin must have at least five warp and weft yarns in each complete weave repeat and thus requires at least five harnesses. Most satin fabrics are made of smooth, lightly twisted yarns that heighten the effect of light unbroken by visible crosswise bindings. The limited number of interlacings allows the weaver to use a proportionately large number of warp yarns and thus produce a heavy textured cloth that can be arranged in smooth, shadowed folds. Satins, having long floats, are susceptible to the wear caused by rubbing and snagging and are, therefore, generally regarded as luxury fabrics.

Damask

Among the variations of satin weave are damask and sateen, a weft-faced satin. Damask is the most important variation of basic satin weave. Classic damask is a patterned, solid-coloured fabric with figures in warp-faced satin and background in weft-faced satin weave. The pattern is created by the difference in light reflection between the warp-faced and weft-faced areas. Silk damasks probably originated in China and came to Europe through Italy, the centre of European silk manufacture between the 13th and 17th centuries. During this period drawloom weavers from the Netherlands and Belgium also developed the art of linen damask weaving. Pictorial linen damasks, unlike most silk damasks of the time, often consisted of a single large repeat, picturing biblical scenes, contemporary events, or the arms of nobles and kings.

Complex weaves. Complex weaves include multiple plane, pile, inlaid, Jacquard, dobby, and gauze (or leno) weaves.

Multiple plain weave. Reversible double-woven cloth is produced by multiple plain weaving. It is woven in two layers, which may be completely independent, may be joined at one or both selvages, may be held together along the edges of a pattern, or may be united by a separate binding weft. Though often tabby weave is employed on both surfaces, any of the basic weaves may be used, depending on the intended use of the fabric.

Double-woven cloths have been used for clothing, but, though warm, they tend to be heavy and to drape poorly. They are most often used as bedcovers or wall hangings. German 18th-century *Beiderwand* is an example of antique double-woven cloth consisting of two layers of tabby weave joined only along the edges of the pattern. A dark-coloured pattern in one layer is set against the light-coloured ground of the other layer; the pattern is seen in negative or the reverse side of the cloth.

Nonreversible cloth with two or more sets of warp and sometimes of weft can also be produced. These cloths have an intricately patterned face, and all warps and wefts that do not appear on the face are carried along and bound into the web on the reverse side. This class includes important historic textiles, such as early Persian and Byzantine figured fabrics, as well as more recent Jacquard-woven imitation tapestries and a wide range of imitation brocaded fabrics.

Pile weave. Pile weaves have a ground fabric plus an extra set of yarns woven or tied into the ground and projecting from it as cut ends or loops. A great range of textures is included in this binding system, from terry pile towelling and corduroy to silk velvets and Oriental rugs.

In warp-pile fabrics the pile is formed by an extra set of warp yarns. To create such a fabric, first one set (sheet) of ground warps is raised, and the weft makes its first interlacing with the ground warp. Next, pile warps are raised, and a rod is inserted through the entire width of the web. The remaining ground warps are raised to form the third shed; then the ground weft is shot across again. This sequence is repeated several times; then the rods are slipped out, leaving a warp pile. To form cut-pile velvet, a knife on the end of the rod cuts the pile warps it passes, creating two fine rows of cut pile. Although the system has many technical variations, the same basic process can be applied to most warp-pile weaving.

If the pile is not cut when the rod is removed, a loop pile fabric results. In weaving terry pile fabrics, the ground warp is under tension, and the pile warp stays slack. When wefts are beaten in, the slack yarns are pushed into loops on both sides of the cloth.

To make velvets by double-cloth construction, two layers of cloth are woven simultaneously face-to-face, with long pile warp yarns connecting the two layers. After the cloth is woven, a knife slices the two layers apart.

Corduroy and velveteen are weft-pile constructions. Weft yarns having long floats are inserted between ground-weave picks. The floats are slit longitudinally after the fabric is completed, thus forming a ribbed surface of cut pile. In manufacture of velveteen the floats are formed over the whole surface of the fabric and cut evenly to imitate velvet.

Hand-knotted Oriental and Scandinavian rugs are constructed on a tabby-weave ground, with each row of knots followed by tightly beaten-in wefts. The pile of fine Oriental rugs may contain 160 knots per inch, thus completely obscuring the knots in the rug's foundation.

Inlaid weave. In all of the fabrics of this class, designs are created by inserting pattern warp or weft yarns between ground warps or wefts.

Brocaded fabric

Brocaded fabric has a pattern of coloured or metallic threads, or both, set in low relief against the ground weave. The ground weave can be any basic weave, since the brocaded pattern is merely inserted between ground wefts and is bound by ground warps. Until the advent of the Jacquard mechanism in the early 19th century, brocaded fabrics were woven by drawloom weavers who inserted the pattern wefts by hand. These weft yarns were wound on small brocading shuttles that travelled across the width of each pattern repeat, a separate shuttle being used for each colour in the repeat. Generally, these extra wefts were found only in the area in which the pattern was located and usually formed long floats on the reverse side of the fabric.

A mechanical process closely corresponding to hand brocading is called swivel, a system of figuring fabrics by using mechanically controlled pattern shuttles. The figures, inserted between ground-weft picks, interlace with the warp. The lappet system produces figured fabrics resembling those made by swivel figuring, but the pattern yarns are extra warps (rather than wefts) brought into play from separate warp beams. Lappet weaving is generally confined to coarse pattern yarns and can be distinguished from swivel by its interlacing with weft rather than with warp yarns.

Jacquard weave. The Jacquard weave, used to make allover figured fabrics such as brocades, tapestries, and damasks, is woven on a loom having a Jacquard attachment to control individual warps. Fabrics of this type are costly because of the time and skill involved in making the Jacquard cards, preparing the loom to produce a new pattern, and the slowness of the weaving operation. The Jacquard weave usually combines two or more basic weaves, with different weaves used for the design and the background.

Dobby weaves. Dobby weaves also produce allover figured fabrics. They are made on looms having a dobby attachment, with narrow strips of wood instead of Jacquard cards. Dobby weaves are limited to simple, small geometric figures, with the design repeated frequently, and are fairly inexpensive to produce.

Gauze or leno weave. Gauze weaving is an open weave made by twisting adjacent warps together. It is usually made by the leno, or doup, weaving process, in which a doup attachment, a thin hairpin-like needle attached to two healds, is used, and the adjacent warp yarns cross each other between picks. Since the crossed warps firmly lock each weft in place, gauze weaves are often used for sheer fabrics made of smooth fine yarns. Although gauze weaving, with its multitude of variations, has been adapted to modern production, it is an ancient technique.

KNITTED FABRICS

Knitted fabrics are constructed by the interlocking of a series of loops made from one or more yarns, with each row of loops caught into the preceding row. Loops running lengthwise are called wales, those running crosswise

courses. Hand knitting probably originated among the nomads of the Arabian Desert about 1000 BC and spread from Egypt to Spain, France, and Italy. Knitting guilds were established in Paris and Florence by the later Middle Ages. Austria and Germany produced heavily cabled and knotted fabrics, embroidered with brightly coloured patterns. In The Netherlands, naturalistic patterns were worked on fabric in reverse stocking stitch, and several Dutch knitters went to Denmark to teach Danish women the Dutch skills. The craft of hand knitting became less important with the invention of a frame knitting machine in 1589, although the production of yarns for hand knitting has remained an important branch of the textile industry to the present day.

The frame knitting machine allowed production of a complete row of loops at one time. The modern knitting industry, with its highly sophisticated machinery, has grown from this simple device.

lomen-
lature of
nitted
abrics

Knitted fabrics were formerly described in terms of the number of courses and wales per unit length and the weight of the fabric per unit area. This system is limited, however, and there is a shift to use of the dimensions and configuration of the single loop, the repeating unit determining such fabric characteristics as area, knitting quality, and weight. The length of yarn knitted into a loop or stitch is termed the stitch length, and in a plain knitted structure this is related to the courses per inch, wales per inch, and stitch density. The two basic equilibrium states for knitted fabrics are the dry-relaxed state, attained by allowing the fabric to relax freely in the air, and the wet-relaxed state, reached after static relaxation of the fabric in water followed by drying.

Knitting machines. The needle is the basic element of all knitting machines. The two main needle types are the "bearded" spring needle, invented about 1589, and the more common latch needle, invented in 1847.

The bearded needle, made from thin wire, has one end bent, forming an operating handle; the other end is drawn out and bent over, forming a long flexible tipped hook resembling a beard. A smooth groove, or eye, is cut in the stem or shank of the needle just behind the tip. In use this needle requires two other units, a sinker to form a loop and a presser to close the needle beard, allowing the loop to pass over the beard when a new stitch is formed. Bearded needles can be made from very fine wire and are used to produce fine fabrics.

The latch needle is composed of a curved hook, a latch, or tumbler, that swings on a rivet just below the hook, and the stem, or butt. It is sometimes called the self-acting needle because no presser is needed; the hook is closed by the pressure of a completed loop on the latch as it rises on the shaft. Needles differ greatly in thickness, in gauge, and in length, and appropriate types must be selected for specific purposes. A 4-gauge needle, for example, is used for heavy sweaters, but an 80-gauge needle is required for fine hosiery.

Weft knitting. The type of stitch used in weft knitting affects both the appearance and properties of the knitted fabric. The basic stitches are plain, or jersey (see Figure 8); rib; and purl. In the plain stitch, each loop is drawn through others to the same side of the fabric. In the rib stitch, loops of the same course are drawn to both sides of the fabric. The web is formed by two sets of needles, arranged opposite to each other and fed by the same thread, with each needle in one circle taking up a position between its counterparts in the other. In a 2:2 rib, two needles on one set alternate with two of the other. The interlock structure is a variant of the rib form in which two threads are alternately knitted by the opposite needles so that interlocking occurs. In the purl stitch, loops are drawn to opposite sides of the fabric, which, on both sides, has the appearance of the back of a plain stitch fabric. Jacquard mechanisms can be attached to knitting machines, so that individual needles can be controlled for each course or for every two, and complicated patterns can be knitted. To form a tuck stitch, a completed loop is not discharged from some of the needles in each course, and loops accumulating on these needles are later discharged together. The plaited stitch is made by feeding two threads into the same hook, so that one thread shows on the one side of the fabric and the other on the opposite side. A float stitch is produced by missing interlooping over a series of needles so that the thread floats over a few loops in each course.

From Isabel B. Wingate, *Textile Fabrics and Their Selection*, 6th ed. (© 1970), Prentice-Hall, Inc.

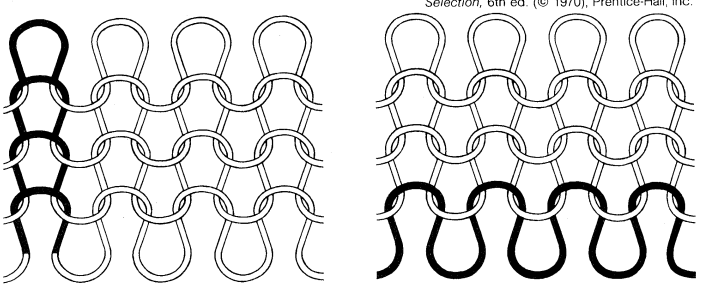


Figure 8: *Stitches in a plain weft-knitted fabric.*
(Left) A wale in a plain, circular-knit fabric. (Right) A course in a plain circular-knit fabric.

Types of knitting machines

Knitting machines can be flat or circular. Flat machines have their needles mounted in a flat plate or needle bed or in two beds at right angles to each other and each at a 45° angle to the horizontal. The knitted fabric passes downward through the space between the upper edges of the plates, called the throat. In the knitting process, the needles are pushed up and down by cams attached to a carriage with a yarn guide, which moves over the length of the machine. The width of the fabric can be altered by increasing or decreasing the number of active needles, allowing production of shaped fabrics, which when sewn together make fully fashioned garments. Although flatbed machines are suited for hand operation, they are power driven in commercial use, and, by selection of colour, type of stitch, cam design, and Jacquard device, almost unlimited variety is possible. The cotton frame, designed to knit fine, fully fashioned goods, shaped for improved fit of such items as hosiery and sweaters, is fitted with automatic narrowing and widening devices.

Circular machine needles are carried in grooves cut in the wall of a cylinder, which may be as small as one centimetre (0.4 inch) in diameter and as large as 1.5 metres (five feet). Some circular machines have two sets of needles, carried in concentric cylinders, so that the needles interlock. During the knitting operation the butts of the needles move through cam tracks, the needles sliding up and down to pick up yarn, form a new loop, and cast off the previously formed loop. In the least complicated of these machines, yarn is supplied from one package, each needle picking up the yarn once per revolution of the cylinder. Modern machines may have as many as 100 feeders, allowing each needle to pick up 100 threads per revolution. Both latch and spring needles are used, with the former more common. Modern, large, circular, plain or jersey machines having 90–100 feeders are frequently used to produce medium-weight fabric. Small bladelike units, or sinkers, are inserted between every two needles to engage and hold the completed fabric, preventing it from riding up with the needles as they are lifted to form new stitches. Machines may be fitted with pattern wheels controlling needle action to produce tuck and float stitches, and a Jacquard mechanism may also be attached. Stop motions are essential to stop the machine when a thread breaks. Because yarn tension affects fabric uniformity, various tension controllers have been devised. An alternative method, positive feed, feeding precisely measured amounts of yarn into the machine, is now considered more satisfactory.

Circular rib machines consist of a vertical cylinder, with needle slots on the outside, and a horizontal bed in the form of a circular plate or dial with needle slots cut radially, so that the two sets of needles are arranged at right angles to each other.

Seamless hosiery, knitted in tubular form, is produced by circular knitting machines. Modern hosiery machines, such as the Komet machine employ double-hooked needles directly opposite each other in the same plane to knit the leg and foot portions, the heel and the toe. The toe is

later closed in a separate operation. In the Getaz toe, the seam is placed under the toes instead of on top of them.

Underwear fabrics are usually knitted on circular machines, and—except for fully fashioned underwear, tights, and leotards, which are knitted to pattern and sewn together—underwear making is a cut, make, and trim operation. Tights or panty hose are a combination of hosiery and underwear and can be fully fashioned. Seamless panty hose are made on circular hose machines modified to make very long stockings with open tops, two of which are cut open at opposite sides and seamed together front and back. The wearing quality and fit of modern panty hose have been greatly improved with the development of stretch nylon and spandex, and greater variety has been introduced with the development of texturized yarn.

Much hosiery is finished by washing, drying, and a boarding process in which the hosiery is drawn over a thin metal or wooden form of appropriate shape and pressed between two heated surfaces. The introduction of nylon fibre led to the development of a preboarding process, setting the loops and the fabric in the required shape before dyeing and finishing. The article, fitted on a form of appropriate shape, is placed in an autoclave or passed through a high-temperature setting unit. Fabric treated in this way does not distort during dyeing.

Circular knitting machines can be adapted to make simulated furs. One type intermeshes plush loops with the plain-stitch base fabric then cuts the loops, producing a pile. A more common method forms the pile with a carded sliver. A plain-stitch fabric is used as the base and loose fibres from a sliver, fed from a brushing or carding device, are inserted by a V-shaped claw, forming the pile. Pile depth is determined by the length of the fibres in the sliver.

One of the most sophisticated knitting machines incorporates electronic selection of sinkers in a Jacquard circular knitting machine.

Warp knitting. The two types of warp knitting are raschel, made with latch needles, and tricot, using bearded needles.

Raschel. Coarser yarns are generally used for raschel knitting, and there has recently been interest in knitting staple yarns on these machines. In the Raschel machine, the needles move in a ground steel plate, called the trick plate. The top of this plate, the verge, defines the level of the completed loops on the needle shank. The loops are prevented from moving upward when the needle rises by the downward pull of the fabric and the sinkers between the needles. Guide bars feed the yarn to the needles. In a knitting cycle, the needles start at the lowest point, when the preceding loop has just been cast off, and the new loop joins the needle hook to the fabric. The needles rise, while the new loop opens the latches and ends up on the shank below the latch. The guide bars then swing through the needles, and the front bar moves one needle space sideways. When the guide bar swings back to the front of the machine, the front bar has laid the thread on the hooks. The needles fall, the earlier loops close the latch to trap the new loops, and the old loops are cast off. Raschels, made in a variety of forms, are usually more open in construction and coarser in texture than are other warp knits.

Tricot. Tricot, a warp knit made with two sets of threads, is characterized by fine ribs running vertically on the fabric face and horizontally on its back. The tricot knitting machine makes light fabrics, weighing less than four ounces per square yard. Its development was stimulated by the invention of the so-called FNF compound needle, a sturdy device that later fell into disuse but that made possible improved production speeds. Although approximately half of the tricot machines in current use make plain fabrics on two guide bars, there is increasing interest in pattern knitting. In this type of knitting, the warp-knitting cycle requires close control on the lateral bar motion, achieved by control chains made of chunky metal links.

Special effects in warp knits. The scope of warp knitting has been extended by the development of procedures for laying in nonknitted threads for colour, density, and texture effects (or inlaying), although such threads may also be an essential part of the structure. For example, in the form called "zigzagging across several pillars," the ground of most raschel fabrics, the front bar makes crochet chains, or "pillars," which are connected by zigzag inlays.

An extension of conventional warp knitting is the Co-We-Nit warp-knitting machine, producing fabrics with the properties of both woven and knitted fabrics. The machines need have only two warp-forming warps and provision for up to eight interlooped warp threads between each chain of loops. These warp threads are interlaced with a quasiweft, forming a fabric resembling woven cloth on one side.

OTHER INTERLACED FABRICS

Net and lace making. The popularity of handmade laces led to the invention of lace-making machines. The early models required intricate engineering mechanisms, and the development of the modern lace industry originated when a machine was designed to produce laces identical with Brussels lace. In the Heathcot, or bobbinet, machine, warp threads were arranged so that the threads moved downward as the beams unwound. Other threads were wound on thin, flat spools or bobbins held in narrow carriages that could move in a groove or comb in two rows. The carriages carrying the bobbins were placed on one side of the vertical warp threads and given a pendulum-like motion, causing them to pass between the warp threads. The warp threads were then moved sideways, so that on the return swing each bobbin thread passed around one of them. Then the warp threads moved sideways in the opposite direction, thus completing a wrapping movement. In addition, each row of bobbins was moved by a rack-and-pinion gearing, one row to the left and one to the right. As these movements continued, the threads were laid diagonally across the fabric as the warp was delivered. Improvements on the Heathcot machine followed through the 19th century: Nottingham-lace machines, used primarily for coarse-lace production, employ larger bobbins, and the pattern threads are wound independently on section spools; in another type, the Barmens machine, threads on king bobbins on carriers are plaited together, sometimes with warp threads.

Early machine-made laces

Schiffli lace, a type of embroidery, is made by modern machines, evolved from a hand version, using needles with points at each end. Several hundred needles are placed horizontally, often in two rows, one above the other. The fabric to be embroidered is held vertically in a frame extending the full width of the machine, and the needles, supplied with yarn from individual spools, move backward and forward through the fabric. At each penetration a shuttle moves upward and interlaces yarn with the needle loop. Movement of both fabric and needles is controlled by Jacquard systems.

Many types of machine-made laces are made, frequently with geometrically shaped nets forming their backgrounds. Formerly made only of cotton, they are now frequently made from man-made fibre yarns. Bobbinet lace, essentially a hexagonal net, is used as a base for appliqué work for durable non-run net hosiery, and, when heavily sized, for such materials as millinery and veilings. Barmens lace has a fairly heavy texture and an angular pattern; flowing lines, heavy outline cords, and fine net backgrounds are not usually made on Barmens machines.

The introduction of light-resistant polyester yarns led to a revival of Nottingham machine-made curtains. Leavers lace is available in an infinite variety of patterns, since the manufacturing technique allows use of almost any type of yarn. The high strength and comparatively low cost of man-made fibre yarns has made sheer laces widely available.

Net, an open fabric having geometrically shaped, open meshes, is produced with meshes ranging from fine to large. Formerly made by hand, the various types are now made on knitting machines. Popular types include bobbinet, made with hexagonal-shaped mesh and used for formal gowns, veils, and curtains, and tulle, a closely constructed, fine net having similar uses. Fishnet, a coarse type with knots in four corners forming the mesh formerly made by fishermen, is now a popular machine-made curtain fabric.

Braiding or plaiting. Braid is made by interlacing three or more yarns or fabric strips forming a flat or tubular narrow fabric. It is used as trimming and for belts and is also sewn together to make hats and braided rugs. Plaiting, usually used synonymously with braiding, may be used in a more limited sense, applying only to a braid made from such materials as rope and straw.

NONINTERLACED FABRICS

With the exception of felt, nonwoven materials are in the early stages of development. There is controversy about the precise meaning of the term nonwoven, but one authority defines nonwoven fabrics as textile fabrics made of a fibrous layer having randomly laid or oriented fibres or threads.

Felt. Felts are a class of fabrics or fibrous structures obtained through the interlocking of wool, fur, or some hair fibres under conditions of heat, moisture, and pressure. Other fibres will not felt alone but can be mixed with wool, which acts as a carrier. Three separate industries manufacture goods through the use of these properties. The goods produced are wool felt, in rolls and sheets; hats, both fur and wool; and woven felts, ranging from thin billiard tablecloths to heavy industrial fabrics used for dewatering in the manufacture of paper. Felts of the nonwoven class are considered to be the first textile goods produced, and many references may be found to felts and their uses in the histories of ancient civilizations. The nomadic tribes of north central Asia still produce felts for clothing and shelter, utilizing the primitive methods handed down from antiquity.

Bonding. Several methods for making nonwoven materials are now firmly established, and others are being developed.

Adhesive and mechanical bonding

In adhesive bonding, fabrics are made by forming a web of fibres, applying an adhesive, then drying and curing the adhesive. The web can be produced by a garnett machine or a conventional card, several layers being piled up to obtain the required thickness. Such webs are weak across the width, but this does not limit their use for certain end products. A more uniform product results from cross laying the web. Other machines, such as the Rando-Webber, lay down the fibres by an airstream.

The fibres in the web may be stuck together in various ways. The web may be sprayed with an emulsion of an adhesive—*e.g.*, a latex based on synthetic rubber, acrylic derivatives, or natural rubber—or, alternatively, may be carried on a mesh screen through a bath of latex, the excess being squeezed out by a pair of rollers. Adhesives may also be applied as a foam or a fine powder. Thermoplastic fibres can be incorporated in the blend and on heating will bond together, giving strength to the mass of fibres.

Mechanically bonded nonwoven products (or fibre-bonded nonwovens) are webs strengthened by mechanical means. The web, sometimes reinforced by a thin cotton scrim in the middle or by texturized yarns distributed lengthwise through it, is punched by barbed needles mounted in a needle board. The fibres in the web are caught up by the needle barbs, and the resulting increased entanglement yields a compact product sufficiently strong for many purposes. Modern needle-felting or punching machines perform 900 punches per minute, and selection of appropriate needles is based on the fibre being processed and the desired product.

Stitch bonding

The Arachne machine, the best known unit for stitch bonding, operates much like a warp-knitting machine. Fibrous web is fed into the machine, and stitches are made by a series of needles placed about eight millimetres apart, giving the web longitudinal strength; lateral strength is provided by the fibre interactions. The products are attractive for many purposes and can be improved by treatment with polyester resins to increase their wear resistance and with thermosetting precondensates to reduce their tendency to pill (*e.g.*, to form small tangles). A new device attached to the Arachne machine permits introduction of weft ends at every single course, making colour effects possible. Araloop machines yield loop-pile fabric suitable for towels and floor coverings.

Three sewing-knitting machines were invented in East Germany in 1958. In the Malimo machine process, warp yarns are placed on top of filling yarns and stitched together by a third yarn. The Maliwatt machine interlaces a web of fibres with a sewing thread, giving the effect of parallel seams. The Malipol machine produces a one-sided pile fabric by stitching loop pile through a backing fabric. A new British process makes double-sided terry fabric, called Terrytuft, by inserting pile yarn into a backing and knotting it into position.

Webs made of yarns having a core of one polymer and an outer sheath of another material having a lower softening point may be lightly pressed and then heated to an appropriate temperature. The core yarn will "spot weld" together at the junction points, binding the mass of fibres together. Products made in this way find uses as industrial fabrics, coatings, and interlinings.

Laminating. The joining of one fabric to another by an adhesive such as natural rubber has long been practiced in rainwear manufacture. Composite materials were later joined by bonding a layer of polyurethane or other foam to a conventional textile fabric. The two components were stuck together by flame bonding or by an adhesive in the form of a continuous coating, in spots, or as a powder. This laminating process has been extended to the joining of two layers of fabric. Each fabric layer can be quite thin, and the amount and type of adhesive are chosen to add only minimum stiffening. Such materials offer a variety of applications. A coating fabric, for example, may be joined to a lining; dimensionally stable composites can be made from cloth layers that are in themselves dimensionally unstable. Acetate knitted fabrics are frequently used as backing material in laminates.

Textile finishing processes

BASIC METHODS AND PROCESSES

The term finishing includes all the mechanical and chemical processes employed commercially to improve the acceptability of the product, except those procedures directly concerned with colouring. The objective of the various finishing processes is to make fabric from the loom or knitting frame more acceptable to the consumer. Finishing processes include preparatory treatments used before additional treatment, such as bleaching prior to dyeing; treatments, such as glazing, to enhance appearance; sizing, affecting touch; and treatments adding properties to enhance performance, such as preshrinking. Newly formed cloth is generally dirty, harsh, and unattractive, requiring considerable skill for conversion into a desirable product. Before treatment, the unfinished fabrics are referred to as gray goods, or sometimes, in the case of silks, as greige goods.

Finishing formerly involved a limited number of comparatively simple operations evolved over the years from hand methods. The skill of English and Scottish finishers was widely recognized, and much British cloth owed its high reputation to the expertise of the finisher. More sophisticated modern finishing methods have been achieved through intense and imaginative research.

Preparatory treatments. It is frequently necessary to carry out some preparatory treatment before the application of other finishing processes to the newly constructed fabric. Any remaining impurities must be removed, and additives used to facilitate the manufacturing process must also be removed. Bleaching may be required to increase whiteness or to prepare for colour application. Some of the most frequently used preparatory processes are discussed below.

Burling and mending. Newly made goods, which frequently show imperfections, are carefully inspected, and defects are usually repaired by hand operations. The first inspection of woollen and worsted fabrics is called perching. Burling, mainly applied to woollen, worsted, spun rayon, and cotton fabrics, is the process of removing any remaining foreign matter, such as burrs and, also, any loose threads, knots, and undesired slubs. Mending, frequently necessary for woollens and worsteds, eliminates such defects as holes or tears, broken yarns, and missed warp or weft yarns.

Scouring. When applied to gray goods, scouring removes substances that have adhered to the fibres during production of the yarn or fabric, such as dirt, oils, and any sizing or lint applied to warp yarns to facilitate weaving.

Bleaching. Bleaching, a process of whitening fabric by removal of natural colour, such as the tan of linen, is usually carried out by means of chemicals selected according to the chemical composition of the fibre. Chemical bleaching is usually accomplished by oxidation, destroying colour by the application of oxygen, or by reduction, removing colour by hydrogenation. Cotton and other cellulosic fibres are usually treated with heated alkaline hydrogen peroxide; wool and other animal fibres are subjected to such acidic reducing agents as gaseous sulfur dioxide or to such mildly alkaline oxidizing agents as hydrogen peroxide. Synthetic fibres, when they require bleaching, may be treated with either oxidizing or reducing agents, depending upon their chemical composition. Cottons are frequently scoured and bleached by a continuous system.

Mercerization. Mercerization is a process applied to cotton and sometimes to cotton blends to increase lustre (thus also enhancing appearance), to improve strength, and to improve their affinity for dyes. The process, which may be applied at the yarn or fabric stage, involves immersion under tension in a caustic soda (sodium hydroxide) solution, which is later neutralized in acid. The treatment produces permanent swelling of the fibre.

Drying. Water, used in various phases of textile processing, accumulates in fabrics, and the excess moisture must eventually be removed. Because evaporative heating is costly, the first stage of drying uses mechanical methods to remove as much moisture as possible. Such methods include the use of centrifuges and a continuous method employing vacuum suction rolls. Any remaining moisture is then removed by evaporation in heated dryers. Various types of dryers operate by conveying the relaxed fabric through the chamber while festooned in loops, using a frame to hold the selvages taut while the fabric travels through the chamber, and passing the fabric over a series of hot cylinders. Because overdrying may produce a harsh hand, temperature, humidity, and drying time require careful control.

Finishes enhancing appearance. Treatments enhancing appearance include such processes as napping and shearing, brushing, singeing, beetling, decating, tentering, calendering or pressing, moiréing, embossing, creping, glazing, polishing, and optical brightening.

Napping and shearing. Napping is a process that may be applied to woollens, cottons, spun silks, and spun rayons, including both woven and knitted types, to raise a velvety, soft surface. The process involves passing the fabric over revolving cylinders covered with fine wires that lift the short, loose fibres, usually from the weft yarns, to the surface, forming a nap. The process, which increases warmth, is frequently applied to woollens and worsteds and also to blankets.

Shearing cuts the raised nap to a uniform height and is used for the same purpose on pile fabrics. Shearing machines operate much like rotary lawn mowers, and the amount of shearing depends upon the desired height of the nap or pile, with such fabrics as gabardine receiving very close shearing. Shearing may also be applied to create stripes and other patterns by varying surface height.

Brushing. This process, applied to a wide variety of fabrics, is usually accomplished by bristle-covered rollers. The process is used to remove loose threads and short fibre ends from smooth-surfaced fabrics and is also used to raise a nap on knits and woven fabrics. Brushing is frequently applied to fabrics after shearing, removing the cut fibres that have fallen into the nap.

Singeing. Also called gassing, singeing is a process applied to both yarns and fabrics to produce an even surface by burning off projecting fibres, yarn ends, and fuzz. This is accomplished by passing the fibre or yarn over a gas flame or heated copper plates at a speed sufficient to burn away the protruding material without scorching or burning the yarn or fabric. Singeing is usually followed by passing the treated material over a wet surface to assure that any smoldering is halted.

Beetling. Beetling is a process applied to linen fabrics and to cotton fabrics made to resemble linen to produce a hard, flat surface with high lustre and also to make texture less porous. In this process, the fabric, dampened and wound around an iron cylinder, is passed through a machine in which it is pounded with heavy wooden mallets.

Decating. Decating is a process applied to woollens and worsteds, man-made and blended fibre fabrics, and various types of knits. It involves the application of heat and pressure to set or develop lustre and softer hand and to even the set and grain of certain fabrics. When applied to double knits it imparts crisp hand and reduces shrinkage. In wet decating, which gives a subtle lustre, or bloom, fabric under tension is steamed by passing it over perforated cylinders.

Tentering, crabbing, and heat-setting. These are final processes applied to set the warp and weft of woven fabrics at right angles to each other, and to stretch and set the fabric to its final dimensions. Tentering stretches width under tension by the use of a tenter frame, consisting of chains fitted with pins or clips to hold the selvages of the fabric, and travelling on tracks. As the fabric passes through the heated chamber, creases and wrinkles are removed, the weave is straightened, and the fabric is dried to its final size. When the process is applied to wet wools it is called crabbing; when applied to synthetic fibres it is sometimes called heat-setting, a term also applied to the permanent setting of pleats, creases, and special surface effects.

Calendering. Calendering is a final process in which heat and pressure are applied to a fabric by passing it between heated rollers, imparting a flat, glossy, smooth surface. Lustre increases when the degree of heat and pressure is increased. Calendering is applied to fabrics in which a smooth, flat surface is desirable, such as most cottons, many linens and silks, and various man-made fabrics. In such fabrics as velveteen, a flat surface is not desirable, and the cloth is steamed while in tension, without pressing. When applied to wool, the process is called pressing, and employs heavy, heated metal plates to steam and press the fabric. Calendering is not usually a permanent process.

Moiréing, embossing, glazing and ciréing, and polishing. These are all variations of the calendering process. Moiré is a wavy or "watered" effect imparted by engraved rollers that press the design into the fabric. The process, applied to cotton, acetate, rayon, and some ribbed synthetic fabrics, is only permanent for acetates and resin-treated rayons. Embossing imparts a raised design that stands out from the background and is achieved by passing the fabric through heated rollers engraved with a design. Although embossing was formerly temporary, processes have now been developed to make this effect permanent.

Glazing imparts a smooth, stiff, highly polished surface to such fabrics as chintz. It is achieved by applying such stiffeners as starch, glue, shellac, or resin to the fabric and then passing it through smooth, hot rollers that generate friction. Resins are now widely employed to impart permanent glaze. Ciré (from the French word for waxed) is a similar process applied to rayons and silks by the application of wax followed by hot calendering, producing a high, metallic gloss. Ciré finishes can be achieved without a sizing substance in acetates, which are thermoplastic (*e.g.*, can be softened by heat), by the application of heat.

Polishing, used to impart sheen to cottons without making them as stiff as glazed types, is usually achieved by mercerizing the fabric and then passing it through friction rollers.

Creping. A crepe effect may be achieved by finishing. In one method, which is not permanent, the cloth is passed, in the presence of steam, between hot rollers filled with indentations producing waved and puckered areas. In the more permanent caustic soda method, a caustic soda paste is rolled onto the fabric in a patterned form; or a resist paste may be applied to areas to remain unpuckered and the entire fabric then immersed in caustic soda. The treated areas shrink, and the untreated areas pucker. If the pattern is applied in the form of stripes, the effect is called plissé; an allover design produces blister crepe.

Optical brightening. Optical brightening, or optical bleaches, are finishes giving the effect of great whiteness

and brightness because of the way in which they reflect light. These compounds contain fluorescent colourless dyes, causing more blue light to be reflected. Changes in colour may occur as the fluorescent material loses energy, but new optical whiteners can be applied during the laundering process.

Finishes enhancing tactile qualities. Finishes enhancing the feel and drape of fabrics involve the addition of sizing, weighting, fulling, and softening agents, which may be either temporary or permanent.

Sizing. Sizing, or dressing, agents are compounds that form a film around the yarn or individual fibres, increasing weight, crispness, and lustre. Sizing substances, including starches, gelatin, glue, casein, and clay, are frequently applied to cottons and are not permanent.

Weighting. Weighting, in the processing of silk, involves the application of metallic salts to add body and weight. The process is not permanent but can be repeated.

Fulling. Also called felting or milling, fulling is a process that increases the thickness and compactness of wool by subjecting it to moisture, heat, friction, and pressure until shrinkage of 10 to 25 percent is achieved. Shrinkage occurs in both the warp and weft, producing a smooth, tightly finished fabric that may be so compact that it resembles felt.

Softening. Making fabrics softer and sometimes also increasing absorbency involves the addition of such agents as dextrin, glycerin, sulfonated oils, sulfated tallow, and sulfated alcohols.

Finishes improving performance. The performance of fabrics in use has been greatly improved by the development of processes to control shrinkage, new resin finishes, and new heat-sensitive synthetic fibres.

Shrinkage control. Shrinkage control processes are applied by compressive shrinkage, resin treatment, or heat-setting. Compressive, or relaxation, shrinkage is applied to cotton and to certain cotton blends to reduce the stretching they experience during weaving and other processing. The fabric is dampened and dried in a relaxed state, eliminating tensions and distortions. The number of warp and weft yarns per square inch is increased, contributing greater durability, and fabrics treated by this method are usually smooth and have soft lustre. The process involves spraying the fabric with water, then pressing the fabric against a steam-heated cylinder covered with a thick blanket of woollen felt or rubber. The manufacturer is often required to specify the residual shrinkage, or percentage of shrinkage, that may still occur after the preshrinking process.

Residual shrinkage

Rayons and rayon blends may be stabilized by the use of resins, which impregnate the fibre. Such fabrics may also be stabilized by employing acetals to produce cross-linking, a chemical reaction. Such synthetics as polyesters and nylons, which are heat sensitive, are usually permanently stabilized by heat-setting during finishing.

Shrinkage of wools is frequently controlled by treatment with chlorine, partially destroying the scales that occur on wool fibres, thus increasing resistance to the natural tendency of wool to felt. Other methods employ coating with resins that attach to the scales in order to discourage felting shrinkage.

Durable press. Durable press fabrics have such characteristics as shape retention, permanent pleating and creasing, permanently smooth seams, and the ability to shed wrinkles, and thus retain a fresh appearance without ironing. Such fabrics may be safely washed and dried by machine. These useful characteristics are imparted by a curing process. Depending upon composition and desired results, fabrics may be precured, a process in which a chemical resin is added, the fabric is dried and cured (baked), and heat is applied by pressing after garment construction; or fabrics may be postcured, a process in which resin is added, the fabric is dried, made into a garment, pressed, and then cured.

Wash-and-wear fabrics

Wash-and-wear was an early durable press process employing chemical treatment and curing of fabrics; at least light ironing was required to restore appearance. Later, however, processes were developed that allowed such fabrics to regain smoothness after home machine washing at moderate temperature, followed by tumble drying.

Crease resistance. Crease, or wrinkle, resistance is frequently achieved by application of a synthetic resin, such as melamine or epoxy.

Soil release. Soil release finishes facilitate removal of waterborne and oil stains from fabrics such as polyester and cotton blends and fabrics treated for durable press, which usually show some resistance to stain removal by normal cleaning processes. Other finishes have been developed that give fabrics resistance to water and oil stains.

Antistatic finishes. The accumulation of static electricity in such synthetic fibres as nylon, polyesters, and acrylics produces clinging, which may be reduced by application of permanent antistatic agents during processing. Consumers can partially reduce static electricity by adding commercial fabric softeners during laundering.

Antibacterial and antifungus finishes. Antibacterial finishes are germicides applied to fabrics to prevent odours produced by bacterial decomposition, such as perspiration odours, and also to reduce the possibility of infection by contact with contaminated textiles. Fabrics may also be treated with germicides to prevent mildew, a parasitic fungus that may grow on fabrics that are not thoroughly dried. Both mildew and rot, another form of decay, may also be controlled by treatment with resins.

Moth-repellent treatments. Wool and silk are subject to attack by moths but may be made moth repellent by the application of appropriate chemicals either added in the dye bath or applied to the finished fabric.

Waterproofing and water repellence. Waterproofing is a process applied to such items as raincoats and umbrellas, closing the pores of the fabric by application of such substances as insoluble metallic compounds, paraffin, bituminous materials, and drying oils. Water-repellent finishes are surface finishes imparting some degree of resistance to water but are more comfortable to wear because the fabric pores remain open. Such finishes include wax and resin mixtures, aluminum salts, silicones, and fluoro-chemicals.

Flameproof, fireproof, and fire-resistant finishes. Flameproof fabrics are able to withstand exposure to flame or high temperature. This is achieved by application of various finishes, depending upon the fabric treated, that cause burning to stop as soon as the source of heat is removed. Fireproofing is achieved by the application of a finish that will cut off the oxygen supply around the flame. Fire-resistant finishes cause fabrics to resist the spread of flame.

DYEING AND PRINTING

Dyeing and printing are processes employed in the conversion of raw textile fibres into finished goods that add much to the appearance of textile fabrics.

Dyeing. Most forms of textile materials can be dyed at almost any stage. Quality woollen goods are frequently dyed in the form of loose fibre, but top dyeing or cheese dyeing is favoured in treating worsteds. Manufacturers prefer piece dyeing, which allows stocking of white goods, reducing the risk of being overstocked with cloth dyed in colours that have not been ordered.

The dye used depends on the type of material and the specific requirements to be met. For some purposes, high lightfastness is essential; but for others it may be inconsequential. Factors considered in dye selection include fastness to light, reaction to washing and rubbing (crocking), and the cost of the dyeing process. Effective preparation of the material for dyeing is essential.

Types of dyes. Textile dyes include acid dyes, used mainly for dyeing wool, silk, and nylon; and direct or substantive dyes, which have a strong affinity for cellulose fibres (see Table). Mordant dyes require the addition of chemical substances, such as salts, to give them an affinity for the material being dyed. They are applied to cellulosic fibres, wool, or silk after such materials have been treated with metal salts. Sulfur dyes, used to dye cellulose, are inexpensive, but produce colours lacking brilliance. Azoic dyes are insoluble pigments formed within the fibre by padding, first with a soluble coupling compound and then with a diazotized base. Vat dyes, insoluble in water, are converted into soluble colourless compounds by means of alkaline sodium hydrosulfite. Cellulose absorbs these colourless compounds, which are subsequently oxi-

Fibres and Dyes

fibres	application classes
Natural fibres	
Animal	
Wool	acid, basic, mordant, reactive, (solubilized vat)
Wool blends (wool–cotton, wool–viscose, etc.)	acid, direct, mordant, reactive
Silk	acid, basic, direct, mordant, (reactive), (solubilized vat)
Vegetable	
Cotton	azoic, basic, direct, mordant, oxidation, reactive, sulfur, vat
Bast (linen, flax, hemp, jute, ramie)	acid, direct, (disperse), reactive, vat, solubilized vat
Modified cellulose fibres	
Viscose	direct, mordant, pigment, reactive, sulfur, vat, solubilized vat
Secondary acetate	disperse
Triacetate	disperse
Synthetic fibres	
Polyamide (nylon, Perlon, Rilsan)	acid, disperse, mordant, pigment, reactive
Polyester (Dacron, Terylene)	disperse, pigment
Polyacrylonitrile (Acrilan, Courtelle, Orlon)	basic, disperse, pigment
Polyvinyl chloride (Envilon, Thermovyl)	basic, disperse
Polyolefines (Meraklon, Prolene)	disperse
Elastomers (Glospan, Lycra)	acid, disperse, reactive, (wool), vat

dized to an insoluble pigment. Such dyes are colourfast. Disperse dyes are suspensions of finely divided insoluble, organic pigments used to dye such hydrophobic fibres as polyesters, nylon, and cellulose acetates.

Reactive dyes combine directly with the fibre, resulting in excellent colourfastness. The first ranges of reactive dyes for cellulose fibres were introduced in the mid-1950s. A wide variety is now available. (C.S.Wh.)

Application process. The dyeing of a textile fibre is carried out in a solution, generally aqueous, known as the dye liquor or dyebath. For true dyeing (as opposed to mere staining) to have taken place, the coloration must be relatively permanent; that is, not readily removed by rinsing in water or by normal washing procedures. Moreover, the dyeing must not fade rapidly on exposure to light. The process of attachment of the dye molecule to the fibre is one of absorption; that is, the dye molecules concentrate on the fibre surface.

Binding forces

There are four kinds of forces by which dye molecules are bound to fibre: (1) ionic forces, (2) hydrogen bonding, (3) van der Waals' forces, and (4) covalent chemical linkages. In the dyeing of wool, which is a complex protein containing about 20 different α-amino acids, the sulfuric acid added to the dyebath forms ionic linkages with the amino groups of the protein. In the process of dyeing, the sulfate anion (negative ion) is replaced by a dye anion. In the dyeing of wool, silk, and synthetic fibres, hydrogen bonds are probably set up between the azo, amino, alkylamino, and other groups, and the amido $-CO-NH-$, groups. Van der Waals' forces (the attractive forces between the atoms or molecules of all substances) are thought to act in the dyeing of cotton between the molecular units of the fibre and the linear, extended molecules of direct dyes. Covalent chemical links are brought about in the dyebath by chemical reaction between a fibre-reactive dye molecule, one containing a chemically reactive centre, and a hydroxy group of a cotton fibre, in the presence of alkali.

In any dyeing process, whatever the chemical class of dye being used, heat must be supplied to the dyebath; energy is used in transferring dye molecules from the solution to the fibre as well as in swelling the fibre to render it more receptive. The technical term for the transfer process is exhaustion. Evenness of dyeing, known as levelness, is an important quality in the dyeing of all forms of natural and synthetic fibres; it may be attained by control of dyeing conditions, that is, by agitation to ensure proper contact between dye liquor and substance being dyed, and by use of restraining agents to control rate of dyeing, or strike.

Serious consideration has recently been given to methods of dyeing in which water as the medium is replaced by solvents such as the chlorinated hydrocarbons used in dry cleaning. There are a number of technical advantages in solvent dyeing, apart from the elimination of effluent (pollution) problems associated with conventional methods of dyeing and finishing. Advantages include more rapid wetting of textiles, less swelling, increased speed of dyeing per given amount of material, and savings in energy, because less heat is required to heat or evaporate perchlorethylene, for example, than is needed for water.

For each application the dyer selects the combination of dyes best suited to the particular fibre or blend he plans to dye and best able to withstand the conditions the textile will encounter in further processing and in use in the finished article. In general, the higher the standard of fastness, the more expensive the dye, and the final choice may be a compromise between the desired fastness standards and the cost of the dyes. Fastness tests and standards have been the subject of work by the American Association of Textile Chemists and Colorists (AATCC), Europäisch-Continentale Echtheitsconvention (ECE), and the Society of Dyers and Colourists (SDC), Bradford, West Yorkshire. Efforts have been made to set up a unified system by the International Organization for Standardization (ISO). Lightfastness is assessed on a scale of 8; 1 represents the poorest fastness and 8 the best. Fastness to other agents, among them water, bleach, acid, alkali, detergent solution, and perspiration, is measured on a scale of 5.

Fastness standards

Dyes are generally used in combination to achieve a desired hue or fashion shade. If the substance to be dyed consists of only one type of fibre, such as wool, the dye mixture will be made up solely of wool dyes. But if the fabric contains more than one kind of fibre and they differ in dyeing properties, then mixtures of different application classes of dyes are used.

Forms in which textiles are dyed. Loose stock consists of randomly distributed wool or cotton fibres; tow is the corresponding term for synthetic fibres. Sliver is a more orderly arrangement of fibres in a loosely connected, continuous form suitable for spinning. It is wound into either hanks or tops, loose balls about one foot in diameter. After spinning, the yarn is either made up into hanks or into packages weighing about two pounds each, by winding the yarn round perforated metal tubes. The packages are curiously named, some according to their shapes; for example, cones, cheeses, cakes, beams, and rockets. Piece goods, woven cloth or textiles knitted in rope form, and garments, a term that includes stockings, tights, hose, and half hose, are also dyed as such.

Hanks and packages

Machinery and equipment. Modern dyeing machines are made from stainless steels. Steels containing up to 4 percent molybdenum are favoured to withstand the acid conditions that are common. A dyeing machine consists essentially of a vessel to contain the dye liquor, provided with equipment for heating, cooling, and circulating the liquor into and around the goods to be dyed or moving the goods through the dye liquor. The kind of machine employed depends on the nature of the goods to be dyed. Labour and energy costs are high in relation to total dyeing costs; the dyer's aim is to shorten dyeing times to save steam and electrical power and to avoid spoilage of goods.

A widely used machine is the conical-pan loose-stock machine; fibres are held in an inner truncated-conical vessel while the hot dye liquor is mechanically pumped through. The fibre mass tends to become compressed in the upper narrow half of the cone, assisting efficient circulation. Levelling problems are less important because uniformity may be achieved by blending the dyed fibres prior to spinning.

The Hussong machine

The Hussong machine is the traditional apparatus; it has a long, square-ended tank as dyebath into which a framework of poles carrying hanks can be lowered. The dye liquor is circulated by an impeller and moves through a perforated false bottom that also houses the open steam pipe for heating. In modern machines, circulation is improved especially at the point of contact between hank and pole. This leads to better levelling and elimination of irregularities caused by uneven cooling.

In package-dyeing machines dye liquor may be pumped in either of two directions: (1) through the perforated central spindle and outward through the package, or (2) by the reverse path into the outer layers of the package and

out of the spindle. In either case levelness is important. In the case of soluble dyes the dye liquor must be free of suspended matter. In the case of disperse dyes, in which particles of dye are dispersed in, rather than dissolved in, the solution, no gross aggregates can be allowed; otherwise the packages would retain undesirable solids on the outer and inner surfaces. Some package-dyeing machines are capable of working under pressure at temperatures up to 130° C.

Piece-dyeing machines

The winch is the oldest piece-dyeing machine and takes its name from the slatted roller that moves an endless rope of cloth or endless belt of cloth at full width through the dye liquor. Pressurized-winch machines have been developed in the U.S. In an entirely new concept, the Gaston County jet machine circulates fabric in rope form through a pipe by means of a high-pressure jet of dye liquor. The jet machine is increasingly important in high-temperature dyeing of synthetic fibres, especially polyester fabrics.

Another machine, the jig, has a V-shaped trough holding the dye liquor and guide rollers to carry the cloth at full width between two external, powered rollers; the cloth is wound onto each roller alternately, that is, the cloth is first moved forward, then backward through the dye liquor until dyeing is complete. Modern machines, automatically controlled and programmed, can be built to work under pressure.

Solutions or suspensions of colorants or their precursors may be padded onto piece goods by passing the cloth through a trough containing the liquor and then between rollers under pressure. Development and fixation processes such as steaming or dry-heat treatment can be carried out in other apparatus. The method is used in semicontinuous and continuous operations. (E.M.A./Ed.)

Printing. Printing is a process of decorating textile fabrics by application of pigments, dyes, or other related materials in the form of patterns. Although apparently developed from the hand painting of fabrics, such methods are also of great antiquity. There is evidence of printing being carried out in India during the 4th century BC, and a printing block dated at about AD 300 has been unearthed in the burial grounds of Akhmīn in Upper Egypt. Pre-Columbian printed textiles have been found in Peru and Mexico. Textile printing has become highly sophisticated and has involved the skills of many artists and designers.

Principal methods of textile printing

The four main methods of textile printing are block, roller, screen, and heat transfer printing. In each of these methods, the application of the colour, usually as a thickened paste, is followed by fixation, usually by steaming or heating, and then removal of excess colour by washing. Printing styles are classified as direct, discharge, or resist. In direct printing, coloured pastes are printed directly on the cloth. For discharge printing, the cloth is first dyed with a background colour, which is destroyed by reagents, or reducing agents, carried in a print paste. This action may leave the discharged design white on a coloured background, although print pastes may also contain colouring matters not destroyed by the discharging agent, producing a coloured design. In the resist process, the cloth is first printed with a substance called a resist, protecting these printed areas from accepting colour. When the cloth is dyed or pigment padded only those parts not printed with the resist are dyed. A special application of this technique, imparting plissé effects, is the printing of the fabric with a resist, followed by treatment with caustic soda.

Block printing. Wooden blocks, carved with a design standing out in relief, are made from solid pieces of wood or by bonding closely grained woods with cheaper ones. When designs include large areas, these are recessed and the space filled with hard wool felt. Fine lines are usually built up with copper strips, and other effects are obtained with copper strips interleaved with felt. To facilitate registration of successive prints, or lays, each block has several pitch pins arranged to coincide with well-defined points in the pattern. Cloth is printed on a table covered with several thicknesses of fabric or blanket, the whole covered with a thick sheet of tightly stretched synthetic rubber. The cloth to be printed is spread on the rubber, either gummed in position or pinned to a backcloth attached to the table. Colour is applied evenly to the block, and the pattern is stamped on the fabric to be printed, using the handle of a small heavy hammer, or maul, to aid penetration of the paste. More colour is then applied to the block and the process is repeated using the pitch pin to obtain true registration. After the fabric has been entirely printed with one colour, other colours are applied in the same way until the design is complete. Although block printing is becoming too laborious and costly for commercial use, some of the most beautiful prints have been made in this way.

Roller printing. This technique is used whenever long runs of fabric are to be printed with the same design. The modern machine, based on one originally devised in 1783, consists of a large central cast-iron cylinder over which passes a thick endless blanket providing a resilient support for the fabric. Backing fabrics, called back grays, are placed between the blanket and the fabric to prevent undue staining of the blanket. Although formerly made of cotton fabric, most modern back grays are continuous belts of nylon. The blanket and back gray are appropriately tensioned, so that the fabric moves through the machine as the central cylinder rotates. Engraved printing rollers, one for each colour, press against the fabric and the central cylinder. The pattern on the roller is etched on the surface of a copper shell supported on a mandrel. High-quality engraving is essential for good printing. Each printing roller is provided with a rotating colour-furnishing roller, partially immersed in a trough of printing paste. Finely ground blades (doctor blades) remove excess colour paste from the unengraved areas of these rollers, and each also has a lint blade. The printed fabric passes from the main cylinder and through a drying and steaming chamber to fix the colour. Although this machine prints only one side of the fabric, the Duplex roller machine, essentially a combination of two roller machines, prints both sides. Modern printing machines are smooth-running precision machines fitted with carefully designed roller bearings and hydraulic or pneumatic mechanisms to ensure uniform pressure and flexibility. Pressure is regulated from an instrument panel, and each roller is controlled independently. Automatic registration is effected by electromagnetic push-button control, and modern electric motors provide smooth-running, variable-speed drives. The washing of back grays and printer's blankets has also been automated.

Spray printing is the application of colour from spray guns through stencils and has limited but occasionally profitable use.

Screen printing. Screen printing may be a hand operation or an automatic machine process. The cloth is first laid on a printing table, gummed in position or pinned to a back gray, and then the design is applied through a screen made of silk or nylon gauze stretched over a wooden or metal frame, on which the design for one colour has been reproduced. This is usually a photographic process, although hand painting with a suitably resistant blocking paint is an alternative. A screen is placed over the fabric on the table against registration stops, ensuring accurate pattern fitting. Print paste is poured on to the screen edge nearest the operator and is spread with a squeegee over the surface of the screen so that colour is pushed through the open parts. The screen is moved until one colour has been applied to the cloth. For application of other colours, the process is repeated with different screens.

With the growing importance of screen printing, the hand operation has been largely replaced by mechanical methods. In some machines, the screens are flat, as in hand printing; others employ rotary screens.

Heat transfer printing. The popularity of polyester fabrics has led to the development of a completely new form of printing termed heat transfer printing, which prints the pattern on paper with carefully selected dyes. The paper is then applied to the fabric by passing the two together through a type of hot calender, and the pattern is transferred from one to the other. This method opens up new possibilities such as the production of halftone effects.

Print paste viscosity

In all textile printing, the nature and, particularly, the viscosity of the print paste are important, and the thickeners employed must be compatible with all the other components. For conventional methods the thickeners are such reagents as starch, gum tragacanth, alginates, methyl cel-

lulose ethers, and sodium carboxymethyl cellulose. Many types of dye can be applied, including direct cotton, vat, mordant, and reactive dyes, as well as pigment colours. Most dyes are fixed by steaming or aging, by a batch or continuous method, and more rapid fixation is effected by flash aging—*e.g.*, allowing a shorter steaming period by employing smaller machines. After steaming, the fabric must be thoroughly washed to remove loose dye and thickener, ensuring fastness to rubbing.

Most textile materials can be printed without special pretreatment, but wool cloths are generally chlorinated before printing. Tops (long, parallel wool fibres), printed in stripes, are used for mixed effects, and printed warps produce shadowy effects. Tufted carpets are printed by a process designed to ensure good penetration.

Textile consumption

Textiles are commonly associated with clothing and soft furnishings, an association that accounts for the great emphasis on style and design in textiles. These consume a large portion of total industry production.

Changing uses of fabric in apparel. Great changes have occurred in the fabrics used for clothing, with heavy woollen and worsted suitings being replaced by lighter materials, often made from blends of natural and man-made fibres, possibly due to improved indoor heating. Warp-knitted fabrics made from bulked yarns are replacing woven fabrics, and there is a trend away from formality in both day and evening dress to more casual wear, for which knitted garments are especially appropriate. The use of synthetic fibre fabrics has established the easy-care concept and made formerly fragile light and diaphanous fabrics more durable. The introduction of elastomeric fibres has revolutionized the foundation-garment trade, and the use of stretch yarns of all types has produced outerwear that is closely fitting but comfortable.

Interlinings

Manufacturers of tailored garments formerly used interlinings made of horsehair, which was later replaced by goat hair, and then by resin-treated viscose rayon. Today fusible interlinings and various washable synthetics are widely used. The performance of a garment is greatly influenced by such factors as the interlining used and the sewing threads employed.

The care required by a textile fabric depends upon both fibre content and the application of various finishing processes. In 1972 the United States Federal Trade Commission passed regulations requiring fabric manufacturers to provide the consumer with care labels to be sewn into homemade garments, and requiring ready-to-wear manufacturers to sew permanent care information labels into clothing (see also INDUSTRIES, MANUFACTURING: *Clothing and footwear industry*).

Household textiles. Household textiles, frequently referred to as soft furnishings, are fabrics used in the home. They include items frequently classified as linens, such as bath and dish towels, table linens, shower curtains, and bathroom ensembles. Related items include sheets, pillowcases, mattresses, blankets, comforters, and bedspreads. In addition, textile products contributing to the atmosphere and comfort of the home include rugs and carpeting, draperies, curtains, and upholstery fabrics.

Most of these items are also used in hotels and motels, and many are used in offices, showrooms, retail stores, restaurants, recreational facilities, and various other commercial establishments.

Industrial fabrics. This class of fabrics includes composition products, processing fabrics, and direct use types.

Composition products. In composition products, the fabrics are used as reinforcements in compositions with other materials, such as rubber and plastics. These products—prepared by such processes as coating, impregnating, and laminating—include tires, belting, hoses, inflatable items, and typewriter-ribbon fabrics.

Processing fabrics. Processing fabrics are used by various manufacturers for such purposes as filtration, for bolting cloths used for various types of sifting and screening, and in commercial laundering as press covers and as nets segregating lots during washing. In textile finishing back grays are used as backing for fabrics that are being printed.

Direct-use fabrics. Direct-use fabrics are manufactured or incorporated into finished products, such as awnings and canopies, tarpaulins, tents, outdoor furniture, luggage, and footwear.

Fabrics for protective clothing. Fabrics for military purposes must frequently withstand severe conditions. Among their uses are Arctic and cold-weather clothing, tropical wear, rot-resistant material, webbing, inflated life vests, tent fabrics, safety belts, and parachute cloth and harnesses. Parachute cloth, for example, must meet exacting specifications, air porosity being a vital factor. New fabrics are also being developed for garments used in space travel. In protective clothing a subtle balance between protection and comfort is required.

The many uses of textiles enter into almost every aspect of modern life. For some purposes, however, the role of textiles is being challenged by developments in plastic and paper products. Although many of these presently have certain limitations, it is likely that they will be improved, presenting a greater challenge to textile manufacturers, who must be concerned with both retaining present markets and expanding into completely new areas.

(C.S.Wh./Ed.)

BIBLIOGRAPHY. Works of general interest include: GEORGE E. LINTON, *Applied Basic Textiles* (1966); ISABEL B. WINGATE, *Textile Fabrics and Their Selection,* 7th ed. (1976); EVELYN E. STOUT, *Introduction to Textiles,* 3rd ed. (1970); NORMA HOLLEN, JANE SADDLER, and ANNA L. LANGFORD, *Textiles,* 5th ed. (1979); and KAX WILSON, *A History of Textiles* (1979). Helpful dictionaries of the textile industry are: GEORGE E. LINTON (ed.), *The Modern Textile and Apparel Dictionary,* 4th ed. rev. (1973); ISABEL B. WINGATE (ed.), *Fairchild's Dictionary of Textiles,* 6th ed. (1979); and CAROLYN A. FAIRCHILD and P.J. ALVEY (eds.), *Textile Terms and Definitions,* 7th ed. (1975). The history and development of the textile industry is treated in ADÈLE C. WEIBEL, *Two Thousand Years of Textiles* (1952, reissued 1972); and the TEXTILE INSTITUTE, *Review of Textile Progress,* 18 vol. (1949–67). Specific fibres, their processing, and their characteristics are surveyed in: J. GORDON COOK, *Handbook of Textile Fibres* (1968); H.R. MAUERSBERGER (ed.), *Matthews' Textile Fibres,* 6th ed. (1954); R.W. MONCRIEFF, *Man-Made Fibres,* 6th ed. (1975); J.J. PRESS (ed.), *Man-Made Textile Encyclopedia* (1959); and H.F. MARK, S.M. ATLAS, and E. CEARNIA (eds.), *Man-Made Fibers: Science and Technology,* 3 vol. (1967–68). Yarn production is the subject of P.R. LORD, (ed.), *Spinning in the Seventies* (1970); and G.R. WRAY (ed.), *Modern Yarn Production from Man-Made Fibres* (1976). Works treating fabric production include: A.T.C. ROBINSON and R. MARKS, *Woven Cloth Construction* (1967); P.R. LORD and M.H. MOHAMED, *Weaving: Conversion of Yarn to Fabric* (1973); D.G.B. THOMAS, *An Introduction to Warp Knitting* (1976); H. WIGNALL, *Knitting* (1964); V. DUXBURY and G.R. WRAY (eds.), *Modern Developments in Weaving Machinery* (1962); and R. KRČMA, *Non-Woven Textiles* (1967; originally published in Czech, 1962). Textile finishing is the subject of A.J. HALL, *Handbook of Textile Finishers,* 2nd ed. (1957); and works treating the dyeing and printing of textiles include T. VICKERSTAFF, *The Physical Chemistry of Dyeing,* 2nd ed. rev. (1954); E.R. TROTMAN, *Dyeing and Chemical Technology of Textile Fibres,* 5th ed. (1975); C.L. BIRD, *The Theory and Practice of Wool Dyeing,* 3rd ed. (1963); H.U. SCHMIDLIN, *Preparation and Dyeing of Synthetic Fibres* (Eng. trans. 1963); A.J. HALL, *Handbook of Textile Dyeing and Printing* (1955); S.R. COCKETT, *Dyeing and Printing* (1964); and L.W.C. MILES, *Textile Printing* (1971). Works concerned with textile quality control include: E.B. GROVER and D.S. HAMBY, *Handbook of Textile Testing and Quality Control* (1960); BRITISH STANDARDS INSTITUTION, *Methods of Test for Textiles, B.S. Handbook No. 11,* 4th ed. (1974); AMERICAN SOCIETY FOR TESTING AND MATERIALS, *1966 Book of A.S.T.M. Standards,* part 24–25 (1966); SOCIETY OF DYERS AND COLOURISTS, *Standard Methods for the Determination of the Colour Fastness of Textiles and Leathers,* 4th ed. (1978). Studies of the potential of textile industries in developing countries include the UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION, *Industrialization of Developing Countries: Problems and Prospects, Textile Industry,* vol. 7 (1969); and the *ECAFE Seminar on the Man-Made Fiber Industry: Japan's Man-Made Fibers* (1966). Periodicals and trade papers concerned with the textile industry include: *CIBA-GEIGY Review* (quarterly); *American Fabrics* (quarterly); *Textile Institute Journal* (irregular); *Textile Organon* (monthly); and *Textile Progress* (quarterly). For the history see also DAVID J. JEREMY, *Transatlantic Industrial Revolution: The Diffusion of Textile Technologies Between Britain and America, 1790–1830s* (1981).

(C.S.Wh./E.N.A./Ed.)

Infectious Diseases

In human beings, an infectious disease is a process caused by an organism, most commonly a microorganism, that impairs a person's health. An infection, in contrast, is the invasion of the body by organisms and the reaction of tissues to their presence or to the toxins that they produce, whether or not health is affected. (An infection that does not alter health is called a subclinical infection.) Thus, a person may be infected but not have an infectious disease. This principle is illustrated by the use of vaccines for the prevention of infectious diseases. For example, a virus such as that which causes measles may be partially inactivated (attenuated) and used as an immunizing agent. The immunization is designed to produce a measles infection in the recipient but generally causes no discernible alteration in the state of health. It produces immunity to measles without producing a clinical illness (an infectious disease).

Among the agents of infectious disease are microorganisms, the microscopic or submicroscopic "germs" such as viruses, bacteria, and protozoans. Parasites such as the acarids, or itch mites, of scabies are barely visible, and helminths, or worms, may grow to be several feet long. The only characteristic that all of these have in common is their dependence, for all or part of their life cycles, on their ability to parasitize, or live on, human beings. To survive and multiply, a virus must get inside a living cell, but many bacteria can live and multiply for a time on nonliving matter, while helminths, for part of their lives, lead an independent existence. One function of an organism that does not correlate with size is the ability to cause serious disease. The virus of poliomyelitis, for example, is only about 25 millimicrometres (a micrometre, or micron, is 0.001 millimetre [0.00004 inch] and a millimicrometre is 0.001 micrometre) in diameter.

The most important barriers to invasion of the human host by microorganisms are the skin and mucous membranes (the tissues that line the nose, mouth, and upper respiratory tract). When these tissues have been broken or affected by earlier disease, invasion by microorganisms may occur. These microorganisms may produce a local infectious disease, such as boils, or may invade the bloodstream and be carried throughout the body, producing generalized bloodstream infection (septicemia) or localized infection at a distant site, such as meningitis (an infection of the coverings of the brain and spinal cord). Infectious agents can be swallowed in food and drink to attack the wall of the intestinal tract and cause local or general disease. The conjunctiva, which covers the front of the eye, may be penetrated by viruses that cause a local inflammation of the eye or pass into the bloodstream and cause a severe general disease, such as measles or smallpox. Microorganisms can enter the body through the genital tract, setting up the acute inflammatory reaction of gonorrhea in the genital and pelvic organs or spreading out to attack almost any organ of the body with the more chronic and destructive lesions of syphilis. Even before birth, viruses and other infectious agents can pass through the placenta and attack developing cells, so that an infant may be diseased or deformed at birth.

From conception to death, then, human beings are targets for attack by multitudes of other living organisms, all of them competing for a place in the common environment. The air people breathe, the soil they walk on, the waters and vegetation around them, the buildings they inhabit and work in, all can be populated with forms of life that are dangerous. Domestic animals may harbour organisms that are a threat, and wildlife teems with agents of infection that can afflict human beings with serious disease. The human body is not without defenses against these threats, for it is equipped with sensitive mechanisms that react quickly and specifically against disease organisms when they attack, and survival throughout the ages has depended largely on these reactions. The human environment, in the biological sense, is mainly a hostile one. Human beings have learned to partially control it, but they cannot subdue it, and sometimes a minor change in the environment may lead to unforeseen alterations in the balance between people and their biological competitors.

(A.B.C./R.D.F./Re.G.)

The article is divided into the following sections:

Infectious agents 527
- Categories of organisms 527
- Modes of survival 528
- Commensal organisms 528

Host–parasite relationships 529
- Immune response 529
- Immunization 530
- Temperature regulation 533

Infections of specific organ systems 533
- Upper respiratory tract infections 533
- Lower respiratory tract infections 534
- Infections of the cardiovascular system 535
- Central nervous system infections 535
- Genitourinary tract infections 536
- Gastrointestinal infections 536
- Infections of the bones and joints 538
- Skin infections 538
- Ocular infections 538

Infections with specific organisms 539
- Bacteria 539
- Viruses 543
- Chlamydia 546
- Rickettsias 546
- Mycoplasmas and ureaplasmas 547
- Fungi 547
- Parasites 547
- Acquired immunodeficiency syndrome 549

Effects of environment on human disease 550
- Human activity 550
- The inanimate environment 551

Bibliography 552

Infectious agents

CATEGORIES OF ORGANISMS

The agents of infection can be divided into different groups on the basis of their size, biochemical characteristics, or manner in which they interact with the human host. The groups of organisms that cause infectious diseases are categorized as bacteria, viruses, *Chlamydia,* rickettsias, mycoplasmas and ureaplasmas, fungi, and parasites.

Bacteria

Although bacteria are generally large enough to be seen under the light microscope, viruses are not. Streptococci, the bacteria that cause scarlet fever, are about 0.75 micrometre (0.00003 inch) in diameter. The spirochetes, which cause syphilis, leptospirosis, and rat bite fever, are five to 15 micrometres long. Bacteria can survive within the body but outside of individual cells. Some bacteria, classified as aerobes, require oxygen for growth, while others, such as those normally found in the small bowel of healthy persons, only grow in the absence of oxygen and, therefore, are called anaerobes. Most bacteria are surrounded by a capsule that appears to play an important role in their ability to produce disease. A number

of bacterial species give off toxins that, in turn, may damage tissues. Bacteria contain both deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) and can be treated with antibiotics.

Viruses are typically intracellular parasites, and for that reason they cannot be grown in the laboratory unless the growing medium contains living cells. Viruses are classified into various groups on the basis of their size or shape or whether they contain DNA or RNA (all contain either DNA or RNA, but not both). Viruses are visible by electron microscopy; they vary in size from about 25 millimicrometres for poliovirus to 250 millimicrometres for smallpox virus. Viruses cannot be treated with antibiotics.

Chlamydia are intracellular parasites that are 250 to 500 millimicrometres in diameter. When they interact with the host's cells, intracellular particles are produced. *Chlamydia* contain DNA and RNA and can be treated with antibiotics. Human diseases caused by *Chlamydia* include conjunctivitis (inflammation of the covering of the eye) in newborn infants, pneumonia, and genital infections.

Rickettsias are small organisms that range in size from 250 millimicrometres to more than one micrometre. They have no cell wall but are surrounded by a cell membrane. Except for *Coxiella burnetii,* the cause of Q fever, they are intracellular parasites, most of which are transmitted to human beings by an arthropod carrier. *C. burnetii,* however, can survive in milk, sewage, and aerosols and can be transmitted to human beings by the tick or by inhalation, causing pneumonia in the latter case. Rickettsial diseases can be treated with antibiotics.

Mycoplasmas and ureaplasmas, which range in size from 150 to 850 millimicrometres, are the smallest organisms capable of growing in a cell-free medium. These organisms can cause pneumonia and genital infections, as well as infection of the brain, spinal cord, liver, and other organs. Diseases due to mycoplasmas and ureaplasmas can be treated with antibiotics.

Fungi

Fungi may exist as yeasts or molds and may alternate between the two forms, depending upon environmental conditions. Yeasts are simple cells, three to five micrometres in diameter. Molds, however, consist of filamentous branching structures (called hyphae), two to 10 micrometres in diameter, that are formed of several cells lying end to end. Fungal diseases in human beings are called mycoses; they include such disorders as histoplasmosis, coccidioidomycosis, and blastomycosis. These diseases can be mild, characterized by an upper respiratory infection, or severe, characterized by involvement of the bloodstream and every organ system. Antibiotics are effective in their treatment.

Among the infectious parasites are the protozoans, unicellular organisms that have no cell wall, that cause such diseases as malaria. The various species of malarial parasites are about four micrometres in diameter. At the other extreme, the tapeworm can grow to several feet in length; treatment is designed to either kill the worm or dislodge it from its host. (R.D.F./Re.G.)

MODES OF SURVIVAL

Infectious agents have various methods of survival. Some depend on rapid multiplication and rapid spread from one host to another. For example, when the measles virus enters the body it multiplies for a week or two, then enters the bloodstream and spreads to every organ. For several days before a rash appears, the surface cells of the respiratory tract are bursting with measles virus, and vast quantities are shed every time the infected person coughs or sneezes. A day or two after the rash appears, the amount of antibody (protein produced in response to a pathogen) rises in the bloodstream, killing the virus and stopping further shedding. The patient rapidly becomes noninfectious but has already spread the virus to others. In this way an epidemic rapidly builds up, and there is an enormous increase in the amount of measles virus in circulation. Many other infectious agents—for example, influenza virus—survive in this manner. How such viruses exist between epidemics is less clear, but the infection certainly smolders, perhaps in human or animal carriers.

The picture is different in more chronic infections. In tuberculosis there is no overwhelming multiplication and no rapid shedding of the tubercle bacillus. Rather, the bacilli remain in the infected person's body for a long period, slowly forming areas of chronic inflammation that from time to time break down and allow them to escape.

Some organisms form spores, a resting or dormant stage that is resistant to heat, cold, drying, and chemical action. Spore-forming organisms can survive for months or years under the most adverse conditions and are not, in fact, highly infectious. The germ of tetanus, *Clostridium tetani,* is present everywhere in the environment—in soil, in dust, on window ledges and floors—and yet tetanus is an uncommon disease, especially in developed countries. The same is true of the anthrax germ, *Bacillus anthracis.* Although usually present in abundance in factories in which rawhides and animal wool and hair are handled, it rarely causes anthrax in employees. *Clostridium botulinum,* the germ of botulism, produces one of the most lethal toxins that can afflict human beings, and yet the disease is one of the rarest because the germ depends for its survival on its resistant spore, which is not parasitic on human beings.

Tetanus, anthrax, and botulism

In contrast to these relatively independent organisms, there are others that cannot exist at all outside the human body. The germs of syphilis and gonorrhea, for example, depend for survival on their great ability to infect and their complete adaptation to a parasitic existence.

Some organisms have complicated life cycles and depend on more than one host. The malarial parasite must spend a portion of its life cycle inside a mosquito, while the liver fluke, an occasional human parasite, spends part of its life in the body of a sheep, part in a water snail, and part in the open air as a cyst attached to grass.

COMMENSAL ORGANISMS

All of the outer surfaces of the human body are covered with microorganisms that normally do no harm and may, in fact, be beneficial. Those commensal organisms on the skin help to break down dying skin cells or to destroy debris secreted by the many minute glands and pores that open on the skin. Many of the organisms in the intestinal tract break down complex waste products into simple substances, and others help in the manufacture or synthesis of chemical compounds, such as some of the vitamins that are essential to human life.

It should be understood that the gastrointestinal canal is considered in this regard to be one of the "outer" surfaces of the body. From its upper part, the stomach and duodenum, food is absorbed into the body, while in its lower part, the large intestine, waste products are excreted from the body. Embryologically, the intestinal canal is an external structure formed by the intucking, or invagination, of the ectoderm, or outer surface, of the body. The mouth, nose, and sinuses (spaces inside the bones of the face) are also considered to be external structures because of their direct contact with the outside environment. Both the gastrointestinal canal and the mouth, nose, and sinuses are heavily populated with microorganisms, some of which are true commensals—living in human beings and deriving their sustenance from the surface cells of the body without doing any harm—others of which are indistinguishable from disease germs. The latter may live like true commensals in a particular tract in a human being and never cause disease, despite their potential to do so. When the environment is altered, however, they are capable of causing severe illness in their host or, without harming their host, they may infect a second person with a serious disease.

True commensals

It is not known why, for example, the hemolytic streptococcus bacterium can live for months or years in the throat without causing harm and then suddenly cause an acute attack of tonsillitis, or how an apparently harmless pneumococcus gives rise to pneumonia. Similarly, it is not understood how a person can harmlessly carry *Haemophilus influenzae,* type B, in the throat but then become ill when the organism invades the body and causes one of the most severe forms of meningitis. It may be that external influences, such as changes in temperature or humidity, are enough to upset the balance between host and parasite, or that a new invader, perhaps a virus, enters

and, by competing for some element in the environment, forces the original parasite to react more violently with its host. The term "lowered resistance," often proffered to help describe conditions at the onset of infectious disease, is not specific and simply implies any change in the immune system or inflammatory system of the host.

A microorganism's environment can be changed radically, of course. If antibiotics are administered, the body's commensal organisms can be killed and other organisms, less innocuous, may take their place. In the mouth and throat, penicillin may eradicate pneumococci, streptococci, and other bacteria that are sensitive to the drug, while microorganisms that are insensitive, such as *Candida albicans,* may then proliferate and cause thrush (an inflammatory condition of the mouth and throat). In the intestinal canal, an antibiotic may kill most of the microorganisms that are normally present and allow dangerous organisms, such as *Pseudomonas aeruginosa,* to multiply and perhaps to invade the bloodstream and the tissues of the body. If an infectious agent—for example, *Salmonella*—reaches the intestinal canal, treatment with an antibiotic may have an effect that differs from what was intended. Instead of attacking and destroying the salmonella, it may kill the normal inhabitants of the bowel and allow the salmonella to flourish and persist in the absence of competition from other microorganisms.

Host–parasite relationships

When a pathogenic (disease-causing) microorganism invades the body for the first time, the clinical (observable) response may range from nothing at all, through various degrees of nonspecific reactions, to specific infectious disease. Immunologically, however, there is always a response, the purpose of which is defense. If the defense is completely successful, there is no obvious bodily reaction; if partially successful, the affected person suffers but recovers from an infectious disease; if unsuccessful, the person is overwhelmed by the intensity of the infectious process and dies.

Natural history of poliomyelitis

The two responses—the clinical and the immunological—can be illustrated by the natural history of the disease poliomyelitis. When the virus of this disease enters the body for the first time, it multiplies in the throat and in the intestinal tract. In some people, it gets no further; virus is shed to the outside from the throat and the bowel for a few weeks, and then the shedding ceases and the infection is over. The host, however, has responded and has developed circulating antibodies to a type of poliovirus. These antibodies are specific globulin elements in the blood that subsequently prevent disease should that poliovirus again be encountered.

In other people, the same process occurs but some virus also gets into the bloodstream, where it circulates for a short time before being eliminated. In a few individuals, the virus passes from the bloodstream into the central nervous system, where it circulates for a short time before being eliminated. In a few individuals, the virus passes from the bloodstream into the central nervous system, where it may enter and destroy some of the nerve cells that control movement in the body and so cause paralysis. Such paralysis is the least common result of infection with poliomyelitis virus; most infected persons have no symptoms at all. Those whose bloodstream becomes infected often have a mild illness, consisting of no more than malaise, slight headache, and possibly a sore throat. This so-called minor illness of poliomyelitis is unlikely to be recognized except by those in close contact with someone paralyzed by the disease.

If the central nervous system becomes invaded by the virus, the infected person suffers from severe headache and other symptoms suggesting meningitis. This is known as the "major illness" of poliomyelitis; such persons are acutely ill, but most recover their normal health after about one week. Only a few of those with the major illness suffer from paralysis; the central nervous system has a large reserve of cells, as have most human organs, and only if a great number of cells in any area are destroyed by virus is the nervous system unable to send out its motor impulses to the muscles served by that area, resulting in paralysis. Of all the people infected with poliomyelitis virus, not more than one in 100, possibly as few as one in 1,000, suffers from paralysis, though paralysis is the dominant feature of the fully developed clinical picture of poliomyelitis and though, to most people, the disease means paralysis.

This wide range of response to the poliomyelitis virus is true of most infections, though the proportions may vary. Influenza virus, for example, may cause symptoms ranging from a mild cold to a feverish illness, severe laryngitis (inflammation of the larynx, or voice box) or bronchitis, or an overwhelming and fatal pneumonia, and the proportions of persons suffering from these differing effects may vary from one epidemic to another. In measles, the typical symptoms are a rash and some respiratory problems, although infections without physical manifestations do occur.

Influenza symptoms

IMMUNE RESPONSE

Natural immunity. There is perhaps more uniformity of pattern in the operation of the defense mechanisms of the body, called the immunological response. Every animal species possesses some natural resistance to disease. Human beings have a high degree of resistance to foot-and-mouth disease, for example, while the cattle and sheep with which they may be in close contact suffer in the thousands from it. Rats are highly resistant to diphtheria, whereas children readily contract the disease.

What such resistance depends on is not understood. It must to some degree be genetically determined; it is possible, for example, to produce by selective breeding two strains of rabbits, one highly susceptible to tuberculosis, the other highly resistant. In human beings, there may be racial differences, but it is always difficult to disentangle such factors as climate, nutrition, and economics from those that might be genetically determined. In some tropical and subtropical countries, for example, poliomyelitis is a rare clinical disease, though a common infection, but visitors to such countries often contract serious clinical forms of the disease. The absence of serious disease in the residents is not, however, due to natural resistance but to resistance acquired after repeated exposure to small doses of poliomyelitis virus from infancy onward. Visitors from other countries, with perhaps stricter standards of hygiene, are protected from such immunizing exposures and have no acquired resistance to the virus when they encounter it as adults.

Natural resistance, in contrast to acquired immunity, does not depend upon such exposures. The human skin obviously has great inherent powers of resistance to infection, for most cuts and abrasions heal quickly, though often they are smothered with potentially pathogenic microorganisms. If an equal number of typhoid germs is spread on a person's skin and on a glass plate, those on the skin die much more quickly than do those on the plate, suggesting that the skin has some bactericidal property against typhoid germs. The skin also varies in its resistance to infectious organisms at different ages: ringworm is a common fungal infection of children's skin but is rarer in adults, and acne is a common infection of the skin of adolescents but is uncommon in childhood or in older adults. The phenomenon of natural immunity can be illustrated equally well with examples from the respiratory, intestinal, or genital tracts, where large surface areas are exposed to potentially infective agents and yet infection does not occur. (A.B.C./R.D.F./Re.G.)

White blood cells

If an organism causes local infection or gains entry into the bloodstream, a complicated series of events ensues. White blood cells (polymorphonuclear leukocytes, or granulocytes), which are normally manufactured in the bone marrow and which circulate in the blood, move to the site of the infection. Some of these cells reach the site by chance, in a process called random migration, since every body site is supplied constantly with the blood in which these cells circulate. Additional polymorphonuclear leukocytes are attracted and directed to the sites of infection, in a process called directed migration, or chemotaxis, by various substances normally found as components of human

blood. A number of these components are collectively referred to as the complement system.

When a polymorphonuclear leukocyte reaches the invading organism, the cell attempts to ingest, or eat, the invader. Ingestion of bacteria may require the help of still other components of the blood, called opsonins, which act to coat the bacterial cell wall and prepare it for ingestion. An opsonin generally is a protein substance, such as one of the circulating globulins or complement components.

Once a prepared bacterium is eaten and is inside the white blood cell, a complex series of biochemical events occurs. The bacteria may be killed but its products passed into the bloodstream, where they come in contact with other circulating cells called lymphocytes. Two general types of lymphocytes—T cells and B cells—are of great importance in protecting the human host. When a T lymphocyte encounters bacterial products, it is sensitized to recognize the material as foreign and, once sensitized, it possesses an immunological memory. If the T lymphocyte encounters the same bacterial product again, it immediately recognizes it and sets up an appropriate defense more rapidly than it did on the first encounter. The ability of a T lymphocyte to function normally, providing what is generally referred to as cellular immunity, is dependent on the thymus gland. The lack of a thymus, therefore, impairs the body's ability to defend itself against various types of infections.

After a T lymphocyte has encountered and responded to a foreign bacteria, it interacts with B lymphocytes, which are responsible for producing the circulating proteins of normal blood called immunoglobulins (gamma globulins). There are various types of B lymphocytes, each of which can produce only one of the five known forms of immunoglobulin (Ig). The first immunoglobulin to be produced is IgM. Later, during recovery from infection, the immunoglobulin IgG, which can specifically kill the invading microorganism, is produced. If the same microorganism invades the host again, the B cell immediately responds with a dramatic production of IgG specific for that organism, rapidly killing it and preventing disease. The presence in the bloodstream of IgG specific for a certain bacteria (called humoral immunity), or the ability to rapidly produce the IgG, provides specific protection against that agent only. (R.D.F./Re.G.)

Immunoglobulins

Most infectious agents exist in several types or strains. There are three types of poliomyelitis, for example, known as types I, II, and III. Antibody produced against type I protects only against type I, which is why vaccines against poliomyelitis must contain all three viruses. There are nearly 100 types of common cold viruses. An effective vaccine against the common cold would have to contain all of these types, and there are technical reasons why this cannot be done, or the person to be protected would require a large number of injections, each containing a few common cold viruses.

Duration of active immunity. In many cases, active immunity is lifelong, as with measles or German measles (rubella). In other instances, it can be short-lived, lasting not more than a few months. Most measurements of antibody are of antibody in the blood, or serum antibody, but, in some diseases, antibody produced elsewhere may be more important. In respiratory viral infections, for example, antibody produced by some of the surface cells of the respiratory tract appears to be a more important part of the defense mechanism than serum antibody, and serum antibody may not be able to reach the respiratory surfaces from the bloodstream. (A.B.C./R.D.F./Re.G.)

The persistence of acquired immunity is related not only to the level of circulating antibody (immunoglobulin) but also to sensitized T lymphocytes (cellular immunity). Although both cellular immunity and humoral (B-cell) immunity are important, their relative significance in protecting a person against disease varies with particular microorganisms. For example, antibody is of great importance in protection against common bacterial infections such as pneumococcal pneumonia or streptococcal disease, whereas cellular immunity is of greater importance in protection against viruses such as measles or against the bacteria that cause tuberculosis. (R.D.F./Re.G.)

IMMUNIZATION

Antibodies are produced in the body in response to the stimulus either of infection with an organism or, through vaccination the administration of a suspension of an organism, dead or alive, by mouth or by injection. When given alive, the organisms are weakened, or attenuated, by some laboratory means so that they still stimulate antibodies but do not produce their characteristic disease. However stimulated, the antibody-producing cells of the body remain sensitized to the infectious agent and can respond to it again, pouring out more antibody. One attack of a disease, therefore, often renders a person immune to a second attack, providing the theoretical basis for active immunization by vaccines.

Antibody can be passed from one person to another, conferring protection on that second person. In such a case, however, the antibody has not been produced in the body of the second person, nor have his antibody-producing cells been stimulated. The antibody is then a foreign substance and is eventually eliminated from the body, and protection is short-lived. The most common form of this type of passive immunity is the transference of antibodies from a mother through the placenta to her unborn child. This is why a disease such as measles is uncommon in babies less than one year old. After that age, the infant has lost all of its maternal antibody and becomes susceptible to the disease unless protective measures are taken. Sometimes antibody is extracted in the form of immunoglobulin from blood taken from immune persons and is injected into susceptible persons to give them temporary protection against a disease, such as measles or German measles.

Many respiratory viral infections, including influenza, are primarily surface infections, attacking and multiplying in respiratory surface cells. Vaccines given by injection may stimulate serum antibody only, and this may be one reason why influenza vaccine is not highly successful. Influenza vaccines inhaled as mists or sprays may induce the production of local surface antibody and so be more effective than vaccines by injection. (A.B.C./R.D.F./Re.G.)

Generally, active immunization is offered before the anticipated time of exposure to an infectious disease. When unvaccinated people are exposed to an infectious disease, two alternatives are available: active immunization may be initiated immediately in the expectation that immunity can be developed during the incubation period of the disease, or passive immunity can be provided for the interim period and then active immunization can be given at an appropriate time. The antigens (foreign substances in the body that stimulate the immune defense system) introduced in the process of active immunization can be live attenuated viruses or bacteria, killed microorganisms or inactivated toxins (toxoids), or purified cell wall products (polysaccharide capsules, protein antigens).

The ideal vaccine

There are five basic requirements for an ideal vaccine. The agents used for immunization should not in themselves produce disease. The immunizing agent should induce long-lasting, ideally permanent, immunity. The agent used for immunization should not be transmissible to susceptible contacts of the person being vaccinated. The vaccine should be easy to produce, its potency should be easy to assess, and the antibody response to it should be measurable with common and inexpensive techniques. Finally, the agent in the vaccine should be free of contaminating substances.

The route by which an antigen is administered frequently determines the type and duration of antibody response. For example, intramuscular injection of killed poliomyelitis virus (Salk vaccine) produces transient serum antibody production and induces temporary systemic immunity; it does not produce local gastrointestinal immunity and, therefore, does not prevent the carrying of the virus in the gastrointestinal tract. Live, attenuated, oral poliomyelitis virus (Sabin vaccine) induces both local gastrointestinal and systemic antibody production and, thus, immunization by mouth is preferred.

The schedule by which a vaccine is given depends upon the epidemiology of the naturally occurring disease, the

duration of immunity that can be induced, the immunological status of the host, and, in some cases, the availability of the patient. Measles, for example, is present in many communities and poses a potential threat to many children over five months of age. A substantial number of infants, however, are born with measles antibody from their mothers, and this maternal antibody interferes with an adequate antibody response until they are between 12 and 15 months of age. Generally, the immunization of infants after the age of 15 months benefits the community at large. In measles outbreaks, however, it is advisable to alter this schedule and immunize all infants between six and 15 months of age.

Diphtheria toxoid. The introduction of diphtheria toxoid more than 50 years ago led to a dramatic reduction in the incidence of the disease in many parts of the world. Primary prevention programs consisting of community-wide routine immunization of infants and children have largely eliminated the morbidity and mortality previously associated with diphtheria. Although the reported annual incidence of diphtheria has been relatively constant since the 1960s, local epidemics continue to occur. A complacent attitude in some nations largely reflects a lack of awareness of the public health hazard that can arise if the proportion of susceptible individuals is significant enough to allow renewed outbreaks. Also, diphtheria will continue to represent a significant public health problem because adequate immunization does not completely eliminate the potential for transmission of the bacteria *Corynebacterium diphtheriae.* Carriage of *C. diphtheriae* in the nose or throat has been well documented in fully immunized persons who clearly may transmit the disease to susceptible individuals.

Diphtheria vaccines

The vaccine—prepared by the treatment of *Corynebacterium diphtheriae* toxin with formaldehyde—is available in both fluid and adsorbed forms, the latter being recommended. Diphtheria toxoid is also available combined with tetanus toxoid and pertussis vaccine (DTP), combined with tetanus toxoid alone (DT), and combined with tetanus toxoid for adults (Td). The Td preparation contains only 15 to 20 percent of the diphtheria toxoid present in the DTP vaccine and is more suitable for use in older children and adults.

Whooping cough vaccine. The number of cases of whooping cough (pertussis), a serious disease that is frequently fatal in infancy, can be dramatically reduced by the use of the pertussis vaccine. The pertussis immunizing agent is included in the DTP vaccine. Active immunity can be induced by three injections given eight weeks apart. Adverse reactions can include high fever, collapse or a shocklike state, or neurological symptoms such as screaming episodes, isolated convulsions, and an altered level of consciousness. Serious neurological complications following pertussis immunization occur approximately once in every 180,000 doses administered. Other side effects are mild and self-limited.

Tetanus toxoid. The efficacy of active immunization against tetanus was illustrated most dramatically during World War II, when the introduction of tetanus toxoid among military personnel virtually eliminated the occurrence of the disease as a result of war-related injuries. Since then, the routine immunization of civilian populations with tetanus toxoid has resulted in decreased incidence of tetanus. In the United States in 1981, for example, only 72 cases of tetanus were reported, the majority of them occurring in persons more than 50 years of age. In virtually all cases, the disease has been reported in unimmunized or inadequately immunized individuals.

Because it provides long-lasting protection and relative safety in human beings, tetanus toxoid has proven to be an ideal vaccine. Tetanus toxoid is available in both vaccine fluid and alum-precipitated preparations. Commercially, tetanus toxoid is available in DTP, Td, and T (tetanus toxoid, adsorbed) preparations. DTP is recommended in infancy, while the Td form is recommended at 12 and 18 years of age and only once every 10 years thereafter. If a person sustains a wound prone to tetanus (such as puncture wounds or wounds contaminated with animal excreta), Td is given along with tetanus immune globulin (TIG) to prevent occurrence of the disease.

Polio vaccine. The value of primary prevention of disease through active immunization programs has been most convincingly demonstrated in the case of poliomyelitis. Before the vaccine was known, more than 20,000 cases of paralytic disease occurred in the United States alone every year. With use of the vaccine after 1961, by 1981 a record low of only six cases was recorded. Such achievement toward eliminating the clinical disease does not justify a casual attitude toward compliance with the recommended polio vaccine schedules, however. Low immunization rates are still evident in children in certain disadvantaged urban and rural groups, among which most of the cases of paralytic disease continue to occur.

Live trivalent oral poliovirus vaccine (TOPV) is used for routine mass immunization but is not recommended for patients with altered states of immunity (for example, those suffering from cancer or a congenital or acquired immune deficiency disease, or those receiving immunosuppressive therapy) or for children whose siblings are known to suffer an immune deficiency disease. Inactivated poliovirus vaccine (PV) is used for immunodeficient or immunosuppressed patients and for the primary immunization of adults because of their greater susceptibility to paralytic disease.

German measles (rubella) vaccine. A major epidemic in the United States in 1964 resulted in more than 20,000 cases of congenital German measles. In consequence, active immunization programs with attenuated German measles vaccine were initiated in 1969 in an attempt to prevent an expected epidemic in the early 1970s. The immunization of all children from one to 12 years of age was aimed at reducing the reservoir and transmission of wild German measles virus and, secondarily, at diminishing the risk of clinical German measles in susceptible pregnant women. This national policy contrasts with that of the United Kingdom, where only girls from 10 to 14 years of age who do not have detectable antibody levels to German measles virus are immunized. Proponents of the British policy have argued that natural German measles infection, which confers lifelong immunity, is more effective than vaccine-induced protection of uncertain duration and that continued outbreaks of German measles in the United States (largely in older children and young adults) since the introduction of German measles vaccines attest to the difficulty of achieving herd (group) immunity for this disease.

German measles vaccines

Live attenuated German measles vaccine is available in combination with measles vaccine (MR) and in combination with measles and mumps vaccines (MMR). For routine infant immunization MMR is given one time at 15 months of age. German measles vaccination can be accompanied by mild joint pain and fever in 5 percent of those who receive it. Vaccination is recommended for all children between the ages of 12 months and puberty. Vaccination is not recommended for pregnant women because fetal infection with the attenuated virus is potentially damaging. A number of women, however, have inadvertently received German measles vaccine during pregnancy with no harm to their fetuses being noted.

Measles (rubeola) vaccine. The licensing and distribution of killed measles vaccine in 1963, followed by the development and widespread use of live attenuated vaccine, have sharply reduced the prevalence of measles and its related morbidity and mortality in many parts of the world. Additional attenuated vaccine preparations have been developed; though not fully evaluated, they appear to be safe and highly effective and to confer prolonged, if not lifelong, protection.

Despite the introduction of effective immunizing agents, measles continues to be a major public health concern. The continued occurrence of outbreaks of measles, especially among young adults, emphasizes the probable failure of herd immunity to eliminate measles transmission, despite high local immunization rates in young children. The outbreaks also indicate the possibility that a small number of appropriately immunized individuals may not develop solid immunity.

Measles vaccine is commercially available in live attenuated form and is used routinely in the MMR preparation

for infant immunization at 15 months of age. Susceptible older children or adults who have not had measles or have not previously received measles vaccine also should receive a single dose.

Mumps vaccine. Mumps is generally a self-limited disease in children but occasionally is moderately debilitating. A live attenuated mumps vaccine is available alone or in combination with measles and German measles vaccines. No serious adverse reactions have been reported following mumps immunization.

Pneumococcal vaccine. *Streptococcus pneumoniae* (pneumococcus) is the most frequent cause of bloodstream infection, pneumonia, and ear infection and the third most common cause of bacterial meningitis in children. There are 84 different types of pneumococci. Pneumococcal infection is particularly serious among the elderly and among those children with sickle-cell anemia, with congenital or acquired defects in immunity, without spleens, or with abnormally functioning spleens.

Immunity after pneumococcal disease is type-specific and lifelong. The pneumococcal vaccine now available consists of antigens from 14 of the most common types of pneumococci that cause disease. It can be given to children two years of age or older or to adults in a single intramuscular injection. The duration of protection is unknown.

Meningococcal vaccine. *Neisseria meningitidis* can cause severe bloodstream infection and meningitis (infection of the coverings of the brain and spinal cord), known as meningococcal disease. In the general population, fewer than one per 400,000 persons is attacked by the microorganism, while among those less than one year of age, the ratio rises to one per 100,000. In a day-care centre in which a primary case of meningococcal disease has occurred, the ratio has been reported at two to 100 per 100,000, and the danger from household contact from an infected person is believed to be 200 to 1,000 times greater than that of the general population. Because of these statistics, anyone who has had contact with the disease in a home, day-care centre, or nursery school should receive antibiotic treatment as soon as possible, preferably within 24 hours of the diagnosis of the primary case of the disease.

Group-specific meningococcal polysaccharide vaccines can be used to control outbreaks of disease and may benefit travellers to countries where these diseases are endemic. Certain of the vaccines are administered routinely to military recruits in the United States.

Hepatitis B vaccine. Hepatitis B virus (HBV) produces an illness characterized by jaundice, poor appetite, malaise, and nausea. Chronic liver disease may follow the infection.

Risk factors for hepatitis B

Hepatitis B vaccine is recommended for persons who are at a greater risk of contracting the disease because of their life-styles or jobs. These include health-care personnel who are exposed to blood products, hemodialysis patients, institutionalized patients and their staffs, patients receiving multiple transfusions, prostitutes and the sexual partners of individuals with the disease, users of illicit intravenous drugs, homosexual males, and newborn infants of mothers with hepatitis B infection.

Hepatitis B vaccine, which is quite expensive, consists of inactivated viral surface antigen particles. Its immediate side effects are minimal and no long-term reactions have been reported.

Influenza vaccine. Influenza is characterized by the acute onset of fever, chills, malaise, headache, diffuse muscle ache, cough, and sore throat. There are high morbidity and mortality rates among patients with recurrent heart failure; chronic pulmonary disease such as cystic fibrosis, poorly controlled asthma, bronchopulmonary dysplasia, or bronchiectasis; chronic kidney or nephrosis disease; metabolic disorders such as diabetes mellitus; or malignancy or other immunodeficiency states.

The manufacture of influenza vaccine is complicated by the many influenza viruses and by the major shifts in chemical configuration that these viruses undergo every five to 10 years. Routine immunization against influenza viruses is not recommended for healthy individuals. During epidemics, those at high risk of developing severe influenza or complications should receive the currently available vaccine before the respiratory disease season.

Experimental vaccines. The low morbidity of chicken pox (varicella) in healthy children does not support the routine use of chicken pox vaccine. In certain persons, especially those with immunodeficiency disease or cancer, however, chicken pox can be devastating. Work is therefore being done to develop a chicken pox vaccine for this highly susceptible population. A live attenuated vaccine has been found to be safe and immunogenic in healthy children, and investigators are testing the vaccine in individuals with compromised immunity.

Cytomegalovirus is the leading cause of congenital deafness and retardation as well as the cause of other defects. A live attenuated virus has been shown to be safe and immunogenic in a limited trial. Until the interaction between virus, mother, and fetus is better understood, however, the vaccine is of questionable usefulness.

The bacteria *Haemophilus influenzae* is a major cause of morbidity and mortality in children, particularly in those under six years of age. Because it is highly contagious among people in close contact with one another, antibiotics are used to prevent infection. Work is proceeding on development of a suitable vaccine. Studies show that an antibody response to a vaccine containing components of the capsule of *H. influenzae* is not reliable in children under 18 months of age. Whether this antigen can be coupled with a suitable adjuvant, such as a lipid or DTP, to make it more immunogenic has not been determined.

Passive immunity. Passive immunity is the administration of antibodies to an unimmunized person from an immune subject to provide temporary protection against a microbial agent, poison, or cell. The process is also used in the treatment of certain disorders associated with toxins (diphtheria, tetanus), in certain bites (snake and spider), and as a specific (Rho-GAM) or nonspecific (antilymphocyte serum) immunosuppressant. Other antibody preparations are available under specific conditions for specific disorders. Passive immunization is not always effective; the duration of immunity provided is brief and variable, and undesirable reactions may occur, especially if the antiserum is of nonhuman origin. Several preparations are available for use as passive immunizing agents.

Human immune serum globulin

Human immune serum globulin (HISG) is prepared from human serum. Special treatment of the serum removes various undesirable proteins and infectious viruses, thus providing a safe product for intramuscular injection. HISG is used for the treatment of antibody deficiency conditions and for the prevention of hepatitis A and hepatitis B viral infections, measles, chicken pox, German measles, and poliomyelitis.

The most widespread use of HISG is in the prevention of hepatitis A infection, a disease for which there is no active immunization available, in individuals known to have had intimate exposure to the disease. Hepatitis B immunoglobulin should be given immediately to persons who are exposed to contaminated blood or who have had intimate physical contact with a person who has hepatitis B infection. Because of the scarcity of the product and the lack of its proven efficacy, hepatitis B immune globulin is not recommended routinely for those who are continuously at high risk of exposure to hepatitis B. It should be given, however, to infants born to mothers who have serological evidence of hepatitis B viral infection.

Passive immunity can be conferred on persons who are exposed to measles, mumps, whooping cough, poliomyelitis, rabies, German measles, tetanus, chicken pox, and shingles (herpes zoster). Several investigators have claimed a beneficial effect of HISG in persons suffering asthma and other allergic disorders; evidence confirming its efficacy in these conditions is lacking, however. Monthly HISG is not beneficial in the prevention of upper respiratory infections, otitis, skin infections, gastrointestinal disorders, or fever of undetermined cause. HISG has been used inconclusively in the treatment of infants with significantly low levels of immunoglobulins and patients with severe burns who are at an increased risk of infection. Antivenoms derived from horses are used effectively to treat snake or spider bites, but not without significant risk of reaction to the equine antibody preparation.

Rho-GAM is a human anti-Rhesus immune serum globu-

lin used in the prevention of Rh hemolytic disease of the newborn. Rh positive is given to an Rh-negative mother after the delivery of an Rh-positive infant or after miscarriage or abortion.

Botulism, a severe paralytic poisoning, results from the ingestion or absorption of the toxin of the bacteria *Clostridium botulinum.* As a preventive measure, antitoxin can be given to individuals known to have ingested contaminated food and to patients with symptoms as soon as possible after exposure.

Most of the damaging effect of diphtheria results from the toxin produced by the bacteria *Corynebacterium diphtheriae.* This toxin not only has local effects but also is distributed through the blood to the heart, nervous system, kidneys, and other organs. Diphtheria antitoxin of animal origin remains the principal treatment, along with antibiotics.

Gas gangrene is caused by infection with clostridial organisms, usually following a traumatic injury that has caused extensive local tissue damage. An antitoxin derived from horses is available as an adjunct to surgical and other treatment of these infections.

TEMPERATURE REGULATION

Although it has been generally accepted that the usual range of deep body temperature is from 97.2° F to 100° F (36.2° C to 37.8° C), temperatures of 95.9° F to 101.2° F (35.5° C to 38.4° C) have been reported in normal individuals. The rectal temperature of a person may read 101° F (38.3° C) in the afternoon or after exercise. Core body temperature, in fact, varies with time of day, environmental temperature, physical activity, food intake, and emotional state. If fever is defined as any elevation of body temperature "above normal levels," it cannot be equated necessarily with the presence of disease.

Body temperature. Body temperature depends upon a balance between heat production and heat loss. In human beings, an elaborate system maintains temperature within a narrow range despite wide variations in heat production or environmental temperature.

Each organ in the body contributes to heat production. The brain and muscles produce about one-sixth of the body's heat, the viscera one-sixth, and the heart one-tenth. During exercise, heat production by muscle can increase markedly. The newborn infant possesses an emergency heat-producing mechanism in the form of brown fat, which makes possible the doubling of the infant's basal metabolic rate without shivering.

Body heat may be lost through radiation, evaporation, convection, or conduction. In general, 60 percent of total heat loss occurs by radiation from the body's surface at a rate affected by the rate of blood flow to the skin's surface. Evaporation is the method of heat loss that can be regulated most readily; the sweat glands can dissipate as many as 1,700 kilocalories per hour. Insulation provided by fat and by intact skin acts to maintain normal body temperature when the body is exposed to cold or warm environments.

Hypothalamus

The thermoregulatory region of the human brain is called the hypothalamus. Chemical substances called prostaglandins may serve as molecular transmitters of stimuli to the area of the hypothalamus concerned with temperature control. Elevation in the temperature of blood going to the hypothalamus is the most important factor in the central control of body temperature, initiating heat loss by causing dilation of blood vessels and increasing heat loss by sweating.

Fever. Fever generally results from a disturbance in the regulation of body temperature by the central nervous system. The hypothalamic thermostat reacts as if the "set point" had been raised. When the temperature of the body reaches the newly established set point, heat loss and gain are balanced again, but at a higher level. During disease, chemicals called endogenous pyrogens, which are derived from white blood cells, are the principal factors responsible for fever.

Whether fever should be treated or not is a question of considerable interest. It is unclear whether fever is an essential defense mechanism for human beings or represents an important biological response to infection. Favourable effects of fever have been noted in patients with syphilis of the central nervous system and in gonococcal infections. Major elevations in body temperature during the course of these infections may help to slow the growth of, or kill, these organisms. There is no firm evidence, however, to support the concept that fever enhances the specific mechanisms of resistance. Conversely, selected studies suggest that fever may be associated with an increased mortality rate during the course of certain infections. It is recommended that "normal" body temperature be maintained by the use of medications such as acetaminophen or aspirin once a specific diagnosis had been established, thereby helping to ensure the comfort of the patient.

Infections of specific organ systems

UPPER RESPIRATORY TRACT INFECTIONS

The common cold. The term common cold reflects the feeling of chilliness on exposure to cold that is part of the onset of symptoms. The feeling was originally believed to have a cause-and-effect relationship with the disease, but this is now known to be incorrect. More than 200 years ago, Benjamin Franklin pointed out that colds are caught from other people, not from exposure to a cold environment. The common cold is an acute, communicable viral disease characterized by nasal stuffiness, sneezing, runny nose, throat irritation, and, sometimes, fever. There are more than 100 agents that cause the illness, including parainfluenza, influenza, and respiratory syncytial viruses, as well as reoviruses. Rhinoviruses, however, are the most frequent cause of the common cold.

The common cold is an illness that occurs seasonally. Young children can contract between three and eight colds a year, usually coming into contact with the infectious agents in day-care centres or preschools. All available evidence indicates that cold weather, chilled wet feet, and drafts do not cause or increase the susceptibility of people to colds. The usual duration of the illness is about five to seven days, but lingering cough and postnasal discharge may persist for two weeks or more.

Diagnosis of a cold is usually made by medical history alone, although it is possible to take a culture for viruses. There is no effective antiviral agent available for the common cold. Therapy consists of treating the symptoms—relieving aches, fever, and nasal obstruction. One of the greatest medical controversies in the 1970s concerned the efficacy of vitamin C in the prevention or treatment of the common cold. In two carefully controlled studies, administration of ascorbic acid (vitamin C) failed to prevent or decrease the symptoms of the common cold.

Sinusitis. Sinusitis is the inflammation of the mucosal lining of one or more sinuses. It commonly accompanies upper respiratory viral infections and in most cases requires no treatment. Purulent (pus-producing) sinusitis can occur, however, requiring treatment with antibiotics.

Origin of sinusitis

The origin of sinus infection is much like that of ear infection. Normally the middle ear and the sinuses are sterile, but the adjacent mouth and nose have a varied bacterial flora. Under the usual conditions, very small hairs called cilia move mucus along the lining of the nose and respiratory tract, keeping the sinuses clean. When the ciliary function is damaged, infection can be established. Following a common cold caused by a viral agent, a decrease in ciliary function may permit bacteria to remain on the mucous membrane surfaces within the sinuses and to produce a purulent sinusitis. The organisms usually involved are *Haemophilus influenzae, Streptococcus pneumoniae, Staphylococcus aureus, Streptococcus pyogenes,* and many other penicillin-sensitive anaerobes. Common symptoms include facial pain, headache, and fever following previous upper respiratory viral illness.

On physical examination, persons suffering sinusitis are usually found to have an elevation in body temperature, nasal discharge, and sinus tenderness. The diagnosis can be confirmed by X-rays of the sinuses and cultures of material obtained from within the sinuses. Antibiotic treatment appropriate to the organism that is recovered in cultures is provided for a period of two weeks or longer.

Pharyngitis (tonsillitis). Pharyngitis is an inflammatory illness of the mucous membranes and underlying structures of the throat. Inflammation usually involves the nasopharynx, uvula, and soft palate. The illness can be caused by bacteria, viruses, mycoplasmas, fungi, and parasites, and by recognized diseases of uncertain causes. More than 90 percent of all tonsillitis is caused by *Streptococcus pyogenes,* adenoviruses, influenza viruses, parainfluenza viruses, Epstein-Barr virus, enteroviruses, and *Mycoplasma pneumoniae.* The streptococcus by far receives the most attention from physicians because it is recovered readily by throat culture and can be treated with penicillin. The association of this bacterium with scarlet fever, erysipelas (a contagious infection of the skin and underlying tissue), rheumatic fever, and kidney problems requires aggressive treatment of streptococcal pharyngitis or streptococcal skin infection.

Symptoms of pharyngitis are often sore throat, fever, and headache. There may be redness and swelling of the tonsils, fever, and palpable lymph nodes in the neck. The diagnosis is established by a detailed medical history and by physical examination; the cause of pharyngeal inflammation can be determined by throat culture. Usually only the symptoms can be treated—with throat lozenges to control sore throat and acetaminophen or aspirin to control fever. If a diagnosis of streptococcal infection is established by culture, appropriate antibiotic therapy, usually with penicillin, is instituted.

Otitis media. Otitis media is an inflammation of the lining of the middle ear and is one of the most common infections in childhood. In its acute form, it commonly develops in association with an infection of the upper respiratory tract that extends from the nasopharynx to the middle ear through the eustachian tube.

The organisms that cause the disease in children under six years of age most commonly are the bacteria *Streptococcus pneumoniae* and *Haemophilus influenzae.* Symptoms include fever and earache. The diagnosis is established by a careful examination of the tympanic membrane (eardrum) and by techniques (tympanometry) that can provide evidence of fluid behind the eardrum. Antibiotics generally are given for acute otitis because the infection can spread to the nearby bones and the central nervous system. The disease can be complicated by perforation of the eardrum and, in rare cases, by permanent hearing losses that lead to delay in the development of speech and language.

Epiglottitis. Acute epiglottitis is a serious bacterial infection almost always caused by *Haemophilus influenzae,* type B. It is characterized by marked swelling of the epiglottis and high fever. The onset is usually abrupt, with high fever and breathing difficulties. Because of the marked swelling of the epiglottis, there is obstruction at the opening of the trachea (windpipe), making it necessary for the patient to sit and lean forward to maximize the airflow. Diagnosis is made by medical history and can be established by X-rays of the neck or by seeing an enlarged, swollen, red epiglottis.

Epiglottitis is a life-threatening disorder and physicians must keep the airway unobstructed and keep the patient quiet. Antibiotics active against *Haemophilus influenzae* are given intravenously and a hollow tube may be passed from the patient's nose into the trachea to maintain the flow of air. Acute, sudden obstruction of the airway is the most serious complication, but with appropriate treatment recovery is almost always achieved.

Types of croup

Croup. The word croup is a nonspecific term used to identify several different respiratory illnesses that are characterized by varying degrees of difficulty in breathing or by cough and hoarseness resulting from obstruction in the region of the larynx. Many different viruses are responsible, the most frequent being the parainfluenza and influenza viruses. Croup occurs most frequently in children under three years of age, with most cases occurring in the early winter months.

The symptoms of croup are often those of a cold: noisy breathing, hoarseness, and low-grade fever. The diagnosis is established by medical history, physical examination, and characteristic findings in X-rays of the neck. Treatment consists of adding humidity to the air to be breathed by the patient and observation to make certain that the respiratory tract does not suddenly become blocked. Aerosolized adrenalin (epinephrine) and steroids have also been used, although there is no conclusive proof of their effectiveness.

LOWER RESPIRATORY TRACT INFECTIONS

Bronchitis. Acute bronchitis is a communicable disease of infancy characterized by cough, runny nose, low-grade fever, and wheezing. It can be caused by many viruses, bacteria, and mycoplasmas. The symptoms include fever and cough, with production of sputum. Diagnosis is established by medical history and physical examination of the chest. Treatment usually consists of the administration of fluids, humidified air, decongestants, and antibiotics, the last if a bacterial cause is suspected.

Asthmatic bronchitis is an acute, recurrent illness characterized by cough and wheezing; it commonly occurs in children under five years of age who are allergic to dust, molds, or other agents. It is usually caused by viruses, such as rhinoviruses and respiratory syncytial virus, but other infectious agents, such as bacteria and mycoplasmas, may also be involved. In asthmatic bronchitis, there is a tendency to wheeze with each upper respiratory viral infection. Children who have a propensity for asthmatic bronchitis usually improve with age. Treatment consists of the administration of oxygen, humidity, and agents that dilate the bronchii.

Bronchiolitis. Bronchiolitis is an acute, communicable disease of infancy, characterized by cough, runny nose, low-grade fever, and wheezing. The major cause of the illness, respiratory syncytial virus, is also responsible for community epidemics of bronchiolitis. Treatment usually consists of the administration of humidified oxygen and observation for possible respiratory distress, although almost all cases are treated without complications. If bacterial pneumonia occurs secondarily to the bronchiolitis, antibiotics are required. An aerosolized antiviral drug shortens the course of the disease, but its use is considered to be experimental.

Pneumonias. Pneumonia is an inflammation of the lungs that can be caused by viruses, bacteria, fungi, mycoplasmas, rickettsias, *Chlamydia,* or parasites. It can also occur as a hypersensitivity, or allergic response, to agents such as mold, humidifiers, and animal excreta.

Viral pneumonias

Viral pneumonias are primarily caused by respiratory syncytial, parainfluenza, and influenza viruses. Pneumonias with similar symptoms are also caused by mycoplasmas (the most common cause of pneumonia in young adults), two species of *Chlamydia,* and the rickettsial disease known as Q fever. Symptoms of these pneumonias include runny nose, decreased appetite, and low-grade fever, usually followed by respiratory congestion and cough. Diagnosis is established by physical examination and chest X-rays. Nonbacterial pneumonia is treated primarily with supportive care, while specific antimicrobial agents are used for mycoplasmas, *Chlamydia,* and rickettsias. In general, the prognosis is excellent.

Tuberculosis should always be considered a possibility in any patient with pneumonia, and skin testing is included in the initial examination of patients with lung problems. Fungal infections such as coccidioidomycosis and histoplasmosis should also be considered, particularly if the patient was recently exposed to excavations, backyard swimming pools, old sheds or barns, or dust storms. Other fungal and protozoan parasites (such as *Pneumocystis carinii*) are common in patients receiving immunosuppressive drugs or in patients with cancer or other chronic diseases.

Hypersensitivity pneumonias are a spectrum of disorders that arise from an allergic response to the inhalation of a variety of organic dusts. Interest in these disorders began in 1932 with the first description of "farmer's lung," a disease caused by exposure to moldy hay. Since then, similar diseases have been seen following exposure to moldy sugarcane, room humidifiers, and air-conditioning ducts, all of which contain the fungus *Actinomyces.* Other fungi found in barley, maple logs, and wood pulp may cause similar illnesses. In addition, people exposed to rats,

gerbils, pigeons, parakeets, and doves may develop manifestations of hypersensitivity pneumonia. Initially, these patients experience fever with chills, cough, shortness of breath, headache, muscle pain, and malaise, all of which may subside in a day if there is no further exposure. A more insidious form of hypersensitivity pneumonia is associated with persistent malaise, fever, weight loss, and cough. Diagnosis is established by medical history, physical examination, and specific laboratory tests. Treatment consists of removing the patient from the offending environment, bed rest, and supportive care. Bacterial pneumonia has been a frequent and serious human illness throughout recorded history. Examination of Egyptian mummies has revealed that some of them died of the illness. The disease was well known to the Greeks and Romans; the symptoms and treatment of bacterial pneumonia were described by Hippocrates (*c.* 460–*c.* 377 BC).

Streptococcus pneumoniae is the leading bacterial cause of pneumonia. *Staphylococcus aureus, Haemophilus influenzae,* and *Legionella pneumophila* are other important causes. Patients with bacterial pneumonias typically suffer a sudden onset of high fever with chills, cough, chest pain, and difficulty in breathing. Diagnosis usually can be established by taking a culture of the organism from the patient's sputum and by chest X-ray examination. Treatment is with specific antibiotics and supportive care. In the United States the mortality rate of bacterial pneumonia is less than 1 percent, except in the elderly or in individuals with rapidly progressing, overwhelming disease.

Legionnaire's disease

Legionella pneumophila is the bacterium responsible for Legionnaire's disease. In 1976, 200 cases of respiratory illness caused by a then unknown agent occurred among 4,500 participants at an American Legion convention in Philadelphia. The disease was marked by high fever, chills, muscle aches, abnormal liver function, and changes in behaviour, in addition to cough and other pulmonary symptoms. X-rays of the chest revealed pneumonia. Most of the patients responded well to the antibiotic erythromycin; some died because, when the outbreak occurred, it was not known what the most appropriate type of antimicrobial therapy would be. Autopsies of those who died revealed the organism that is now called *Legionella pneumophila.* The form of pneumonia that occurred with Legionnaire's disease is particularly frequent among people who have cirrhosis of the liver caused by excessive ingestion of alcohol.

INFECTIONS OF THE CARDIOVASCULAR SYSTEM

Endocarditis. When bacteria circulating in the bloodstream adhere to the valves of the heart or other parts of the heart's inner lining (the endocardium), bacterial endocarditis occurs. With adherence, vegetations of bacteria and fibres become vegetations of dead tissue. Bacteria may adhere to the endocardium of damaged heart valves and cause erosion and destruction. Infections of the heart valves occur most often with bacteria that gain access to the blood after tooth extraction, intravenous therapy, surgery, or anemia, or after the use of bacteria-contaminated material in drug abuse. Bacterial endocarditis is frequent in patients who have suffered acute rheumatic fever or congenital heart disease.

Patients with bacterial endocarditis have symptoms of high fever, hemorrhages in the skin, meningitis, infectious arthritis, and a newly developed heart murmer. The disease is very serious and can cause cardiac failure. Small clots of thrombin (the enzyme of the blood that causes clotting) may break off and travel by the bloodstream to other parts of the body, cutting off the blood supply and causing tissue death. Diagnosis is established by blood culture and physical examination. Treatment consists of the use of intravenous antibiotics for four to six weeks or longer. If severe erosion of the valves and resultant cardiac failure occur, surgical replacement of the damaged valve with an artificial one may be necessary.

Fungal endocarditis primarily strikes people who have undergone heart surgery and can also occur in intravenous-drug abusers. Symptoms are the same as those of bacterial endocarditis. Treatment consists of the use of appropriate antifungal agents. Prognosis is poor, however, and survival rates are less than 50 percent.

Pericarditis. In pericarditis, the outer layer of the heart, or pericardium, is inflamed. Symptoms include chest pain, shortness of breath, and rapid, shallow respiration. Various infectious agents, as well as noninfectious diseases, can cause pericarditis. Treatment is supportive; appropriate antibiotics are used when specific bacteria are shown to be the cause of the infection.

Myocarditis. An inflammation of the muscular middle layer of the heart, or myocardium, is called myocarditis. The disease may result in no symptoms at all or in significant cardiac dysfunction. Viruses, bacteria, and noninfectious diseases can cause myocarditis. Diagnosis is made by physical examination and electrocardiographic tests. Treatment is primarily supportive and intensive medical care is the rule. Drugs may be used to improve the heart's ability to contract and to remove extra fluids from the body.

Rheumatic fever. Acute rheumatic fever is an inflammatory disease of the heart, joints, central nervous system, and subcutaneous tissues that develops after a throat infection with group A beta-hemolytic streptococcal organisms or after untreated "strep throat." Prevention is possible with penicillin but specific treatment is not available. Rheumatic fever is particularly important because of the heart disease that usually ensues.

Rheumatic fever symptoms

When a streptococcal throat infection goes untreated, most patients recover without complications. Some, however, develop fever, malaise, anorexia, and joint pain two to three weeks later. Upon examination, the striking findings are pallor, rapid pulse, muffled heart sounds, and heart murmur or murmurs. Other symptoms can include a rash called erythema marginatum and a movement disorder called chorea (Sydenham chorea, or St. Vitus' dance). Diagnosis is established on the basis of the specific symptoms and is supported by documentation of an earlier streptococcal infection. Selective laboratory tests support the clinical findings.

Treatment of rheumatic fever usually consists of the administration of penicillin and salicylates, along with supportive care. Rheumatic fever patients must receive antibiotics (usually penicillin) on a regular basis for the remainder of their lives because their damaged heart valves predispose them to the development of bacterial endocarditis.

CENTRAL NERVOUS SYSTEM INFECTIONS

Meningitis. Meningitis is an inflammation of the outer covering of the brain and spinal cord, or the meninges. It can be caused by virtually any known infectious agent, bacterial organisms producing the most life-threatening infections. The patient usually suffers fever, headache, intractable vomiting, irritability, anorexia, and stiffness in the neck.

Diagnosis of meningitis is made by examination and confirmed by the performance of a test called a spinal tap (or lumbar puncture). In this test a needle is inserted into the lower part of the patient's back and a small sample of cerebrospinal fluid (CSF)—the liquid that surrounds the entire central nervous system—is removed. If a bacterial, tuberculous, or fungal infection is found, patients need intensive medical care; appropriate antibiotics must be administered as soon as meningitis is suspected. The mortality and morbidity of the bacterial disease are substantial, even with the prompt use of appropriate antibiotic therapy.

The results of bacterial meningitis can include seizures, deafness, blindness, paralysis, and various degrees of impairment of the intellect. To control the spread of meningitis caused by *Haemophilus influenzae* or *Neisseria meningitis,* the antibiotic derivative rifampin should be administered to any who have come in contact with the disease.

In contrast, if meningitis is caused by viruses, no specific treatment is required. Patients recover from this illness in three to five days, rarely with any serious result.

Encephalitis and meningoencephalitis. An inflammation of the brain itself is called encephalitis. It is most often caused by a group of arboviruses, such as those that cause the Western equine, Eastern equine, St. Louis, California,

and Powassan encephalitides. Common symptoms include drowsiness, paralysis, coma, convulsions, seizures, mental confusion, and speech problems. Meningoencephalitis is an inflammation of the brain, the meninges, and the spinal cord. It is generally accompanied by fever, headache, vomiting, and stiffness in the neck and back.

Diagnosis of encephalitis can be confirmed by a spinal tap. The cause of the disease can be established by culture for the particular virus involved or by isolation of the virus by a biopsy of the brain tissue. Treatment of encephalitis is largely supportive because no specific antiviral agents, except for that which works against herpes simplex virus, are available for therapy. Patients require intensive medical care, with continuous monitoring of their heart and respiratory functions and management of their fluid and electrolyte balances. Prognosis depends upon the specific viral agent that caused the encephalitis; mortality can be as high as 70 percent.

The most common results of encephalitis include convulsions, hearing loss, speech difficulties, paralysis, spasticity, and mental retardation. Prevention can be achieved by immunizing the horses that carry the arboviruses and by mosquito control.

GENITOURINARY TRACT INFECTIONS

Urethritis. Urethritis is an inflammation of the duct that leads from the bladder to the body's exterior. Its causes vary with age, sexual practices, and hygienic standards. Urethritis due to fecal contamination or irritation due to physical or chemical substances is common in young children. After puberty, the most common cause of the condition is gonorrhea, an infection spread through sexual intercourse (see below *Sexually transmitted diseases*). It can also be caused by trauma, the introduction of foreign bodies into the urethra, bubble bath, or soap.

Diagnosis of urethritis is established by taking a culture of urethral discharges. Treatment may simply involve the withdrawal of the offending chemical agent, when the inflammation is caused by chemical irritation, or the administration of antibiotics, when *Neisseria gonorrhoeae* is involved. If it is suspected that the bloodstream is also involved, the patient must be hospitalized and treated with intravenous antibiotics.

Cystitis. Cystitis is an inflammation of the bladder, usually caused by bacteria. The term is also often used to denote infection of the urinary tract, a common illness in females. When cystitis occurs in very young males, an underlying structural abnormality of the genitourinary system should be suspected.

The symptoms of cystitis usually include frequent and painful urination and low-grade fever. Diagnosis is established by culture of the urine, which is normally sterile. Antibiotics are used to treat bacterial cystitis.

Pyelonephritis. Bacterial infection of the kidney is called pyelonephritis, or upper urinary tract infection. Symptoms usually include fever, chills, and pain in the flanks. Diagnosis is established by culture of the urine. If the organism causing the pyelonephritis escapes into the bloodstream, the patient may require intravenous antibiotics.

Prostatitis. When the prostate, the gland in males located just below the bladder and surrounding the urethra, is inflamed, the condition is called prostatitis. Symptoms include fever, malaise, muscle aches, groin pains, a sensation of pressure in the lower abdomen, and burning upon urination. Physical examination reveals a very tender, warm, swollen, and firm prostate.

Prostatitis may be caused by bacteria, viruses, or fungi, and the specific infectious agent is seldom identified. Antibiotics are used if a bacterial cause is suspected, and analgesics, stool softeners, and bed rest are also prescribed.

Sexually transmitted diseases. The sexually transmitted, or venereal, diseases include syphilis, gonorrhea, herpes simplex viral infection, venereal warts, and lymphogranuloma venereum.

Syphilis

Syphilis is perhaps the oldest known sexually transmitted disease, although the mode of transmission was not recognized until the 18th century. Over the centuries, it has been called the "Italian disease," the "French disease," and the "great pox," among other names. It is caused by a slender, coiled, flexible bacteria called *Treponema pallidum* and is primarily transmitted by sexual intercourse.

Untreated syphilis progresses through three or four stages over a period of many years. Primary syphilis results in a single, firm, nontender nodule, or chancre, that is usually located in the genital area. If left untreated, secondary syphilis may begin several days later, characterized by fever, sore throat, enlargement of the lymph nodes, headaches, and rash. After this phase, syphilis enters a latent period that lasts from three to 10 years. Tertiary syphilis is characterized by involvement of the central nervous and cardiovascular systems, which can be devastating if untreated.

Syphilis is transmitted to the fetuses of untreated pregnant women. About one-fourth of these fetuses die and are spontaneously aborted, while another one-fourth die in the immediate neonatal period. Those fetuses that survive may be born with rashes, pneumonia, and skeletal abnormalities. If congenital syphilis is not treated, blindness, deafness, perforation of the palate, inflammation of the liver, and involvement of the central nervous system may ensue.

The prognosis for syphilis varies with the promptness of diagnosis and treatment. A specific diagnosis is established by testing the blood and by examining the infectious organism with fluorescent electron microscopy. Treatment with penicillin is effective. In several countries, programs to prevent the spread of syphilis provide confidential counseling and treatment for people with the disease and for those who have come in contact with it. Pregnant women are tested for syphilis and, if they are found to be infected, they are treated to protect their fetuses.

Gonorrhea

Gonorrhea is an infection caused by the bacterium *Neisseria gonorrhoeae.* Usual symptoms are a urethral discharge and discomfort during urination. The bacterium can sometimes enter the bloodstream, causing infection in virtually every organ system. Common symptoms after invasion of the bloodstream are fever, swelling of the joints, eye infection, and skin rash. Diagnosis is established by culture, and the disease usually can be treated with penicillin. Tetracycline generally provides effective therapy for penicillin-resistant strains.

Herpes simplex viruses (types 1 and 2) cause infections in virtually all human beings. Herpes simplex type 1 is responsible for oronasal ulcers ("cold sores," or "fever blisters") while herpes simplex 2 usually is the cause of urogenital infections. The latter is highly contagious and may be transmitted by individuals who are carriers of the organism but who remain asymptomatic. When symptoms occur, fever and malaise are followed by burning pain in the genital area and enlargement of the lymph nodes in the groin. Blisters and small ulcers are usually found in the area of infection, and there is severe pain and burning upon urination. A cure is not known, although an antiviral drug may shorten the duration of the symptoms if it is applied early in the course of the disease.

Venereal warts, or condylomata acuminata, are wartlike growths in the pubic area that are accompanied by itching and discharge. Cryotherapy (freezing) or the application of an ointment is used to treat the disease. Warts can also occur as a result of molluscum contagiosum, a viral condition that is treated by cryotherapy or curettage.

Lymphogranuloma venereum is caused by the bacteria *Chlamydia trachomatis.* It is characterized by blisters, ulcers, and abscesses; if it is untreated, local tissue is destroyed and infected lymph nodes become enlarged. Diagnosis is made by a specific serological test or a skin test. Antibiotics are used for treatment.

GASTROINTESTINAL INFECTIONS

Cholera. Since 1817, the history of cholera has been one of periodic devastating epidemics over large geographical areas, particularly in India and Bangladesh. Cholera is caused by the bacterium *Vibrio cholerae* and the toxin that it produces. The disease is accompanied by severe diarrhea; it is possible for a patient to lose a volume of fluid equivalent to total body weight in one or two days. Unless fluid is promptly replaced, the disease can be fatal.

Cholera is acquired by ingesting water that is contami-

nated with *Vibrio cholerae.* When these organisms reach the bowel, they multiply rapidly and adhere to the lining of the intestinal wall, producing a toxin that causes a tremendous outpouring of fluid from the body into the intestinal tract. Symptoms at this stage include vomiting and "rice water" stools. If lost fluids are not replaced, the patient rapidly dehydrates, and sunken eyes, dry mouth, cold extremities, and depressed blood pressure develop. Children may also suffer fever, lethargy, and seizures as a result of the extreme dehydration. If fluid is not provided, the patient develops shock and dies.

Proposed regimens for the treatment of cholera filled 15 pages of one treatise on the subject in 1850. Adequate therapy, however, is simply that of replacing fluids and electrolytes. In the United States, for example, where treatment is given intravenously, mortality from cholera has been reduced to almost zero. Fluids can be provided entirely by mouth if the patient is sufficiently alert and can be enticed to drink the large volumes of fluid required. Oral therapy has proven extremely effective if intravenous therapy is not available. Antibiotics such as tetracycline shorten the duration and severity of the illness. Cholera immunization is not entirely effective and the duration of immunity is brief.

Escherichia coli. *Escherichia coli* are ubiquitous organisms that are an integral part of the normal gastrointestinal flora. They can be transmitted by water, milk, food, or flies and other insects. The organisms cause diarrhea in three ways: by giving off toxins, by directly invading the mucosal lining of the intestinal wall, or by adhering to the intestinal wall. When disease is due to *E. coli* toxin, the patient develops watery, yellow stools, without mucus or blood, associated with fever and vomiting. With invasive *E. coli* disease, the patient develops dysentery with fever, malaise, abdominal cramps, and a bloody diarrhea with mucus in the stool. When *E. coli* adheres to the intestinal wall, the diarrhea may be clear or filled with blood and mucus. Therapy is largely that of replacing lost fluids, although specific antimicrobial agents can be effective in some cases. The illness is self-limited and there is no evidence of long-lasting effects.

Salmonella. *Salmonella* bacteria are pathogens (microbes that cause disease) important to human beings and animals. Most infections by these bacteria result from the ingestion of contaminated food or water. Many different diseases are caused by the more than 1,400 serotypes of this bacteria genus.

Salmonella gastroenteritis results in the abrupt onset of nausea, cramplike abdominal pain, and loose, watery diarrhea with blood and mucus. It is generally a self-limited illness; no treatment is required and death is rare. In infants under one year of age, however, invasion of the bloodstream frequently occurs and antibiotic treatment is recommended.

Typhoid fever

Typhoid fever, caused by a specific serotype called *Salmonella typhosa,* is endemic in many countries. The illness is accompanied by high fever, chills, headache, abdominal pain, and, usually, constipation. An enlarged spleen, abdominal tenderness, and a typical rash also often occur. Diagnosis is made by blood culture (because the microbe invades the bloodstream at some time during the course of the illness), stool culture, and serological testing. The treatment of typhoid fever requires specific antibodies, which reduces mortality to less than 1 percent.

Life-threatening complications of typhoid fever can include intestinal perforation or hemorrhage. *Salmonella typhosa* can reach any organ system through the bloodstream and establish severe local infections of the joints and bones, heart and lungs, or the meninges. Patients with such diseases as cancer or sickle-cell anemia are particularly prone to develop serious and prolonged infection with *Salmonella.* Typhoid bacteria can persist in the bile passages of patients for an indefinite period of time. These carriers can pass the infection to healthy persons if they practice poor hygiene or if they are food handlers.

Shigella. Infection of the gastrointestinal tract by bacteria of the genus *Shigella* is called shigellosis. The illness produces diarrhea consisting of either watery stools or scant stools containing mucus and blood, and cramplike abdominal pain. The widespread occurrence of the syndrome in Japan led to the discovery of the *Shigella* bacteria by Shiga Kiyoshi in 1897.

Shigellosis occurs throughout the world, especially where overcrowding is a problem and personal hygiene is poor. Very few organisms are needed to cause the disease and they survive well on fomites (inanimate objects, such as clothing, that can convey infection). In addition, humans may be carriers.

Diagnosis of shigellosis is made by stool cultures. Treatment with appropriate antibiotics shortens the duration of the illness and eliminates the *Shigella* organism from the stool, thereby limiting further transmission of the disease.

Viral gastroenteritis. Viral gastroenteritis, or viral diarrhea, is perhaps the most common type of diarrhea in the world. Symptoms include vomiting, fever, and diarrhea. Potentially life-threatening dehydration can occur in young infants and children. Rotaviruses, Norwalk agent, adenoviruses, and enteroviruses are the most common causes. There is no specific therapy; treatment is mainly that of fluid and electrolyte replacement.

Hepatitis. Hepatitis is an inflammation of the liver, usually caused by viral invasion. The first description of hepatitis has been attributed to the Hippocratic school in the 2nd century BC. The second reference to the disease was an account by St. Zacharias about 1,000 years later. It has been estimated that about 80 epidemics of jaundice, a common result of hepatitis, occurred throughout the world from the early 1600s to 1874.

Hepatitis is still one of the world's major health problems. The disease is generally caused by one of three or more related viruses, called hepatitis A, hepatitis B, and non-A, non-B hepatitis. Infections with any of these viruses usually results in low-grade fever, malaise, nausea, jaundice (discoloration of the skin, eyes, and mucous membranes because of an overproduction of bile pigments by the liver), and abdominal pain. Many cases of hepatitis are so mild that a person may be unaware of infection by one of the viruses. Hepatitis B, however, has a tendency to persist in the liver and to cause a prolonged infection that can result in severe damage to the liver's cells. Approximately one in 10 patients with hepatitis B infection becomes a carrier of the virus and may transmit it to others. Fulminant hepatitis is a life-threatening infection that results in severe liver failure, impaired kidney function, difficulty in the clotting of blood, and marked changes in neurological function. Such patients rapidly become comatose; mortality is as high as 90 percent.

Symptoms of hepatitis

Other viruses that cause hepatitis include Epstein-Barr (EB) virus; herpes simplex viruses; measles, mumps, and chicken pox viruses; and cytomegaloviruses. In all cases, there is no specific treatment. Vaccines have been developed to prevent hepatitis B infection, and human immune serum globulin (HISG) is available to prevent infection in those who have had contact with hepatitis patients.

Cholecystitis. Cholecystitis is an inflammation of the gallbladder that is usually associated with one or more predisposing conditions. Patients often have had an earlier infection in the gallbladder that led to bile retention and, in turn, to acute cholecystitis. Cholecystitis without the formation of gallstones, called acalculous cholecystitis, is caused by bacteria such as *Salmonella, Staphylococcus, Streptococcus* (as part of scarlet fever), and leptospirosis. Calculous cholecystitis occurs in people predisposed to form gallstones, such as those who are obese or who suffer from hemolysis (destruction of the red blood cells).

Symptoms are usually those of severe abdominal pain together with vomiting and nausea. Jaundice, dark urine, and fever may also be present. Diagnosis is established by physical examination and certain laboratory and X-ray findings. Treatment involves the surgical removal of the gallbladder. Complications can include perforation of the gallbladder, which can cause overwhelming bloodstream infection requiring intravenous antibiotics and surgery.

Pancreatitis. Inflammation of the pancreas, or pancreatitis, may follow abdominal trauma or the formation of gallstones that obstruct the common bile duct. It can be associated with excessive ingestion of alcohol; with disorders such as cystic fibrosis or Reye's syndrome; or with

scorpion stings. Infectious causes of pancreatitis include mycoplasmas, Epstein-Barr viruses (the cause of infectious mononucleosis), Coxsackie viruses, leptospirosis, hepatitis viruses, mumps, congenital German measles, *Ascaris* worms, and syphilis.

Symptoms are severe abdominal pain, bruises on the flanks or umbilical area, nausea, and, occasionally, fever. Diagnosis is established by certain laboratory tests and X-ray findings. Treatment is usually supportive and consists of the management of pain and intravenous feeding, allowing the gastrointestinal tract to rest.

INFECTIONS OF THE BONES AND JOINTS

Osteomyelitis. Osteomyelitis is the inflammation of bones, generally caused by bacteria. Microorganisms can be introduced to the bone by direct inoculation, following trauma or surgery; by the bloodstream; or by invasion from a contiguous site of infection. Once the organisms gain entry to the bone they multiply and cause microscopic bone death. The most frequent organism involved is *Staphylococcus aureus.*

Symptoms of osteomyelitis include fever, bone pain, and warmth, tenderness, and swelling of the overlying skin and soft tissues. Diagnosis is confirmed by radionuclide bone scans. The precise cause of the infection is determined by cultures of the blood and bone. In addition to bacteria, the disease can be caused by fungi or viruses.

Treatment of osteomyelitis requires the long-term administration of intravenous antibiotics; some patients also require surgery to remove dead bone tissue. If the disease is not treated appropriately, acute osteomyelitis can progress to a chronic disease with extensive bone damage, possibly requiring amputation of the affected limb.

Arthritis. Infection of the joints by bacteria is called septic arthritis. The common infecting organisms are *Staphylococcus aureus, Haemophilus influenzae* (in young children), *Salmonella, Neisseria gonorrhoeae,* and *Pseudomonas.* Patients suffer fever, swelling, heat, and tenderness over the involved joints. Diagnosis is established by physical examination, X-rays, and culture of the fluid from the affected joint and from the blood. Intravenous antibiotics are often used to treat this condition, although open surgical draining may be required.

SKIN INFECTIONS

Many illnesses caused by infectious agents have associated skin conditions. In some cases an exanthem, or rash, may be the hallmark of the disease, as in the case of measles, smallpox, and chicken pox (varicella). In others the rash is only a vague indicator of the more important underlying disease, as with meningitis, syphilis, or Rocky Mountain spotted fever.

Dermatophytosis. The dermatophytes comprise a highly specialized group of fungi that live on skin and can only rarely subsist in deeper tissues. The growth of these organisms on hair, nails, and skin elicits changes that vary with the location of the infection.

Ringworm

Superficial skin lesions are referred to as ringworm or tinea, both names referring to the round shape of most of the lesions, similar to the holes made in woollen garments by the clothes moth, genus *Tinea.* The tineas are usually subclassified according to the area of the body that is involved. Tinea capitis involves the head; tinea corporis, the body surfaces; tinea axillaris, the armpit (or axilla); tinea cruruis ("jock strap itch"), the genital area; tinea pedis ("athlete's foot"), the feet; and tinea unguium, the nails. Diagnosis of these conditions is made by observation and by microscopic examination. Treatment with topical antifungal agents or with oral antifungal agents may be provided. Prognosis is excellent.

Candidiasis (Moniliasis). *Candida* is a fungus capable of causing skin infections as well as serious systemic disease. It is a ubiquitous organism that covers the skin and is harboured by many people. It can cause skin infections such as diaper rash or thrush (infection of the mucous membrane lining of the mouth in infants). Diagnosis is made by the appearance of the infection. Treatment involves the use of specific antifungal drugs.

Warts. Warts are infectious, benign skin tumours induced by viruses. Although warts have been studied at least since the time of Aulus Cornelius Celsus, the great Roman medical writer, in the 1st century AD, they continue to interest researchers and challenge the therapeutic skills of experienced physicians.

Warts are common, their presence often associated with cosmetic problems, in the case of common and flat warts; physical disability, with plantar warts; and embarrassment, with venereal warts. The giant *Condylomata acuminata* can cause destruction and certain warts can, rarely, mutate and so produce cancer. The papilloma virus causes common warts. Among the numerous ways to remove warts are cryotherapy, electrodessiccation, and the application of podophyllin.

Impetigo. Impetigo, a simple superficial infection, is the most common skin infection among children. Symptoms may include thick, crusted skin lesions caused primarily by streptococcal bacteria and blisters associated with staphylococci. Poor hygiene, crowding, and humid, hot weather have been implicated in the spread of impetigo from one child to another.

Impetigo is generally diagnosed by observation. The most effective treatment is the use of oral or intramuscular penicillin. Topical treatment with special soaps alone generally is not effective except in very mild cases.

Cellulitis. Infections of the skin and subcutaneous tissue are called cellulitis. The condition is characterized by an area of redness, warmth, swelling, and tenderness, usually well differentiated from the surrounding tissues. A wide variety of microbes can cause cellulitis. The infection may follow trauma, insect bites, puncture wounds, or bloodstream infection (when organisms from the blood invade a particular localized soft tissue site). Erysipelas, or St. Anthony's fire, is a specific, debilitating form of cellulitis caused by *Streptococcus pyogenes* bacteria. Antibiotics are used for the treatment of cellulitis.

OCULAR INFECTIONS

Blepharitis. Blepharitis is a common chronic inflammation of the eyelids that is caused by a staphylococcus. The illness is marked by red, scaling, itching eyelids and the loss of the eyelashes. Diagnosis is established by examination and by culture of ocular discharge. Treatment consists of the application of warm compresses and topical antibiotics.

Sty

Hordeolum. Hordeolum, or sty, is a pus-forming infection within the hair follicle or oil gland along the margin of the eyelid. It is usually caused by staphylococci. Frequent application of warm, wet compresses usually provides effective therapy; the infection may resolve or drain spontaneously. Severe pain or failure to resolve may require surgical drainage.

Dacryocystitis and dacryoadenitis. Inflammation of the tear sac, called dacryocystitis, is a common pediatric infection. It usually occurs after a long-term obstruction of the nasolacrimal duct (a tube that connects the eye with the nose, allowing proper drainage of the eye's secretions). This infection, caused by staphylococci or streptococci, produces tender swelling in the inner corner of the eye and an area of surrounding redness associated with a purulent (puslike) discharge from the eye. Warm compresses and antibiotics are used in treatment.

Dacryoadenitis, an inflammation of the tear gland, is uncommon. It can be a manifestation of infectious mononucleosis, mumps, gonorrhea, or influenza. Symptoms include pain and swelling in the region of the gland, located in the outer half of the upper lid. Treatment employs warm compresses and antibiotics.

Preseptal and orbital cellulitis. Inflammation of the tissues around the eye is called preseptal cellulitis. Orbital cellulitis is an inflammatory process involving the layer of tissue that separates the eye itself from the eyelid. These infections occur most commonly in children under three years of age and in older people with diabetes or other debilitating diseases. The offending organism is usually a staphylococcus.

Symptoms of preseptal cellulitis are severe edema (accumulation of fluid) in the eyelid, a characteristic purplish-blue hue around the eye, and high fever; the patient ap-

pears to be critically ill. A person with orbital cellulitis has severe pain upon movement of the eye, redness and swelling of the sclera (outer membrane of the eyeball), and fever. Diagnosis is established by physical examination, and the specific bacterial cause of the infection is determined by cultures of the ocular area and of the blood.

Preseptal and orbital cellulitis are serious and life-threatening infections. In general, patients should be hospitalized and should receive intravenous antibiotics and supportive care. Complications include extension of the infection into the brain and occlusion of the central retinal vein, with permanent loss of vision.

Conjunctivitis (pink eye). Inflammation of the conjunctiva (mucous membrane lining the inner surface of the eyelid) is a common disorder. It can be caused by many microorganisms, including staphylococci, *Haemophilus influenzae,* streptococci, gonococci, and viruses such as adenoviruses. The symptoms are weeping red eyes. The patient complains that a foreign body is present within the eye; occasionally there is blurring of vision because of the presence of pus.

One of the most severe forms of conjunctivitis is that caused by *Neisseria gonorrhoeae.* In this form there is marked swelling of the eyelid and aching pain in the eye. It can progress rapidly, producing erosion, perforation of the cornea, and subsequent loss of the eye. Systemic antibiotics can be administered in the case of congenital gonorrheal conjunctivitis. Most cases of conjunctivitis are self-limited, but recovery is more rapid in cases of bacterial conjunctivitis treated with topical antibiotics.

Infections with specific organisms

Literally thousands of organisms can produce clinical illness. Some of these organisms have been discussed above in relation to the organs that they infect. In this section, others are discussed further because of their significance as a cause of generalized infection in individual persons or their importance to public health. (R.D.F./Re.G.)

BACTERIA

Anthrax. The disease anthrax, named from the Greek word for coal, is so called because, when it attacks a person's skin, a sore with a coal-black centre develops. In more than 90 percent of the cases of anthrax in human beings, the bacilli remain within the skin sore. In cutaneous (skin) anthrax, there is always considerable swelling around the lesion and bouts of shivering and chills, but there is little other disability. The organism that is responsible for causing anthrax, *Bacillus anthracis,* may escape from the sore, however, and spread up a lymph channel to the nearest lymph node, where it is usually halted. Only seldom does it invade the bloodstream, causing rapidly fatal septicemia (blood poisoning), internal bleeding, and, sometimes, anthrax meningitis.

When the anthrax bacillus—which comes from infected animal products, usually hides, wool, and hair or bones—is breathed in and attacks the lungs, a suffocating bronchitis results from which few persons survive. This illness is sometimes called "woolsorters' disease." Intestinal anthrax is a very rare and fatal form of the disease that comes from eating the flesh of animals that have died of anthrax. The disorder causes inflammation of the stomach and intestines with ulcers much like the sores that appear in cutaneous anthrax.

Intestinal anthrax

Bacillus anthracis is a spore-bearing germ capable of remaining alive in pastures for years. In damp, warm pastures the bacilli multiply rapidly, but when the pastures dry out, the bacilli form spores. The disease is often fatal to cattle, sheep, goats, and other herbivores, and they sometimes die in the thousands. Their hides, wool, and bones are sometimes heavily contaminated with *B. anthracis.* The bacillus is not highly infectious in human beings and does not depend upon multiplication within them for survival.

Cutaneous anthrax is treated with penicillin and the outlook is good if it is treated early. It can be prevented by the vaccination of exposed workers and by the sterilization of contaminated materials, such as imported wool.

Tularemia. Tularemia is a disease found in a large variety of animals. It was named for Tulare County, California, where *Pasteurella tularensis,* the germ that causes the infection, was first isolated in ground squirrels. In animals, *P. tularensis* usually causes an illness characterized by a caseous (cheeselike) degeneration of various glands and the spleen, liver, and lungs.

Tularemia can be spread to human beings by the bite of an infected animal; by contact with the blood or fine dust from the animal's body during skinning or similar operations; by the bite of a tick, mosquito, or deerfly; or by the ingestion of infected animal products that have not been properly cooked. The most common form of the disease in human beings is the ulceroglandular form, in which there is a painful sore at the site of the infection and a swelling of the lymph node that drains the area; the sore is often on the finger and the swelling, or bubo, in the armpit. The bubo can break down and discharge pus, but it sometimes remains hard and tender for weeks. Along with these local signs, the infected person has a fever that may persist for two or three weeks, with headache, vomiting, body pains, and general weakness. Tularemia also occurs in a typhoidal form marked by an exhausting, or feverish, illness and a pneumonic form caused by inhalation of dust contaminated by *Pasteurella tularensis.* Although penicillin is of no use, treatment with other antibiotics is usually successful. Mortality is sometimes as high as 5 to 7 percent in the typhoidal or pneumonic forms, but for the other type the prognosis is excellent.

Brucellosis. Brucellosis, also called Malta fever or undulant fever, is an infection caused by several species of *Brucella* organisms that have their natural reservoirs in domestic animals—*B. abortus* in cattle, *B. suis* in pigs, and *B. melitensis* in sheep and goats. The infection may not be apparent in animals, for the brucellae and animals that they infect have become fairly well adapted to one another. In cattle, for example, the only signs of illness may be a drop in milk production or a general malaise, although abortion is also common. Brucellosis is, therefore, of considerable economic importance, although it causes no dramatic loss of livestock.

Human beings are not natural hosts for the *Brucella* germs and they often react violently when infected. Acute brucellosis is an illness characterized by sudden chills and fevers, drenching sweats, pains throughout the body, and great weakness and fatigue. It lasts for two weeks and then may abate, but the symptoms often return with waves of fever (from which the name undulant fever was derived) in recurring bouts of illness for about six months or a year. The infection then ceases in most people, although it can persist, sometimes for years. Chronic brucellosis is perhaps the most difficult form of the disease to deal with because the patient's symptoms are vague and may easily be misdiagnosed as psychological in origin. Brucellosis may be complicated by infection of the joints or spine or involvement of the heart, eyes, kidneys, or lungs.

Brucellosis is treated with antibiotics, particularly the tetracyclines and streptomycin. Therapy in the acute form is fairly satisfactory, but the chronic form is more difficult to treat. The solution to the problem of brucellosis lies in the eradication of the disease in infected animals.

Botulism. After 13 people in Wildbad, Germany, in 1793, shared an uncooked smoked sausage that had been sitting for hours, all became ill and six died. (To describe their illness, the word botulism was derived from the Latin *botulus,* for sausage.) The conditions beneath the skin of the eaten sausage had been anaerobic (*i.e.,* there was very little oxygen in the meat) and enough time had elapsed to allow the clostridial organisms present to multiply and produce a toxin in the sausage substance.

Botulism is a disease caused by *Clostridium botulinum* that has given off a toxin, which in turn produces the symptoms of the disorder. This spore-bearing organism, present in the soil, grows only in conditions of low oxygen tension, such as may be found in canned, smoked, or cured food, but not at all in fresh food. When multiplying, it produces a very powerful toxin; a mere taste of food containing botulinum toxin is enough to cause death. The toxin is easily destroyed by heat, however; two

Botulinum toxin

minutes at 158° F (70° C) destroys it. The multiplying bacilli are also susceptible to heat; 10 minutes at 176° F (80° C) kills them. The spores are highly resistant to heat and withstand boiling for hours. To kill them, moist heat at 248° F (120° C) is required. This temperature can be achieved with certainty only in commercial canning plants. In home canning, this safe temperature may be reached if a pressure cooker is used, but merely boiling the cans is not reliable.

When *Clostridium botulinum* toxin is swallowed in food, it does not irritate the gastrointestinal tract but is absorbed rapidly and carried in the bloodstream to nerve endings in muscles. In normal response to a nerve impulse, the chemical acetylcholine is produced at these sites and causes the muscle to contract. *C. botulinum* toxin attacks the last fine nerve fibrils and stops the impulse from passing along these fibres. No acetylcholine is released and the muscle cannot contract; it is paralyzed.

The first sign of the disease in many people is paralysis of the muscles of the eyelid, a sign that may appear within hours of eating the food. Next, the paralysis affects the muscles of speech. Swallowing becomes difficult and, finally, the respiratory muscles stop. The person remains conscious through most of the illness, until suffocation occurs. Death may come within a day, although people less severely poisoned may live for a week. Few who reach the stage of severe paralysis survive. Others who may have taken less toxic food develop only the milder forms of paralysis. Mortality in outbreaks varies from 2 to almost 100 percent. In every outbreak some patients die, and in the worst there are no survivors.

Treatment of botulism is difficult. *Clostridium botulinum* antitoxin is given in large doses intravenously, but it is doubtful that antitoxin can do anything to dislodge the toxin once it has reached the nerve fibrils. A chemical, guanidine hydrochloride, counteracts the action of *C. botulinum* toxin on nerve endings and has been used successfully in treatment, but it is itself a toxic substance that should be given only with great care. Paralyzed muscle can recover if the patient can be kept alive, and perhaps the best hope of survival in otherwise desperate cases lies in tube feeding, a tracheostomy (making an opening in the windpipe), and use of an artificial respirator. *C. botulinum* is an inhabitant of the soil, not a parasite of human beings. It attacks people only when they disturb its natural environment.

Tetanus (lockjaw). Tetanus is caused by *Clostridium tetani,* a spore-forming organism similar to *C. botulinum.* The tetanus organism lives in the soil and is ubiquitous. It gains access to the human body through wounds, particularly deep punctures. Its toxin is given off under anaerobic conditions in which the organism is deep within tissues and deprived of oxygen.

As the germs multiply, they produce their toxin, which rapidly spreads by way of the bloodstream or directly up a nerve to the central nervous system. There it attacks motor nerve cells and excites them to overactivity. Excessive impulses rush down the nerves, and the muscles at the far end are thrown into severe convulsive spasm. The most common spasms occur in the muscle of the jaw, and the first sign of the illness often is stiffness of the jaw, or trismus. The muscles of the mouth are often affected, pulling the lips out and up over the teeth into a grimace, the mixture of smile and snarl that heralds the onset of the generalized convulsive stage of tetanus.

Any muscles can be attacked. Spasm of the muscles of the throat can make swallowing impossible, whereas the muscles of the larynx or of the chest wall can be thrown into such violent spasm that breathing is impossible and life is threatened. This is a common cause of death if the tetanus is untreated, but there are other toxic effects on the heart, blood pressure, and vital brain centres that may kill later in the disease.

Tetanus is a preventable disease. Tetanus vaccine is one of the most successful, least toxic, and most enduring of available vaccines. It should be given to every child and repeated throughout life, especially after wounds have been received. Tetanus antitoxin (ATS) is a different product that contains antibodies against tetanus derived from the blood of people who have been immunized against the disease. It can be injected into the body of a wounded person to provide some immediate protection. Tetanus antitoxin can also be useful if a person has not been appropriately immunized.

Treatment of tetanus is primarily supportive. Tetanus antitoxin is given to help neutralize the toxin in the bloodstream, but it has little effect once the toxin has affected the nerve endings. Intravenous penicillin kills the organisms that remain within the wound site. Another important component of treatment consists of keeping the patient alive until the disease subsides spontaneously with time. Patients are usually intentionally paralyzed with drugs to prevent muscle spasms caused by the disease; they are maintained by artificial or mechanical breathing because the respiratory muscles are paralyzed. Even the most severely ill patients can be saved if managed by skilled nurses and physicians in an intensive care setting. Sedatives and painkillers are used to keep the patient comfortable. Prevention of this devastating disease is by immunization.

Leprosy (Hansen's disease). Leprosy is one of the most feared of diseases, a dread that stretches back into antiquity, the leper being considered as "unclean." Yet it is not a highly infectious disease, prolonged intimate family contact being needed for its spread from one person to another. Most adults in areas in which leprosy occurs appear to be immune, but children are very susceptible. The disease has almost disappeared from most temperate countries, but it is still common in Asia, Africa, and Central and South America. At least 2,000,000 people are known to have the disease, and the actual number of infected people may be as high as 15,000,000. In sheer numbers, leprosy presents a serious problem, not lessened by the fact that it is a disabling, deforming disease, slowly progressing throughout the life of the leper but not usually cutting that life short. Management of leprosy involves social, vocational, medical, rehabilitative, orthopedic, and reconstructive surgical services.

The leprosy bacillus

The disease is caused by the leprosy (or Hansen) bacillus, *Mycobacterium leprae,* and has two principal forms, the tuberculoid and the lepromatous. How the bacillus gets into the human body is not clearly known. It can be discharged in enormous quantities from the nose or broken-down sores of an infected person and, therefore, can be inhaled or spread from skin to skin.

The human body's first reaction to the leprosy bacillus takes place in the deep layers of the skin in one of two ways. In one form of reaction, body cells crowd into the area in an attempt to seal off the invader, and in these areas very few bacilli can be found. This form of leprosy is known as tuberculoid leprosy because of the hard nodules, or tubercles, that form in the skin. The intense cellular reaction involves all of the thicknesses of the skin and the tissues under it, the sweat glands, the hair follicles, and the nerve fibrils that end in the skin. All of this shows up on the infected person's skin as a firm dry spot in which there is no sense of heat, cold, or touch. The cellular reaction continues to spread into the main trunk of the involved nerve, tending to strangle it so that impulses cannot get up or down and thereby causing loss of power in the muscles of the area, loss of sense of pain, and loss of circulation in the affected part. This is most commonly seen in the forearm or lower leg, and it leads to claw hand and gross deformity of the foot, but paralysis of muscles of the face, eye, and neck may also occur. The patient is unable to feel pain and minor injuries pass unnoticed. Large eroding ulcers can form, causing loss of fingers and toes; sometimes the condition of the limb is so bad that amputation is necessary. It is ironic that this form of leprosy occurs in people whose tissues resist the disease, for the intense cellular response is a reaction of resistance, successful insofar as it prevents local multiplication of the leprosy bacillus and spread through the body but unsuccessful in that it grips and destroys the vital tissues in the invaded areas.

Lepromatous leprosy

In the second form of the disease, called lepromatous, or cutaneous, leprosy, there is very little cellular response and the bacilli can multiply freely. The organisms can always be found in enormous numbers in the deep layers of the

affected skin, and they spread widely through the skin's lymphatic channels. The disease spreads up the nerves but does not adhere to them as in the tuberculoid form. It very often spreads to the skin of the face, where it causes thickening and corrugation of the skin and a typical leonine appearance. Soft nodules appear on the ears, nose, and cheeks and sometimes break down into discharging sores. The nose often is teeming with bacilli, and this sometimes leads to destruction of the septum of the nose and the palate.

The progress of leprosy is slow. It may be years before a child infected by a parent shows the first sign of the disease, often a vague, scarcely noticed spot on the skin. Years may pass before any change is noticed, and the child has often grown to an adult before the disease is recognized. Lepers suffer occasionally from bouts of fever, but the course of the disease is mainly one of increasing disability and disfiguration. Lepers often do not die of leprosy; they can live a normal span of years and, with proper medical and rehabilitative care, can live in some measure of comfort.

Drugs such as the sulfones and rifampin can arrest the progression of leprosy and render the patient noncontagious. Resistance to drug therapy has been noted, however, and the increase in frequency of leprosy may be a result of the failure to respond to the commonly used forms of treatment.

One of the difficulties in managing leprosy is that no one has been able to grow the leprosy bacillus in laboratory plates. Only since 1960 has it been possible to grow the bacillus in a laboratory animal. The footpad of a mouse can be infected, and the disease in mice resembles that in human beings, allowing for the study of its progress and of the effect of new drugs on it in the laboratory.

Apart from the use of drugs, the management of the disease is a vast human problem. The leper must be helped in his disfigurement and his paralysis. The greatest problem is the prevention of infection. A baby born to a leprous mother has little chance of escape unless it is separated from her. A father is almost bound to infect some members of his family unless taken away from them. The fear of separation makes the family conceal the disease and thus increases the danger of its spread. The ideal must be not a colony for lepers but village or community groups in which whole families can live in good conditions and the leper can be given necessary treatment and the encouragement and help to work within the disease's limitations.

Tuberculosis. Tuberculosis is one of the great scourges of mankind, yet most people infected with the germ do not suffer from the disease. Instead, the germ, *Mycobacterium tuberculosis,* becomes trapped in the tissues of the body, surrounded by defensive cells, and finally sealed up in a hard, calcified nodule, or tubercule. This nodule can be seen on an X-ray film of the lungs as a tiny, dense shadow at the edge of the lung field, and often a shadow representing lymph nodes can be seen in the centre of the field. These two shadows represent the primary complex of tuberculosis and show that the tubercle bacillus, on reaching the lungs, travelled a short distance before the defenses of the body brought its movement to an end.

Childhood tuberculosis

Infection with the tubercle bacillus is common in childhood. An infected child often suffers no obvious illness but, despite the mildness of the infection, develops a lifelong immunity to the disease. A skin test taken at any later time will reveal the earlier infection and the immunity. Since the tubercle bacillus may be sequestered quickly in the tissues, the child is not usually infectious. In some cases, however, the tubercle bacillus is not caught in a primary complex but spreads rapidly in the lungs and out into the bloodstream. It can then reach almost any organ of the body, especially the meninges that cover the brain, where it causes tuberculosis meningitis. This rapidly developing disease is called miliary tuberculosis because, when viewed on an X-ray, the many tubercles look like tiny millet seeds. The infection may have resulted from the inhalation of a very large amount of tubercle bacilli or from a lowered body resistance that could be caused by several factors, one of which is malnutrition.

Tuberculosis takes a different form in adults. Whether this is because of a reactivation of an earlier, primary infection or a new heavy infection is unclear. In the adult disease, the bacillus spreads slowly and widely in the lungs and, instead of causing a hard, calcified nodule, forms large cheesy (caseous) masses that break down the respiratory tissues and form cavities in the lungs. The disease process eats its way into the bronchi, or air passages, and the tubercle bacilli can then be coughed or breathed out, making the patient infectious. A blood vessel can be eroded by the advancing disease and then the patient, with great alarm, coughs up bright red blood. This spreading type of the disease was once popularly known as consumption, a name that vividly described the destructive progress of the illness. Before curative drugs were available, this process was usually, but not invariably, fatal. With rest, good food, and nursing care, patients built up resistance against the tubercle bacillus and eventually sealed it off, though with much scarring of the lungs. In those persons whose resistance was overcome, the lungs became a mass of caseous material, and the bacilli often spread to the kidneys, meninges, and other organs before the patient died.

Bovine tuberculosis

Most of the discussion of tuberculosis so far relates to the disease caused by the strain of *Mycobacterium tuberculosis* that infects human beings. The strain that infects cattle is excreted in milk. If the milk is ingested raw, the bovine bacillus readily infects human beings. The bovine bacillus may be caught in the tonsils and may spread from there to the lymph nodes of the neck, where it causes caseation of the node tissue. The node swells and bulges under the skin of the neck, finally breaking down and bursting through the skin as a chronic discharging ulcer. From the gastrointestinal tract, the bovine bacillus may spread into the bloodstream and reach any part of the body. It shows, however, a great preference for bones and joints, where it causes destruction of tissue and eventually gross deformity. Tuberculosis of the spine, or Pott disease, kept children in hospitals, strapped for years to spinal frames, the purpose of which was to prevent deformity while the body slowly overcame the infective process. Despite such measures, many cases of hunchback were caused by the bovine tubercle bacillus. Pasteurization of milk kills tubercle bacilli and this, along with the raising of tuberculosis-free herds, has led to the disappearance of bovine tuberculosis in human beings in many countries.

Tuberculosis remains a disease that is encountered frequently, however. In fact, the prevalence of the disease increased during the 1970s. Tuberculosis is more severe in malnourished people, but it can afflict anyone, regardless of nutrition, hygiene, or socioeconomic status. In many parts of the world, especially the less developed countries, it remains one of the major fatal diseases.

Treatment of tuberculosis represents one of the great advances in 20th-century medicine. Streptomycin was the first drug that proved to be effective, followed by isonicotinic acid and para-aminosalicylic acid. One of the difficulties of treatment is that some tubercle bacilli become resistant to drugs, though there is less danger of this when two or three drugs are given together. Despite such combinations, some tubercle bacilli may still become resistant and newer, but sometimes more toxic, drugs must be used. The drug treatment of tuberculosis has, in the developed countries, removed the need for the sanitariums in which patients were nursed and rested for years while the defensive properties of their bodies dealt with the disease. Drugs have reduced the span of treatment to months instead of years, and patients usually can be treated in their homes.

Diphtheria. Diphtheria is caused by *Corynebacterium diphtheriae,* which enters the body of a susceptible person through the respiratory tract, though the symptoms it causes are not usually respiratory. The germ settles on the tonsils and creates an inflammatory reaction that leads to the formation of a thick membrane that may cover the tonsils and spread to the hard palate and up behind the nose. The diphtheria bacillus remains at this site and multiplies there, all the while producing its powerful toxin, which spreads through the body in the bloodstream and causes great damage to the heart and the nervous system.

The membrane that formed during the infecting process

separates from the throat by the end of a week or 10 days, but toxic complications occur later in severe cases. The heart suffers first, often in the second or third week. The patient develops toxic myocarditis (inflammation of the heart muscle), which can be fatal. If the person survives this dangerous period, the heart recovers completely and the patient appears to be well. But this appearance is deceptive and is one of the most treacherous aspects of the disease because paralysis caused by the action of diphtheria toxin on the nervous system often strikes when the patient seems to have recovered. Paralysis of the palate and some eye muscles develops in about the third week; this is usually transient and not severe. As late as the fifth to eighth week, however, paralysis affecting swallowing and breathing develops in severe cases, and the patient may die after weeks of apparent well-being. Later still, paralysis of the limbs may occur, but this is not life-threatening. If the patient can be supported through this critical phase, recovery is complete.

Diphtheria antitoxin

Antibiotics can destroy the diphtheria bacillus in the throat and are given to every patient. The essential treatment of the patient, however, is the administration of antitoxin to neutralize the circulating toxin. Antitoxin does not neutralize toxin that has already been bound to tissue and that has produced tissue damage. The antitoxin may be lifesaving if given early enough, but the body eventually eliminates it as a foreign substance and it does not give any permanent protection against the disease. To prevent diphtheria, the body must produce its own antitoxin in response to active immunization with diphtheria toxin. Active immunization carried out on a wide scale has markedly decreased the prevalence of this illness.

(A.B.C./R.D.F./Re.G.)

Borrelia (relapsing fever). This disease was first noted in 1868, when organisms were observed in the blood of patients during an attack of fever. Relapsing fever is characterized by recurring episodes of fever separated by periods of relative well-being. It is caused by spirochetes of the genus *Borrelia.* Lice or ticks become infected by feeding on people infected with this organism, and they remain carriers throughout their lives. Relapsing fever spreads under conditions of crowding, cold weather, and poor hygiene, all of which favour the spread of lice. Epidemics have occurred during wars, earthquakes, famines, and floods.

After the spirochete has lived about one week in its newly infected host, the person experiences the sudden onset of high fever, chills, headache, and muscle aches. After approximately one more week, the infected person returns to a normal state of health for seven days or more. As many as 10 repetitive feverish episodes can occur if the disease remains undiagnosed and treatment is not provided. The crisis that terminates each attack is typically characterized by profuse sweating, low blood pressure, low temperature, and malaise. Relapses are usually less severe than the first attack and the intervening periods of good health are more extended.

An enlarged liver or spleen, rashes, and inflammation of the eye, heart, and brain may be noted in patients with borderline infection. The diagnosis can be by specific serologic (blood) tests or by seeing the loosely coiled organism in the patient's blood. Many antibiotics are available for treatment. Prevention by the elimination of the vectors that transmit the disease is of the utmost importance.

Whooping cough (pertussis). Whooping cough is an acute infection of the upper respiratory tract that is most serious in young children. It is caused by the bacterium *Bordetella pertussis;* a similar illness is caused by *B. parapertussis.* Both are highly communicable diseases and, once either is introduced into a susceptible population, virtually everybody becomes infected.

Once the organism is ingested, it remains quiet for about one week. The illness then progresses through three stages—the catarrhal, paroxysmal, and convalescent—which together last six to eight weeks. In the first stage, which lasts about two weeks, the symptoms are those of a cold, dry cough, red eyes, and low-grade fever. In the paroxysmal state, there is a repetitive series of coughs that are exhausting and often result in vomiting. The infected person may appear blue, with bulging eyes, and be dazed and apathetic, but the periods between coughing paroxysms are comfortable. During the convalescent stage there is gradual recovery. The complications of whooping cough include pneumonia and ear infections.

The diagnosis of the disease is usually made on the basis of its symptoms and is confirmed by specific cultures. Antibiotics are available that may help to shorten the duration of illness and the period of communicability. Prevention is by immunization. (R.D.F./Re.G.)

Pneumococcal, staphylococcal, and streptococcal infections. The term coccus denotes a small, round dot or berry. "Pneumo" is from the Greek word for lung, "staphylo" from the Greek word for a bunch of grapes, and "strepto" from the Greek word for twisted, like a chain. Thus, the pneumococcus is a dot-shaped germ that attacks the lungs and other organs, the staphylococci appear under the microscope like a bunch of grapes, and the streptococci form chains. Each of the three microorganisms belongs to a different biological group, but in their relations with human beings they have many things in common. They often live as harmless commensals in the upper respiratory tract, but each can become virulent and cause serious conditions such as pneumonia, septicemia (blood poisoning), or meningitis.

The pneumococcus is the most common cause of lobar pneumonia, the disease in which one or more lobes, or segments, of the lung become solid and airless as a result of inflammation. Staphylococci affect the lungs either in the course of staphylococcal septicemia—when the germs leak out of the circulating blood and cause scattered abscesses in the lungs—or as a complication of a viral infection, commonly influenza—when these organisms invade the damaged lung cells and cause a life-threatening form of pneumonia. Streptococcal pneumonia is the least common of the three and occurs usually as a complication of influenza or other lung disease.

Pneumococci often leak out from inflamed lungs into the bloodstream and cause septicemia, with continued fever but no other special symptoms. Staphylococci produce a hectic type of septicemia with high, swinging fever; the germs reach almost any organ of the body, the brain, the bones, and especially the lungs, and destructive abscesses form in the infected areas. Streptococci also cause septicemia with hectic fever, but the organisms tend to cause inflammation of surface lining cells rather than abscesses, pleurisy (inflammation of the chest lining) rather than lung abscess, peritonitis (inflammation of the membrane lining the abdomen) rather than liver abscess. In the course of either of the last two forms of septicemia, organisms may leak into the nervous system and cause streptococcal or staphylococcal meningitis, but these are rare conditions. Pneumococci, on the other hand, often spread directly into the central nervous system, causing one of the common forms of meningitis.

Skin diseases

Staphylococci and streptococci are common causes of skin diseases. Boils and impetigo (in which the skin is covered with blisters, pustules, and yellow crusts) may be caused by either. The staphylococcus also can cause a vicious inflammation of the skin that strips the outer skin layers off the body and leaves the underlayers exposed, as in severe burns, a condition known as epidermal necrolysis. Streptococci can be the cause of the red inflammation of the skin known as erysipelas.

Some staphylococci produce an intestinal toxin and cause food poisoning. Streptococci settling in the throat produce a reddening toxin that speeds through the bloodstream and produces the symptoms of scarlet fever.

All three organisms can be killed with antibiotics. The diseases they cause have not been conquered by drugs, but, when antibiotics are used with care, much of the damage done by the organisms can be controlled and many lives can be saved.

Meningococcal infections. Meningococci are fairly common inhabitants of the human throat, in most cases causing no illness at all. As the number of healthy carriers increases in any population, however, there is a tendency for the meningococcus to become more invasive. When an opportunity is presented, it launches into the bloodstream, invades the central nervous system, and causes meningo-

coccal meningitis (formerly called cerebrospinal meningitis or spotted fever). Meningococcal meningitis, at one time a dreaded and still a very unpleasant disease, usually responds to treatment with penicillin if diagnosed early enough. When meningococci invade the bloodstream, some leak into the skin and cause blood-stained spots, or purpura. If the condition is diagnosed early enough, antibiotics can clear the bloodstream of the germ and prevent any from getting far enough to cause meningitis. Sometimes the septicemia takes a mild, chronic, relapsing form with no tendency toward meningitis; this is curable once it is diagnosed. The meningococcus can cause one of the most fulminating of all forms of septicemia, in which the body is rapidly covered with a purple rash, purpura fulminans; in this form the blood pressure collapses, the heart and blood vessels are affected by shock, and the infected person dies within a matter of hours. Few are saved, despite treatment with appropriate drugs.

Haemophilus influenzae. *Haemophilus influenzae* is a microorganism named for its occurrence in the sputum of patients suffering from influenza—an occurrence so common that it was at one time thought to be the cause of the disease. It is now known to be a common inhabitant of the nose and throat that may invade the bloodstream, producing meningitis, pneumonia, and various other diseases. In children it is the most common cause of acute epiglottitis, an infection in which tissue at the back of the tongue becomes rapidly swollen and obstructs the airway, creating a potentially fatal condition. *H. influenzae* also is the most common cause of meningitis and pneumonia in children under five years of age, and it is known to cause bronchitis in adults. The diagnosis is established by cultures of blood, cerebrospinal fluid, or other tissue from sites of infection. Antibiotic therapy is generally effective, although death from sepsis or meningitis is still common. (A.B.C./R.D.F./Re.G.)

Plague. Plague was noted in China in the 10th century, but its reputation as the "Black Death" can be attributed to epidemics that swept through the European continent during the 1300s. In 1894 Alexandre Yersin was one of the first to describe the plague bacillus, *Pastuerella pestis* (also called *Yersinia pestis* or *Bacillus pestis*), and rats and fleas were later proven to be responsible for the spread of the disease.

Plague is transmitted to human beings by bites from infected fleas or rats or by close contact with an infected person. Once the organism enters the bloodstream, it travels to the liver, spleen, kidneys, lungs, and brain. It can establish residence at any or all of these places, causing overwhelming infection after a three- or four-day incubation period. Fever, malaise, weakness, headache, and shaking chills quickly develop. The appearance of one or more tender, large nodes (bubos) has provided the clinical name "bubonic plague" for one form of the disease. Pneumonic plague is characterized by a devastating pneumonitis that runs a fulminant course, with neurologic manifestations such as insomnia, stupor, weakness, staggering gait, speech disorder, and loss of memory. Antibiotic therapy must be given promptly to prevent death. Hygienic measures and elimination of rats from human habitations have virtually ended urban epidemics of plague.

Two forms of plague

Rat bite fever. It is possible to consider three distinct infections as forms of rat bite fever. The first, sodoku (Japanese: "rat poisoning"), or classic rat bite fever, is caused by *Spirillum minus.* The symptoms of this infection include relapsing high fevers associated with shaking chills, severe muscle aches, reddish-brown rash, and an ulcer at the site of the bite. Although the major complication is infection of the heart, death is rare. Penicillin is the preferred antibiotic treatment.

Streptobacillus moniliformis is the cause of streptobacillary rat bite fever and of Haverhill fever (the term used to describe streptobacillary fever when it is transmitted by means other than the bite of a rodent). In streptobacillary rat bite fever, the initial symptoms are fever, vomiting, severe headache, rash, and inflammation of the joints. These may be followed by pneumonia, bronchitis, arthritis, heart infections, and abscesses in the brain. Penicillin is effective treatment; without it, mortality is high.

Leptospirosis. "Mud fever," "pea pickers' disease," and "European swamp fever" were all terms formerly used to describe this illness. Most mammals, particularly rodents, can be infected and may transmit the disease to human beings.

Leptospirosis is an acute systemic illness characterized by extensive inflammation of the blood vessels. The first symptoms to arise are the abrupt onset of fever, chills, muscle aches, headache, abdominal pain, vomiting, and red eyes. After a latent period of five to seven days, during which the infected person may improve, fever returns and the infection may involve the brain. In a less common form of the illness known as Weil's syndrome or icteric (associated with jaundice) leptospirosis, the infected person turns yellow as a result of the destruction of red blood cells and liver disease. Kidney failure and heart dysfunction occur.

Diagnosis is established by blood cultures on special media. Antibiotic therapy is effective only if given within four days of the onset of the illness. Supportive care and the maintenance of fluid balance are needed because dehydration, cardiovascular collapse, and acute kidney failure can all occur. Leptospirosis may be prevented by good hygiene and the immunization of animals against certain species of spirochetes of the genus *Leptospira.* (R.D.F./Re.G.)

VIRUSES

The infectious viruses that enter the body through the respiratory tract include those that primarily affect the respiratory tract—such as the cold and influenza viruses—and others that are systemic in their effects—such as the infectious mononucleosis, smallpox, and chicken pox viruses. Of the many viruses that can cause disease in the respiratory tract, at least 200 have been identified. Approximately 100 of them are rhinoviruses, the organisms that cause the common cold, by far the most common infectious disease to afflict mankind.

The common cold. Once a virus becomes established on the respiratory surface of the nose, its activities irritate the nose's cells, which respond by pouring out streams of clear fluid. This fluid acts to dilute the virus and clear it from the nose. The sensory organs in the nose are stung by the inflammatory reaction, thereby setting up sneezing, a second method of expelling the virus. If the virus penetrates more deeply into the upper respiratory tract, coughing is added to the infected person's symptoms in a further effort to get rid of the virus. Symptoms abate as the host's defenses increase, the clear fluid often changing to a thick fluid that is full of the debris of dead cells.

The common cold is the expression of an open fight between a rhinovirus and the host's defenses. It takes place in a restricted field, the nasal epithelium (covering layer), but it has, in miniature, all of the characteristics of the wider encounters between human beings and viruses that mark some of the more serious infectious diseases.

Influenza. There are three families of influenza viruses, called the A, B, and C viruses. The A viruses cause the great influenza epidemics, and the B viruses cause smaller localized outbreaks; the C viruses are not important causes of disease in human beings.

The A viruses are not a stable family group; in fact, their most constant characteristic is change. The first virus in the family was isolated and identified in 1933 and was termed simply influenza virus A or, sometimes, A0. The virus kept changing but was still recognizable as virus A0 until 1946, by which time its character had changed so much that its designation was changed to A1. This A1 virus recurred for 11 years, continuously changing but still obviously an A1 virus. In 1957 another major change took place, and the A2 virus appeared; it was first detected in China and so was called the A2 Asian influenza virus. The A2 Asian strain continued to appear, changing slightly year by year, until 1968, when a different strain was isolated in Hong Kong. Although it differed considerably from the 1957 strain, it still had most of the A2 characteristics. It was therefore called not A3 but A2 Hong Kong 1968.

The A viruses

These changes, or mutations, occur in the antigenic structure of the virus. An antigen is that part of a microorganism that gives it its own specific character and that

enables it to be identified from a crowd of otherwise similar microorganisms. The antigens of a germ, also closely connected with its infectivity, stimulate specific antibodies in the body of the infected host. These antibodies protect the host for some time against reinfection, and the microorganism finds it difficult to spread rapidly. When, however, a microorganism undergoes a major change in its antigenic composition, the antibodies provoked by the older antigens give no protection against the new, and the population is wide open to attack by what is, in the antigenic sense, a completely new microorganism. This has two important consequences. The new virus can sweep unhindered through unprotected hosts, and vaccines that gave protection against the old infection are useless against the new one. This happened in 1957 and 1968, when the new Asian and Hong Kong influenza virus strains swept across the globe, causing pandemics of the disease.

The symptoms of influenza are always primarily respiratory in character. Some affected people have little more than a severe common cold, while others suffer an overwhelming general infection. The reaction depends upon where the virus strikes and how much resistance it meets. Reactions of the respiratory lining cells are the same as in the common cold, but location can make a major difference. Whereas fluid poured into the nose is an inconvenience to the host, fluid poured into the air cells of the lungs can be disastrous. In fatal cases there is no inflammatory reaction, as in pneumonia; instead, the lungs fill with fluid and death occurs before such a reaction can take place. Influenza is a rapidly disabling illness, but the worst symptoms are usually over in a week unless complications set in. These are usually not caused by the influenza virus itself but by such microorganisms as pneumococci, staphylococci, and *Haemophilus influenzae.*

Other respiratory illnesses. The common cold is easy to diagnose and so, in an epidemic, is influenza. Many other acute respiratory illnesses, however, are difficult to distinguish clinically. The symptoms are caused by irritation of the cells of the respiratory tract, and these respond in the same way to different irritants. Most affected people have a feeling of chilliness, their temperatures are elevated for a few days, and they may require rest in bed. An exact diagnosis can be made if specimens of sputum, nasal discharge, throat washings, and blood are taken, but this entails a vast amount of tedious and expensive laboratory work, and in many cases the sick person recovers long before the results are known. Mortality is rare except in the severely disabled patient.

Infectious mononucleosis. Infectious mononucleosis incapacitates its host for varying periods of time. Some affected people are physically fit for normal activities within two or three weeks, while others remain ill for as long as two months. The most common symptom of the illness is sore throat, and there is usually a thick white coating, or membrane, on each tonsil. Swelling of the lymph nodes in the neck, armpits, and groin—for which the disease is sometimes called glandular fever—occurs in some people. Changes in the white cells and serum of the blood are detected by a special test.

Epstein-Barr virus

Epstein-Barr virus is the cause of infectious mononucleosis. The virus was first isolated from tumour cells of children suffering from a form of cancer called Burkitt lymphoma. Later research showed that children around the world develop antibodies to this virus early in life, evidence that they have been infected with it, though without suffering any illness and certainly without any signs of tumour growth or of infectious mononucleosis. Infectious mononucleosis is a disease mainly of adolescents and adults and seems to occur only in those who escaped Epstein-Barr virus infection in childhood.

Infectious mononucleosis and Burkitt lymphoma are both diseases of the lymphoid tissues of the body. Some of the unusual blood cells in the former may provide clues to the cause of the latter.

Smallpox and chicken pox. In both smallpox and chicken pox the virus enters by the respiratory route, multiplies in the internal organs, and, at the height of the infection, causes a rash on the skin. The rashes are similar at first glance. In each, the first thing seen is a flat spot, or macule, which changes into a clear blister, or vesicle. When the contents of the blister become turbid, the blister is then called a pustule. The pustule dries up and forms a scab that, upon separation, leaves a scar, or pock. There are two great differences in the diseases, however. In chicken pox, the rash is heaviest on the trunk and upper part of the limbs, displaying what is called a centripetal, or centre-seeking, distribution. In smallpox, the rash is heaviest on the face and lower part of the limbs, displaying a centrifugal, or centre-fleeing, distribution. In chicken pox, the progression from macule to pock is so rapid that the early stages are often not noticed; within a few hours the macules have changed to vesicles and, by the end of 24 hours, the first scabs have appeared. In smallpox, a macule takes three days to change to a vesicle, and it is eight days before the first scab is formed.

(A.B.C./R.D.F./Re.G.)

Eradication of smallpox

Although historically one of the great epidemic diseases of mankind, smallpox has now been conquered. Eradication of the disease by vaccination is possible since only human beings are carriers and the virus does not naturally live outside the human body. From 1967 to 1976 the World Health Organization carried out an eradication program of universal vaccination in Africa, South Asia, South America, and Indonesia, the last areas of endemic smallpox. Except for an epidemic in Somalia in 1977, reports thereafter were only of isolated and scattered cases of the disease.

Chicken pox, or varicella, is still ubiquitous among human beings. It is characterized by fever and rash, as described above. Complications, although uncommon, include secondary invasion of the skin by bacteria and spread of those bacteria to virtually any organ system in the body. Diagnosis is established by the typical symptoms. Chicken pox is a serious illness in adults and in children who suffer other chronic illness, such as leukemia or other malignancies. In these cases, high fever, muscle aches, and malaise are more common.

The chicken pox virus can persist for many years after recovery without causing problems. It can then be activated and cause a unilateral vesicular (blister) eruption in a specific, local area of the body. This is called a herpes zoster infection, or shingles, and severe pain is felt in the area of the blisters, which last from three to five days and then subside on their own. Specific antiviral agents are being evaluated for the treatment of severe chicken pox and herpes zoster.

Measles (rubeola). The viral illness called measles is marked by fever, cough, cold, conjunctivitis, and a characteristic rash. When no complications occur, the illness lasts 10 days and requires only the control of fever and bed rest. Mortality is rare except in parts of the world where malnutrition is severe; there, death rates for measles can approach 50 percent. Immunization to prevent measles should be given routinely at 15 months of age.

Complications of measles are frequent and include a superimposed bacterial ear infection or pneumonia or a primary measles lung infection. Encephalitis is a rare occurrence. Measles has also been associated with a large number of other manifestations and complications. Of historical interest was the outbreak of a severe, often fatal, form of measles called black measles, which was characterized by hemorrhagic skin eruptions. There were also signs of encephalitis and pneumonia and often massive bleeding from the nose, mouth, and bowel.

Measles virus can invade various organ systems and cause hepatitis, appendicitis, and gangrene of the extremities. Persistent measles virus infection can cause a degenerative central nervous system disease called subacute sclerosing panencephalitis (SSPE), in which there is a gradual onset of progressive behavioral and intellectual deterioration. Motor incoordination and impairment of speech and sight subsequently develop. The final stages of stupor, dementia, blindness, and death occur within six to nine months. There is no treatment for SSPE.

Mumps. Infection with mumps virus was once common in childhood, but the frequency of infection was drastically reduced with the introduction of routine immunization for prevention of the disease in 1967. Symp-

toms of the disease are low-grade fever and a typical swelling and inflammation of the parotid glands (those glands located behind the ears that are responsible for the production of saliva).

Mumps virus usually enters various organs and causes a more generalized infection than is frequently recognized. Meningitis, arthritis, and kidney involvement can occur, but the most feared complication is orchitis (an inflammation of the testes that can result in sterility). Diagnosis of mumps is usually made by clinical examination. Treatment is primarily supportive because no specific antiviral therapy is known to be effective. Mortality is rare.

German measles (rubella). German measles is generally a mild illness characterized by fever, sore throat, anorexia, swollen glands, headache, runny nose, and red eyes, all of which generally precede the rash. The duration and severity of the illness is variable and complications are rare. The most significant illness caused by the rubella virus occurs when mothers acquire the infection during the first two months of pregnancy. Infants born after such infections may be congenitally deaf, be blind as a result of cataracts, or have heads that are too large or too small. If a woman of childbearing age has not had a natural infection with rubella virus, she should be immunized prior to pregnancy.

Erythema infectiosum. Erythema infectiosum (also called fifth disease, ring rubella, or large spotted disease) is an acute communicable disease that occurs in local outbreaks and community epidemics. It is associated with a distinctive rash on the face. The rash at first looks as if the infected person has been slapped on the cheek and then acquires a generalized, lacelike appearance. The illness is usually mild and without complications.

Yellow fever. Yellow fever has been one of the great epidemic diseases of tropical Africa and South America. It is caused by a virus that is held in natural reservoir by forest monkeys and transmitted to people by mosquitoes. Yellow fever is characterized by the sudden onset of fever, chills, headache, photophobia (pain upon looking at light), backache, nausea, and vomiting. The fever usually has a two-stage course during which the pulse slows and kidney involvement occurs along with bleeding disorders. The skin and eyes may appear yellow (jaundice), a symptom that gave rise to the disease's popular name. Diagnosis is usually established by isolation of the virus and by the patient's history of having been in an area in which the disease is endemic. Treatment is supportive and is designed to correct the acid–base imbalance and electrolyte abnormalities caused by vomiting, heart failure, and kidney derangements. Mortality is less than 5 percent.

Immunization is the most practical and reliable way to prevent yellow fever in areas in which it is endemic. Mosquito control is an essential component of any program designed to prevent the spread of yellow fever and of other diseases caused by arboviruses such as Eastern equine encephalitis, Western equine encephalitis, St. Louis encephalitis, dengue fever, and Venezuelan equine encephalitis. (R.D.F./Re.G.)

Poliomyelitis. Poliomyelitis is a disease that causes inflammation and destruction of the anterior horn cells (motor cells) of the spinal cord. The virus enters by the mouth and multiplies in the throat and intestinal tract. It is usually a harmless parasite of human beings, for in only one in 1,000 cases does the virus reach the central nervous system. When it does, it attacks nerve cells and prevents them from sending out impulses to muscles. The result is paralysis of muscles, the extent of which depends on where the virus strikes and the number of cells that it destroys.

In some people the virus invades the bloodstream and gets no further, causing a vague flulike illness called the minor illness of poliomyelitis. When the virus enters the central nervous system, patients get signs and symptoms of meningitis, the major illness. They have headache, stiff neck, and vomiting. Many recover at this stage, but a few develop paralysis.

Paralysis of polio

The impulses that constantly move down the motor nerve from the spinal cord to muscles are cut off in poliomyelitis when the nerve cells are destroyed. As a result, muscles become limp and cannot contract. This is known as flaccid paralysis and is the type found in poliomyelitis. Most commonly, some of the limb muscles are paralyzed. The abdominal muscles or the muscles of the back may be paralyzed, and this paralysis may affect posture. The neck muscles may become weak, so that the head cannot be raised. Paralysis of the face muscles may cause twisting of the mouth or drooping eyelids. Paralysis of the throat or of the muscles of breathing can threaten life.

The poliovirus does not always completely destroy nerve cells. By the end of a month power returns to apparently paralyzed muscle, and by the end of six months recovery can be complete. If the nerve cells are destroyed completely, however, paralysis is permanent. There is no treatment for poliomyelitis other than supportive care. Most important in prevention of the disease is routine active immunization.

Rabies. Rabies virus is associated with dogs, foxes, jackals, and wolves, among other smaller animals, and has an odd association with vampire bats. In the plains of South America, these bats feed on the blood of cattle, and the death of cattle from paralytic rabies is a serious economic problem there. The bats themselves become rabid and bite other bats, but the spread of rabies among bats probably takes place mainly through the air in the caves in which they roost.

Rabies virus travels quickly in a bitten animal from the bite to the central nervous system. It multiplies in the nerve cells and after a time spreads down some nerves and reaches the salivary glands. When a dog gets rabies, its temper changes quickly. It is often more affectionate but soon grows restless and wanders off, ready to bite anyone who gets in its way. After a few days it returns home, unable to bite any more because the muscles of its throat are paralyzed; it seeks only a quiet place to hide and dies there from the rapid spread of paralysis. Any human being whom it has bitten during the short illness acquires rabies unless preventive treatment is given.

Rabies in dogs

In human beings, rabies begins with vague symptoms of malaise, headache, and fever, but in a day or two a state of excitement begins. The person becomes anxious, sleepless, and fearful. The muscles of the throat become paralyzed so that the person cannot swallow or drink, and this leads to a dread of water (hydrophobia). The mental state varies from maniacal excitement to dull apathy—the term rabies means "madness"—but soon the person falls into coma and usually dies in less than a week.

Once the earliest symptoms have appeared there is probably no cure for rabies, although the use of curare and artificial breathing may be lifesaving. The incubation period (the time that elapses between the bite and the first symptom) is usually one to three months. This provides a chance to interrupt the otherwise inevitable progress of the infection. The bite should be washed immediately and the bitten patient should receive a dose of antirabies serum. This antibody is derived from horses or human beings; it provides the patient with already prepared antibodies against the rabies antigen. Active rabies immunization is also initiated to allow the patient's body to make its own antibody. (A.B.C./R.D.F./Re.G.)

Herpes simplex viruses. Viruses of the herpes viradae group are DNA viruses that cause a multiplicity of diseases. Those causing infections in human beings are the varicella-zoster virus, which causes chicken pox or shingles; the Epstein-Barr virus, which causes infectious mononucleosis; the cytomegalovirus, which is most often associated with infections of newborn infants and immunocompromised people; and herpes simplex virus, which causes cold sores and herpetic venereal (sexually transmitted) diseases.

The herpex simplex virus (HSV) has assumed increasing importance because of its ability to cause venereal disease. There are two serotypes of herpes simplex, HSV-1 and HSV-2. Primary infection can be an isolated event, with or without the symptoms of ulcerative lesions and systemic illness.

HSV-1 is the common cause of cold sores. The primary infection usually occurs in childhood and is without symptoms in 50 to 80 percent of cases. Between 10 and 20 percent of these individuals have recurrences precipitated by emotion or by other illness. HSV-1 can also cause infec-

tions of the eye, central nervous system, and skin. Serious infections leading to death may occur in immunocompromised persons.

Herpetic venereal diseases

HSV-2 is associated most often with herpetic lesions of the genital area (venereal disease). HSV-1, however, can be recovered from 15 percent of genital lesions. The frequency of asymptomatic venereal infection is not known. Recurrences are the rule, especially with HSV-2.

In herpetic venereal diseases, the lesions are small, red, painful spots that become filled with fluid and then rupture, leaving eroded areas that eventually become scabbed. These primary lesions occur from two to eight days after exposure and may be present for up to three weeks. Viral shedding and pain usually resolve in two weeks. When infections recur, the duration of the pain, lesions, and viral shedding is approximately 10 days. The involved area includes the vagina, cervix, vulva, and, occasionally, the urethra in females and the head of the penis in males. The disease is usually transmitted by sexual contact.

Diagnosis of herpetic venereal disease is important because newborn infants born vaginally to mothers with active HSV-2 infection are at increased risk for generalized infection, including that of the central nervous system. HSV-2 can cause death in 60 percent of infants so affected and severe mental retardation in 20 percent of the surviving infants.

Examination of the fluid from the lesions shows viral particles within the nuclei of tissue cells and the presence of multinucleated giant cells. Diagnosis can also be made by examination of the blood for a change in the serum antibody levels as evidence of active infection. The elevation of antibody levels is not as striking in recurrent infections as it is in primary infections.

Early detection and prevention of the spread of herpetic venereal disease is important. In patients with the primary herpetic infection, an antiviral agent called acyclovir may be effective in diminishing the duration of symptoms and the period of time during which the virus may be recovered from the lesions. It is only effective before the latency state is established, however. During latency, when the virus lives in tissues without causing symptoms, it is protected against destruction. Subsequent reactivation of the virus can occur in response to a multitude of stimuli.

CHLAMYDIA

Chlamydia are intracellular parasites found in many vertebrates, including birds and human beings and other mammals. The diseases they cause were first recognized in the early 20th century.

Clinical illnesses are caused by two species, *Chlamydia trachomatis* and *C. psittaci.* The former is a frequent cause of genital infections in women and, if an infant passes through an infected birth canal, it can produce disease of the eye and pneumonia in the newborn. Young children sometimes develop ear infections, laryngitis, and upper respiratory tract disease from *Chlamydia.* Such infections can be treated with erythromycin.

Chlamydia psittaci produces psittacosis, a disease in human beings that results from exposure to the discharges of infected birds. The illness is characterized by high fever with chills, a slow pulse rate, pneumonia, headache, weakness, fatigue, muscle pains, anorexia, nausea, and vomiting. The diagnosis is usually suspected if the patient has a history of exposure to birds and is confirmed by blood tests. Mortality is rare and specific antibiotic treatment is available. (R.D.F./Re.G.)

RICKETTSIAS

The rickettsias are a family of microorganisms named for a U.S. pathologist, Howard T. Ricketts, who died of typhus in 1910 while investigating the spread of the disease. The rickettsias cause a group of diseases in human beings characterized by fever and a rash, and all depend for survival on passing at some stage through the body of a louse, mite, tick, or flea.

The great epidemic form of the disease, louse-borne typhus, is caused by *Rickettsia prowazekii,* which differs from most other rickettsias, except for that which causes trench fever, in that human beings and lice are its only hosts. It needs no other environment than that which human beings, when their hygiene is poor, provide.

Human beings get the other diseases only when they break into a cycle in nature in which the rickettsias live. In murine typhus, for example, *Rickettsia mooseri* is a parasite of rats conveyed from rat to rat by the rat flea, *Xenopsylla cheopis;* it bites human beings if they intrude into its environment. Scrub typhus is caused by *R. tsutsugamushi,* but it normally parasitizes only rats and mice and other rodents, being carried from one to the other by a small mite, *Leptotrombidium* (previously known as *Trombicula*). This mite is fastidious in matters of temperature, humidity, and food and finds everything suitable in restricted areas, or "mite islands," in South Asia and the western Pacific. It rarely bites human beings in their normal environment, but if people invade its territory en masse it attacks, and outbreaks of scrub typhus follow.

Scrub typhus

The spotted fevers are caused by rickettsias that spend their normal life cycles in a variety of small animals, spreading from one to the other inside ticks; these bite human intruders and cause African, North Asian, and Queensland tick typhus, as well as Rocky Mountain spotted fever. One other spotted fever, rickettsialpox, is caused by *Rickettsia akari,* which lives in the body of the ordinary house mouse, *Mus musculus,* and spreads from one to another inside the house mite *Allodermanyssus sanguineus.* This rickettsia is probably a parasite of wild field mice, and it is perhaps only when cities push out into the countryside that house mice catch the infection.

Typhus. Epidemic typhus has also been called jail fever, war fever, and camp fever, names that suggest overcrowding, underwashing, and lowered standards of living. The body louse, which spreads the disease, has a powerful sucking mouth. As it sucks the blood of a typhus victim, rickettsias pass into the louse's gut, where they invade the intestinal cells. They multiply and the cells burst after a few days, releasing hordes of rickettsias into the louse's intestinal canal. These either reinfect other cells or are passed out in the louse's feces. Lice leave a body when it gets too hot from fever or too cold from death and crawl to another human host. The clothing of a heavily infested typhus patient is contaminated with louse feces, and careless removal of it may raise a cloud of infected dust in the air and, in this way, spread typhus to others, especially physicians and nurses.

About 10 days after being bitten, a person falls ill. Intense headache and fever are always present, and after a few days a rash begins. The sick person is flushed, and his eyes are bleary. If the disease is untreated, circulation becomes sluggish and there may be spots of gangrene on the fingers, genitals, nose, and ears. Signs of pneumonia or kidney failure are common. Soon the person becomes stuporous and then may lapse into coma and die. Much depends on age—the young tend to recover, the aged to die.

Tetracycline and chloramphenicol have a dramatic curative effect, however, and, if treated early enough, few die. The disease can be diagnosed clinically during epidemics and by laboratory tests.

The prevention of epidemic typhus requires the elimination of human body lice. Antityphus vaccine is also effective; two doses are given about one month apart and a third is given after three months. Thereafter, doses at intervals of several months are given to people who are exposed to the disease. The vaccine provides significant protection against attack and almost complete protection against death.

Rocky Mountain spotted fever. Rocky Mountain spotted fever is the form of tick-borne typhus that occurs in the Western hemisphere. It is caused by *Rickettsia rickettsii,* which is spread to human beings by the wood tick (*Dermacentor andersoni*), dog tick (*D. variabilis*), lone star tick (*Amblyomma americanum*), and several others in different parts of the Americas. Despite its name, Rocky Mountain spotted fever is most common on the eastern coast of the United States and, in fact, has been found in every state. It is a disease of the summer and early fall, for only then are ticks active.

The illness begins with headache, fever, and chills, soon followed by pains in the bones and joints and great weak-

Severe rash

ness and prostration. A rash develops in the first week of illness, beginning on the extremities and moving to the trunk. It is more profuse than the rash of epidemic typhus and affects the face as well as the body. In some people the colour of the rash deepens after a day or two, and in the worst cases it turns purple with blood. By the end of a week in severe cases, the patient shows signs of brain irritation and may be agitated, sleepless, or delirious. Breathing becomes laboured and circulation poor, and areas of gangrene may develop on the hands and feet. In the worst cases, the patient may become comatose and die, but in most cases fever gradually abates and the patient slowly recovers. If given early in the course of the disease, chloramphenicol or tetracycline is an effective treatment.

Q fever. Q fever is a disease caused by the rickettsial agent *Coxiella burnetii.* The Q stands for query, for its cause was at first unknown. It spreads rapidly in cows, sheep, and goats, and whole herds, flocks, and areas often are contaminated with the *Coxiella* organisms. The germs are passed in vast quantities in milk, urine, and feces, in vaginal discharges, and on placentas, and they survive for long periods dried in dust.

Q fever in human beings tends to occur in localized outbreaks. In the middle of a heavily infested area, it may break out suddenly on one farm only, affecting everyone on it, and die out again within a few days. The clinical symptoms are those of fever, chills, severe headache, and pneumonia. The disease is usually mild and self-limited and complications are rare. Treatment with tetracycline or chloramphenicol shortens the duration of illness.

(A.B.C./R.D.F./Re.G.)

MYCOPLASMAS AND UREAPLASMAS

Mycoplasmas and ureaplasmas are the smallest free-living microorganisms. They are ubiquitous in nature and capable of causing widespread disease, but the illnesses they produce in human beings are generally milder than those caused by bacteria. There are nine species of the genus *Mycoplasma* and one of the genus *Ureaplasma* that cause disease in human beings.

Mycoplasma pneumoniae is the most important member of its genus. It causes pneumonia, colds, pharyngitis, sinusitis, bronchitis, rashes, heart infection, liver disease, arthritis, and disease of the brain and spinal cord. *M. pneumoniae* is associated with 20 percent of all cases of pneumonia in adults and children over five years of age. Patients have fever, cough, headache, and malaise and, upon physical examination, may found to be suffering pharyngitis (inflamed throat), enlarged lymph nodes, and pneumonia. There may also be an accompanying ear or sinus infection, bronchitis, or croup. Diagnosis is established by chest X-rays and blood tests. Although treatment with erythromycin or tetracycline may shorten the illness, it can last for weeks.

Mycoplasmas may also cause a red, bumpy rash—usually on the trunk or back—that is occasionally vesicular (with blisters). Inflammation of the heart muscle and the covering of the heart (pericardium) is rare but can be caused by mycoplasmas. About one-fourth of the people infected with these organisms experience nausea, vomiting, diarrhea, and cramping abdominal pain. Inflammation of the pancreas (pancreatitis) or the liver (hepatitis) may occur, and infection of the brain and spinal cord is a serious complication.

Ureaplasmas

Ureaplasmas can be recovered frequently from the genital areas of healthy persons. The organism can cause inflammation of the urethra and has been associated with infertility, low birth weight of infants, and repeated stillbirths. In general, ureaplasma infections are mild, however. Tetracycline is the preferred treatment once the organism has been established as the cause of infection by microscopic examination of urethral secretions.

(R.D.F./Re.G.)

FUNGI

Fungi are large organisms that usually live on dead and rotting animal and vegetable matter. They are found mostly in soil, on objects contaminated with soil, and on plants and animals and may also be airborne. Although most fungi do not infect humans, they may invade the bloodstream and cause devasting disease in persons whose defenses against infection have been weakened by malnutrition, cancer, or the use of immunosuppressive drugs.

Coccidioidomycosis and histoplasmosis. Coccidioidomycosis and histoplasmosis are two rather similar fungal infections. Both the *Coccidioides immitis* and *Histoplasma capsulatum* fungi may cause severe generalized and fatal infections, but in both a common manifestation is a mild form of pneumonia and pneumonitis. The infected person may develop fever and chills, cough, and other symptoms that are indistinguishable from those caused by many other acute respiratory infections. X-ray examination of the chest may show a localized shadow or a more diffuse mottling of the lungs. Sometimes the X-ray appearance is similar to that of active tuberculosis, and, as recovery ensues, hard calcified scars can be seen on the X-ray film, very much like those seen in old, healed tuberculosis. In a very few cases there are widespread areas of destruction in the lungs, and the fungi spread through the body to the liver, spleen, heart, and brain, sometimes causing death. Most commonly, however, infection with *C. immitis* or *H. capsulatum* causes no symptoms or discomfort at all.

Both fungi can be found in the soil. Coccidioidomycosis occurs most often during dry spells, and dust storms have caused outbreaks of the infection in human beings. During World War II, when such outbreaks occurred among airmen, the planting of grass in airfields and the oiling of dusty roads kept the *Coccidioides immitis* on the ground and stopped the spread of the disease. *Histoplasma capsulatum* prefers moist, shady conditions and is found in woods, caves, cellars, silos, and old chicken houses. The last seem to be especially important because chickens often are infected and pass enormous numbers of organisms in their droppings. The use of chicken manure in gardens may lead to histoplasmosis in humans. Dogs, rats, mice, bats, pigeons, skunks, and probably many other animals become infected and may help to spread the disease. Diagnosis of both coccidioidomycosis and histoplasmosis is made by serologic tests or by culture of the organism. Most cases do not require treatment, but patients with widespread disease can be treated with amphotericin B.

(A.B.C./R.D.F./Re.G.)

Aspergillosis. After candidiasis, Aspergillosis is the second most common fungal infection in patients with cancer. *Aspergillus* can have a variety of effects on human beings, ranging from no illness, to mild pneumonia, to overwhelming generalized infection. The fungus is usually introduced into the body by inhaling spores or through a cut or a surgical wound.

People with leukemia or other malignancies are unable to contain the organism in the lungs and develop widespread disease involving the liver, kidneys, skin, and brain. Diagnosis is difficult to establish because cultures are usually not positive and can sometimes be made only by tissue biopsy. Specific antifungal agents are available, but the invasive disease is rapidly fatal.

Candidiasis. *Candida albicans* is the fungus that most often infects human beings. It normally lives on the surface of the skin, where it causes no overt illness. When it invades the mucous membranes of the mouth or vagina, it produces a mild infection called thrush. In patients whose resistance to infection is diminished, however, *C. albicans* can gain entry to the bloodstream and cause a generalized disease. Specific treatment is available, but mortality may be high in patients with a generalized infection.

Cause of thrush

(R.D.F./Re.G.)

PARASITES

Trichinosis. The worm that causes trichinosis, *Trichinella spiralis,* is tiny, not more than four millimetres (0.16 inch) long when fully mature. It infects an enormous number of hosts, from mice to rats to polar bears. In one stage of its existence, it lives in the muscle of the host animal, where it becomes surrounded by tough fibrous tissues that form a cyst around it. The worm may live in the cyst for months or years. Unless the muscle of the host animal is eaten by another animal, the trichinella dies in the cyst and the cyst calcifies. If the host animal is eaten,

the hard coat of the cyst is dissolved in the stomach, and the larva pierces and settles down in the wall of the intestine of its new host. It matures quickly and mates, and the female then produces eggs from which minute larvae emerge. These pass into the lymph channels of the intestine, whence they enter the main lymph duct of the body, the thoracic duct. The thoracic duct empties into one of the large veins, and the tiny larvae then reach the heart, pass through the lungs, pass into the heart again, and enter the general circulation to disseminate throughout the body. They encyst in muscle, and the cycle begins again.

Human beings most commonly become infected with *Trichinella spiralis* by eating undercooked pork. A few days after exposure, the infected person becomes feverish and has vague abdominal symptoms. These are followed, a few days later, by pains in the joints, headache, and swelling of the face, a characteristic symptom. Severe pain develops in the muscles of the limbs, in the chest, and in the eyeballs, and breathing is often painful because the diaphragm is heavily infected. The illness continues for a week or two before it gradually subsides, but in some patients the condition worsens. The outlook depends upon the intensity of the infection. In some epidemics, mortality may be as high as 10 to 16 percent, but many people have attacks so mild that they are not recognized.

Examination of the blood is an aid to diagnosis of trichinosis. Pieces of muscle may be taken for microscopic examination, and cysts may then be seen in the muscle fibres. There is no known specific treatment.

Fascioliasis. *Fasciola hepatica,* the liver fluke, is a leaf-shaped worm about four centimetres (1.6 inches) long that grows in the liver of various animals, especially in cattle and sheep. The eggs pass down the bile duct and are excreted in the feces. If the eggs get into pools of water, they hatch after a few weeks, and the larvae must find their way into a small water snail. There, in the course of about two months, they multiply and emerge as free-swimming larvae. These finally attach themselves as cysts to grass or leaves of plants growing in the water. The cysts resist drying when the water recedes, and, if they are later swallowed with the grass or plants, they hatch in the host's intestine, migrate across the abdominal cavity, pierce the liver, and settle down in the bile ducts, where they cause obstruction to the flow of bile and inflammation in surrounding liver tissue.

Life cycle of the liver fluke

Fascioliasis is a disease of cattle and sheep that has serious economic consequences. Human beings can be infected by eating wild watercress. The cysts hatch in the person's intestine and migrate to the liver and other organs. These migrating larvae may cause unexpected complications; they have been found, for example, in the tissues of the larynx. The symptoms of fascioliasis in human beings are fever, sweating, loss of weight, abdominal pain, and anemia, and sometimes a fleeting rash. In the blood there is an increase in white blood cells of the type called eosinophils, a common finding in worm infections. There is no effective treatment. The best way to prevent fascioliasis in human beings is to forbid the sale of watercress unless it is grown in commercial beds that are protected from pollution by animal feces. (A.B.C./R.D.F./Re.G.)

Schistosomiasis. Schistosomiasis is a disease caused by blood flukes and characterized by inflammation of the intestines, bladder, liver, and other organs. It affects at least 200,000,000 people in Africa, Asia, South America, and the Caribbean.

Once the blood fluke enters an animal host, it lives in the veins of the intestines and bladder, where it deposits a large number of eggs. The eggs pass in the host's stool, find their way to freshwater, hatch, and release freely mobile forms called miracidia. The miracidia infect snails that, if ingested, carry the infection to a new host. Humans most often acquire schistosomes by working, bathing, or swimming in water containing the infected snails. In the human host the miracidia mature and migrate to the lungs and liver, depositing eggs. It is the body's reaction to these eggs that leads to chronic illness.

Symptoms of schistosomiasis consist of unexplained gastrointestinal disease, enlarged liver and spleen, the coughing up of blood, and blood in the urine and eosinophils in the blood. Diagnosis is established by a history of travel to the areas of the world where schistosomiasis is endemic and the presence of the fluke in a blood smear.

Treatment requires eradicating the blood fluke from the body at all stages of its life cycle. Death occurs in less than 1 percent of untreated individuals but may result from associated severe bleeding, renal failure, or heart failure. Irreversible damage can occur in the liver, lungs, or urinary tract. Schistosomiasis can be prevented by avoiding contact with water infested by schistosomes.

(R.D.F./Re.G.)

Scabies. The parasite that causes scabies, *Sarcoptes scabiei,* is a mite just visible to the unaided eye. Despite its small size, it has quite formidable mouthparts and eight very powerful legs. With its hind legs, the female of the species hitches itself into the perpendicular to the skin and, with its mouthparts thus closely applied to the skin, cuts its way down into the horny layer. Altering its direction, the mite then tunnels horizontally and gouges out an easily recognizable burrow up to several inches long and usually not very straight. The female mite prefers the skin between the fingers and toes, on the wrist and elbows, in the armpits, below the breast, and on the male genitals. It does most of its moving at night. Sometimes it is visible at the far end of one of its burrows and is easy to pick up on a needle and examine under a microscope. The female mite lays eggs that hatch into larvae and undergo two molts before becoming female adults. The development from egg to laying adult frequently takes about three weeks.

The mite passes from person to person by close contact. Scabies is characteristically a disease of wartime, for living standards then drop, washing may be difficult, and people may be crowded together. The disease afflicts many schoolchildren, with periodic epidemics occurring in classrooms even among persons with strict habits of hygiene and adequate living conditions.

The burrowing of the mite causes irritation; in addition, a rash breaks out in parts of the body where there are no burrows—on the buttocks, over the shoulder blades, and on the abdomen. The cause of the rash is not clear; it is probably a reaction to the presence of the parasite on the body. It is intensely irritating and interferes with sleep.

The means of prevention of scabies are obvious. They are the wearing of clean underclothes and frequent changing of clothing and washing, conditions not always possible in time of war or mass migration. Effective treatment consists of painting the affected person from neck to feet with a lotion that kills the mite; benzyl benzoate is often used. It is desirable to treat an entire family or a group of people living together, for often the early burrowing stage of the disease causes no irritation and a person can be infested without knowing it.

Malaria. Malaria (from Italian, "bad air") is a serious environmental disease caused by a protozoan parasite called *Plasmodium.* Although it has nothing to do with bad air, it does rise from still bodies of water and spread across the surface of the land, conveyed to human beings by the *Anopheles* mosquito.

The mosquito begins its life as an egg deposited on water. The larva hatches out in two or three days and spends the next few weeks near the surface of the water. It feeds on microorganisms in the water and breathes in oxygen from the atmosphere through a spiracle, or breathing orifice, at one end of its body. After molting four times it changes into a pupa and, from the pupa, the adult mosquito emerges after a few days. The young mosquito floats for an hour or so on the pupa case until its wings dry and harden, and then it flies into the air, where the adult female is soon mated by a male. The pair then separates, the male to seek plant juices as its food, the female to feed on human blood. The female is armed with mouthparts especially adapted for cutting into human skin, which it pierces in its search for a blood vessel; at the same time, it squirts some saliva into the wound. If plasmodia, the malaria parasites, were present in the human victim, some will have entered the mosquito's stomach. The sexual forms of the parasite (gametocytes) mate in the mosquito's stomach and the fertilized female gamete (sex cell) encysts in the stomach wall. After one to three

Life cycle of plasmodia

weeks the cyst bursts, and a large number of young sexual parasites, the sporozoites, emerge. These find their way to the mosquito's salivary glands, whence they are injected into a human host the next time the mosquito bites. The parasites multiply in the human liver for a week or so and then spill into the bloodstream, entering many red blood cells. There they multiply asexually and, after two or three days, having destroyed the red blood cells, they enter more red blood cells and the cycle continues. Eventually sexual forms, the gametocytes, emerge. They do not multiply or mate in the human body, but, if the victim is bitten again, they are sucked into the mosquito's stomach, and the life cycle of the malaria parasite begins anew.

Four species of *Plasmodium* infect human beings—*Plasmodium vivax, P. ovale, P. malariae,* and *P. falciparum.* They all cause malaria, the main feature of which is recurring bouts of fever that shake a person with alternating shivering and sweating. According to the rate at which the asexual parasites multiply and burst out of the infected person's blood cells, these attacks may occur every day (quotidian malaria), every third day (tertian malaria), or every fourth day (quartan malaria). Different cycles often take place in the red blood cells at the same time, blurring the periodicity of the shivering attacks. *P. falciparum* tends to sporulate (multiply asexually) in 36 hours rather than 48 hours, and the attacks come on more quickly than every third day; the disease is then called subtertian malaria or, because of its severity, malignant tertian malaria.

Malaria can mimic almost any other disease. The infected person becomes anemic from loss of red blood cells and may be jaundiced, have gastroenteritis, show signs of bronchitis, have a hectic continued fever, or even collapse and be cold, with no fever at all. Any illness in a person who may have been exposed to infection must be regarded as malaria until proved otherwise. In most persons the attacks cease even without treatment, though they recur later. But when *P. falciparum* is the parasite, the initial symptoms of mild fever, headache, aches and pains, and minimal vomiting or diarrhea change suddenly after a day or two to an overwhelming illness, with signs of liver, kidney, or respiratory failure or with coma from invasion and blocking of the small blood vessels of the brain, a characteristic of cerebral malaria. People with these symptoms may die unless they have the most urgent and skillful medical attention, and it is one of the tragedies of malaria that the early illness of malignant malaria, which is easily treatable, so often is not identified until the unmistakable but desperate state of cerebral malaria sets in.

Cerebral malaria

Treatment is simple in theory. Chloroquine, one of several excellent antimalarial drugs, cures early falciparum malaria and brings attacks of the other forms to an end, though they recur later unless the person is treated after the attack with drugs to prevent a relapse. Some strains of falciparum malaria have developed resistance to chloroquine, and other antimalarial drugs must be used. For the treatment of malignant or cerebral malaria, the antimalarial drug must be given intravenously without delay, and measures are taken to restore the red blood cell level, to correct the severe upset of the body's fluids and salts, and to get rid of the urea that accumulates in the blood when the kidneys fail.

The prevention of malaria can be tackled in two main ways—by trying to break contact between human beings and mosquitoes and by administering drugs to people exposed to bites to prevent the onset of the disease. The mosquito can be attacked in the larval and adult stages. Antimalarial drugs, if taken regularly, prevent the onset of the disease. They are taken before entry to an area in which malaria is endemic, regularly while in the area, and for some time after leaving it.

The malaria parasite, like most microorganisms, holds stubbornly to life. Malaria parasites resistant to drugs have appeared in several areas of the world, and combinations of old and new drugs have been tried to combat them. More than 100,000 possible drugs have been tested in the United States alone in the search for an effective one. The struggle against malaria continues. (A.B.C./R.D.F./Re.G.)

Amebiasis (amebic dysentery). *Entamoeba histolytica* is another parasite that causes acute and chronic infections, traditionally in the tropics. Amebiasis is initiated by ingesting food or water contaminated with amoebic cysts. In the proper environment, the cysts set up infection in the intestine, causing bloody diarrhea and abdominal cramps. Deep ulcers in the intestinal wall and pockets of infection in the liver can also occur.

Diagnosis is established by finding cysts in the infected person's stool by microscopic examination, although infection may be undetected in some individuals until an abscess develops in the liver. Specific therapy is available and the mortality rate is low. Spread of the disease can be controlled by boiling drinking water and by adequately disposing of human waste to avoid the contamination of food.

Giardiasis. The protozoan parasite *Giardia lamblia* causes the diarrheal condition known as giardiasis. It is transmitted directly from person to person, and the infection is quite frequent in crowded institutions. Clinical symptoms include large, foul-smelling, fatty stools, heartburn, malaise, and lethargy. The diarrhea is usually self-limiting and many occurrences of the illness are undiagnosed, although a chronic malabsorption of nutrients from the intestines may develop in young infected children.

Diagnosis is established by finding the organism in stool specimens or in intestinal washings. The disease is generally mild and mortality is rare, specific therapy being available.

Common worm infections. The worm *Ascaris lumbricoides* causes one of the most prevalent infections in the world. It lives in the soil and its eggs are ingested with contaminated food. The eggs hatch in the human intestine and the worms then travel through the bloodstream to the liver, heart, and lungs. They can cause pneumonia, perforations of the intestine, or blockage of the bile ducts, but infected people usually have no symptoms beyond the passage of worms in the stool. Specific treatment is available and prognosis is excellent.

Infections are also caused by whipworms, genus *Trichuris,* and pinworms, *Enterobius vermicularis,* each popularly named for its shape. The former is parasitic in the human large intestine and may cause chronic diarrhea. The latter can be found throughout the gastrointestinal tract, especially in children, and can cause poor appetite, loss of weight, anemia, and itching in the anal area (where it lays its eggs). Both conditions are easily diagnosed and treated with drugs. (R.D.F./Re.G.)

Whipworm and pinworm

ACQUIRED IMMUNODEFICIENCY SYNDROME (AIDS)

Acquired immunodeficiency syndrome (AIDS) is the progressive destruction of the immune system by the human immunodeficiency virus (HIV). First diagnosed in 1981 in a small population of homosexual men, AIDS has spread beyond its original risk group to include heterosexual males and females with multiple sex partners, hemophiliacs, and habitual intravenous drug users. Infants and small children have formed one of the fastest growing risk groups. HIV is spread by exchange of infected blood or blood products, by sexual intercourse, and by an infected mother to a fetus through the placental wall or to a suckling infant through breast milk. Although HIV has been recovered in the saliva of infected persons, in amounts lower than in the blood, studies have shown that no infections resulted after skin or mucous-membrane exposure to the infected saliva. There is no evidence to indicate that insect vectors can transmit the virus.

The results of research that was carried on between 1982 and 1984 proved conclusively that HIV is the causal agent of AIDS. HIV is a retrovirus, which means that its genetic material is ribonucleic acid (RNA) rather than deoxyribonucleic acid (DNA, the genetic material of the human host cell, which encodes the instructions for the operations of the cell). HIV has also been called human T-cell lymphotropic virus type III (HTLV-III), after the previously discovered human retroviruses, and lymphadenopathy-associated virus (LAV). An international commission renamed the virus human immunodeficiency virus (HIV) in 1986.

Once viruses infiltrate the body, they generally invade the target cell and infiltrate its DNA. When the normal

biosynthetic pathways are stimulated, viral DNA is transcribed, along with the cell's DNA, into RNA, which then directs the assembly of proteins. Because HIV is a retrovirus, it must translate its RNA-encoded instructions into a form recognizable to the DNA of the host cell. HIV uses an enzyme called reverse transcriptase to bring about the manufacture of a strand of DNA complementary to its RNA genome. The viral DNA infiltrates the host cell's DNA and is then replicated using the body's own biosynthetic machinery. Once HIV infiltrates the DNA of a host cell, the infection is permanent and the host continues to replicate the viral DNA.

HIV targets

The major target cells of HIV are helper T lymphocytes, or T4 cells, and, to a lesser extent, macrophages. T4 cells recognize foreign complex macromolecules (antigens); the antibodies bind the antigens and inactivate or destroy them. T4 cells also moderate cell-mediated immunity, and they influence the activity of phagocytic macrophages.

HIV binds to specific receptors (CD4 molecules) on the surfaces of T4 cells and some macrophages, effectively destroying these cells. The result is a serious impairment of the immune system that renders the host susceptible to pathogens (such as viruses, fungi, parasites, and many bacteria) that are normally controlled by these agents of the host's immune system. The host becomes increasingly vulnerable to infection. The virus also can be found in certain cells of the intestine and in the bone marrow. It has also been found to cross the blood–brain barrier, infecting cells of the central nervous system, such as the brain and spinal cord, and causing disorders of the central nervous system.

AIDS runs a fairly predictable, relatively slow, staged course. Exposure to HIV through any of the modes of transmission discussed above, rather than membership in any of the major risk groups, is the predominant reason for infection. After exposure to the virus, there is an asymptomatic period, estimated to be between six weeks and one year, before the virus or an antibody to the virus is detectable in the blood. At the time the virus is first detectable, it begins to replicate in the T4 cells, a process that eventually destroys them. The level of T4 cells, however, generally remains near normal. This period may be asymptomatic as well, or the infected person may experience fatigue, fever, swollen glands, and occasionally a rash. These symptoms, however, disappear within weeks. An early clinical symptom of infection is the development of chronic lymphadenopathy (chronic swollen lymph glands), a stage in the development of AIDS that lasts from three to five years.

A precipitous drop in the level of T4 cells indicates a decrease in immune functioning and the progression to a new stage. This stage characteristically shows no overt disease; however, tests show a delayed, and later absent, ability of the immune system to begin a cellular immune response against specific proteins injected under the skin. Approximately 1½ years after this stage, the level of T4 cells continues to decrease, causing, initially, infections of the skin and mucous membranes and, later, disseminated systemic infections. The result is an extremely severe decrease in immune function.

Throughout the stages of AIDS, a myriad of nonspecific symptoms, designated AIDS-related complex, or ARC, may develop. ARC includes unexplained fevers, night sweats, chronic diarrhea, and wasting. Disorders of the central nervous system may develop. Cancers associated with AIDS have been reported, including lymphomas, cancers of the rectum or tongue, and Kaposi's sarcoma, which is the development of tumours in the skin and linings of the internal organs. Kaposi's sarcoma is often present during the early stages of HIV infection. *Pneumocystis carinii* pneumonia and cytomegalovirus are extremely common in AIDS patients. Throughout, the patient is susceptible to a variety of opportunistic infections—*i.e.,* infections that are normally controlled by the immune system, but that proliferate if the immune system has broken down. Death usually results from complications that develop from the compromised immune system.

ARC

Since viruses take over the biosynthetic machinery of the host cell, drugs that target and destroy the viruses also generally damage the host cell. As of the late 1980s, no effective vaccine against HIV was available; however, drug therapies to combat and stem the effects of AIDS were in clinical trial stages. Azidothymidine (AZT), originally an anticancer drug, has met with some success in the treatment of AIDS.

Education remains the most immediate way to stem the growing number of AIDS cases. Data show that, although AIDS has expanded past its original risk groups, only a fraction of the new AIDS cases cannot be traced back to one of the three modes of transmission discussed above. Testing for the presence of HIV or antibodies to HIV after exposure remains an important method for prompt diagnosis and thus the administration of early medical care that may delay or prevent complications. An early awareness of infection with HIV also provides an opportunity to change one's behaviour and prevent inadvertent transmission. (Ed.)

Effects of environment on human disease

HUMAN ACTIVITY

Social patterns. Human beings are social animals. As a result, human social habits and circumstances influence the spread of infectious agents. Poorer families tend to live in more crowded conditions, which facilitate the passage of disease-causing organisms from one person to another. This is true whether the germs pass through the air from one respiratory tract to another, or whether they are bowel germs and depend for their passage on close personal hand-to-mouth contact or on lapses of sanitation and hygiene.

The composition of the family unit is also important. In families with infants and preschool children, infection spreads more readily, for these youngsters are both more susceptible to infection and, because of their faulty hygiene habits, more likely to share their microbes with other family members and with friends. Because of this close and confined contact, infectious agents are spread more rapidly.

Distinction must be made between disease and infection. The virus of poliomyelitis, for example, spreads easily in conditions of close contact (infection), but it usually causes no active disease. When it does cause active disease, it attacks older people much more severely than the young. Children in more crowded homes, for example, are likely to be infected at an early age and, if illness results, it is usually mild. In less crowded conditions, young children are exposed less often to infection; when they first encounter the virus at an older age, they tend to suffer more severely. The difference between infection and disease is seen even more clearly in the outcome of the infection. Under conditions of poor hygiene, poliovirus spreads rapidly in early childhood, leading more often to immunity than to illness; under high standards of hygiene, young children are exposed less frequently and fewer develop immunity in early life, with the result that paralytic illness, a rarity under the former conditions, is seen frequently in older children and adults. The pattern of infection and disease, however, can be changed. In the case of the poliomyelitis virus, only artificial immunization can abolish both infection and disease.

Population density. Density of population does not of itself determine the ease with which infection spreads through a population. In New York City, with its many high-rise dwellings, the density of the population per square mile is much greater than in some of the world's older cities, but the hepatitis virus, for example, spreads much faster in the latter. A family in a New York City apartment may never see the inhabitants of most of the other apartments in the block, while neighbours in an ancient Asian city are in daily contact. In New York state, the incidence of infectious hepatitis has been shown to vary inversely with population density—to be lowest in the city itself, higher in other urban areas of the state, and highest in rural areas. This pattern, of course, reflects the population's chances of contact and social habits, as well as its density.

Social habits. The vampire bats of Brazil, which trans-

mit paralytic rabies, bite cattle but not ranchers, presumably because on the plains of Brazil ranchers are few but cattle plentiful. Bat-transmitted rabies, however, does occur in human beings in Trinidad, where herdsmen sleep in shacks near their animals. The mechanism of infection is the same in Brazil and Trinidad, but the difference in social habits affects the incidence of the disease.

Malta fever

During the early 20th century in Malta, goats were milked at the customers' doors, and a *Brucella* species in the milk caused a disease that was common enough to be called Malta fever. When the pasteurization of milk became compulsory, Malta fever almost disappeared from the island. (It continued to occur in rural areas where people still drank their milk raw and were in daily contact with their infected animals.)

No less important alterations in environment occur when children in a modern community first go to school. Colds, coughs, sore throats, and swollen neck glands can occur one after the other. In a nursery school, with young children whose hygiene is less than perfect, outbreaks of dysentery and other bowel infections may occur, and among children who take their midday meal at school, food-borne infection caused by a breakdown in hygiene can sweep through entire classes. These are dangers against which the children are protected to some extent at home but against which they have no defense when they move to the school environment.

Changing food habits among the general population also affect the environment for human beings and microbes. Meals served in restaurants, for example, offer a greater danger of food poisoning if the standard of hygiene for food preparation is flawed. The purchase and preparation of poultry—which is often heavily infected with *Salmonella*—present a particular danger. If chickens are bought fresh from a farm or shop and cooked in an oven at home, food poisoning from eating them is rare. If poultry is purchased while it is deep-frozen and then not fully thawed before it is cooked, there is a good chance that insufficient heat penetration will allow the *Salmonella*—which thrive in the cold—to survive in the meat's centre and infect the people who eat it.

Temperature and humidity. At a social gathering, the human density per square yard may be much greater than in any home, and humidity and temperature may rise to levels uncomfortable for human beings but ideal for microbes. Virus-containing droplets pass easily from one person to another, and an outbreak of the common cold may result.

In contrast, members of scientific expeditions have spent whole winters in the Arctic or Antarctic without any respiratory illness, only to catch severe colds upon the arrival of a supply ship in the early summer. This is because viruses, not cold temperatures, cause colds. During polar expeditions, the members rapidly develop immunity to the viruses they bring with them and, throughout the long winter, they encounter no new ones. Their colds in the summer are caused by viruses imported by the crew of the supply ship. When the members of the expedition return on the ship to temperate zones, they again come down with colds, this time caught from friends and relatives who have spent the winter at home.

Movement. Movement into a new environment often is followed by an outbreak of infectious disease. For example, the crew members of a supply ship can be afflicted with colds upon arrival at an Arctic base as a result of catching viruses from those whom they came to relieve. In mass movements of population, infection is always a hazard. On pilgrimages and in wars, improvised feeding and sanitation lead to outbreaks of such intestinal infections as dysentery, cholera, and typhoid, and more have sometimes died in war from these diseases than have been killed in the fighting.

Measles epidemics

People entering isolated communities may carry a disease such as measles with them, and the disease may then spread with astonishing rapidity and often with enhanced virulence. A traveller from Copenhagen carried measles virus with him to the Faeroe Islands in 1846, and 6,000 of the 8,000 inhabitants caught the disease. Most of those who escaped were old enough to have acquired immunity during a measles outbreak 65 years earlier. In Fiji a disastrous epidemic of measles in 1875 killed one-fourth of the population. In these cases, the change of environment favoured the virus. Nearly every person in such "virgin" populations is susceptible to infection, so that a virus can multiply and spread unhindered.

In a modern city population, by contrast, measles virus mainly affects susceptible young children. When it has run through them, the epidemic must die down because of a lack of susceptible people, and the virus does not spread again until a new generation of children is on hand. With the use of measles vaccine, the supply of susceptible young children is reduced and the virus cannot spread and multiply and must die out.

An innocent change in environment such as camping can lead to infection if it brings a person into contact with sources of infection that are absent at home. Picnicking in a wood, a person may be bitten by a tick carrying the virus of one of several forms of encephalitis; as he swims in a canal or river, his skin may be penetrated by the organisms of Weil's disease. He may come upon some watercress growing wild in the damp corner of a field and may swallow with the cress almost invisible specks of life that will grow into liver flukes in his body, giving him an illness that is common in sheep but that can spread to human beings when circumstances are in its favour.

Occupation and commerce. In occupational and commercial undertakings, people often manipulate their environment and, in so doing, expose themselves to infection. A farmer in his fields is exposed to damp conditions in which microorganisms of disease may flourish. In clearing out a ditch, he may be infected with leptospires passed into the water in rats' urine. In his barns he may be exposed to brucellosis if his herd of cattle is infected or to salmonellosis or Q fever. Slaughterhouse workers run similar risks, as do veterinarians. A worker in a dock or tannery may get anthrax from imported hides, an upholsterer may get the disease from wool and hair, and a worker mending sacks that have contained bone meal may contract the disease from germs still clinging to the sack.

Workers in packing plants and shops often are infected from the raw meat that they handle; they are sometimes regarded as carriers and causes of outbreaks of *Salmonella* food poisoning, but as often as not they are victims rather than causes of the outbreak. Workers in poultry plants can contract salmonellosis, more rarely psittacosis or a viral infection of the eye, from the birds that they handle. Forestry workers who break into a reserve may upset the balance of nature of the area and expose themselves to attack from the undergrowth or the trees by insect vectors of disease that, if undisturbed, would never come into contact with human beings. Whenever people manipulate the environment—by herding animals together, by importing goods from abroad, by draining a lake, or by laying a pipe through swampy land, and in many other seemingly innocent ways—they run the chance of interfering with microbial life and attracting into their own environment agents of disease that they might not otherwise ever encounter.

THE INANIMATE ENVIRONMENT

Dust cannot cause infectious disease unless it contains the living agents of the infection. Yet the term inanimate is a convenient one to use when infectious disease arises from contact with an environment in which there is no obvious direct living contact between source and victim of an infection. A pencil is an inanimate object, but if it is sucked by a child suffering from scarlet fever and then by a second child, the organism causing the disease can be conveyed to the second child. Many such objects—a handkerchief or a towel, for example—may convey infection under favourable conditions, and when they do so, they are known as fomites.

Dust as a fomite

Dust is perhaps one of the most common inanimate elements capable of conveying disease. Organisms present in dust may get on food and be swallowed, settle on the skin and infect it, or be breathed into the respiratory passages. Some germs—of anthrax and tetanus, of Q fever, brucellosis, and psittacosis, for example—can live for long periods dried in dust. Under certain conditions all may

be dangerous, but under different conditions there may be no danger. The germ of tetanus, *Clostridium tetani,* is one of the most common germs in dust, but the incidence of tetanus varies greatly in different parts of the world. In many countries it is rare, in others common, and the difference may be related to slight differences in human behaviour or custom. The wearing of shoes in temperate climates protects the wearer against many wounds, while the barefoot child in the tropics sustains many puncture wounds of the feet that, although minor, may yet carry tetanus spores into the tissues of the foot where conditions may be ideal for germination and the production of toxin. Obstetrical practices in some parts of the world can lead to infection of the newborn child. Other influences, more subtle, such as changes of temperature or humidity, may affect the spread of diseases by dust. A long wet winter followed by a dry summer may encourage the growth of molds in hay, and the dust, when the hay is disturbed, may lead to infection of the farmer's lungs.

Water is not a favourable medium for the growth and multiplication of microorganisms, and yet many can survive in it long enough to carry infection to human beings. Cholera and typhoid are both waterborne diseases, and the virus of hepatitis, too, can survive in water. But the mere presence of a microorganism in water does not necessarily lead to the spread of disease. People have swum in water polluted with *Salmonella typhi* without getting typhoid, while others eating shellfish from the same water have developed the disease. The same may be true of hepatitis; the eating of clams from infected water has often caused the disease, whereas swimming has not proved to be a hazard. Shellfish concentrate germs in their tissues, which is probably why they can transmit diseases.

Water is carried to human beings in pipes for drinking, but it is also carried away in pipes as sewage. When there are defects in the systems, water may pass from one to the other, and microorganisms from sewage may get into drinking water. Wells and other sources of water can be contaminated by badly sited septic tanks, manure heaps, or garbage dumps. If such water is properly treated, it will be safe to drink; if it is drunk while untreated, disease is sure to follow.

The soil of gardens and farmlands can harbour disease organisms, as can food. Smoke, fog, and tobacco can affect the human respiratory tract and render it more vulnerable to disease. Air is probably the most common source of infectious agents. The whole of the environment is, in fact, alive with organisms with which human beings compete for survival.

BIBLIOGRAPHY. PAUL D. HOEPRICH (ed.), *Infectious Diseases,* 3rd ed. (1983); and RALPH D. FEIGIN and JAMES D. CHERRY (eds.), *Textbook of Pediatric Infectious Diseases,* 2 vol. (1981), are advanced postgraduate textbooks containing material to which the general reader might refer. ANDRÉ SIEGFRIED, *Routes of Contagion* (1965; U.K. title, *Germs and Ideas: Routes of Epidemics and Ideologies,* 1965; originally published in French, 1960); and HENRY E. SIGERIST, *Civilization and Disease* (1943, reissued 1970), deal with the effect of disease on human life and history; both are nontechnical. FRANK M. BURNET, *The Background of Infectious Diseases in Man* (1946), outlines the problems presented by infectious diseases in the middle of the 20th century; the fact that many of the problems have since been solved adds to the interest of the book. CLIFFORD HORTON-SMITH (ed.), *Biological Aspects of the Transmission of Disease* (1957), deals with the mechanics and interrelationships of infection in plants, animals, and humans; the level of writing varies from the very technical to the semipopular. WILLIAM H. MCNEILL, *Plagues and Peoples* (1976), a book for the general reader, describes the dramatic impact of infectious diseases on the rise and fall of civilizations. J.F.D. SHREWSBURY, *A History of Bubonic Plague in the British Isles* (1970), gives a scholarly but highly readable description of the plague. PAUL CASSAR, *Medical History of Malta* (1964), deals with the problems of epidemic disease, in war and peace, throughout the history of the island. KENNETH MELLANBY, *Pesticides and Pollution* (1967), reviews the problems presented by pollution in the modern world, many of which derive from attempts to control the spread of disease in humans and nature. JAMES M. ALSTON, *A New Look at Infectious Disease* (1967), shows briefly how, as one problem in the field of infectious disease is solved, another takes its place. J.A. BOYCOTT, *Natural History of Infectious Disease* (1971), discusses infectious disease as just one aspect of parasitism—a study, in simple terms, of host–parasite relationships. ANDREW B. CHRISTIE, *Infectious Diseases: With Chapters on Venereal Diseases,* 5th ed. (1968), gives a simple account of the clinical and social aspects of infectious diseases, and *Infectious Diseases: Epidemiology and Clinical Practice,* 3rd ed. (1980), is an advanced textbook, but with much about the nature of infection that the nontechnical reader can find useful. ANDREW B. and MARY C. CHRISTIE, *Food Hygiene and Food Hazards . . .,* 2nd ed. (1977), discusses nontechnically the problems of foodborne infections. FRANK L. HORSFALL and I. TAMM (eds.), *Viral and Rickettsial Infections of Man,* 4th ed. (1965), is an advanced postgraduate textbook but, especially on epidemiology, contains material to which the general reader might refer.

(A.B.C./R.D.F./Re.G.)

Information Processing and Information Systems

In popular usage, the term information refers to facts and opinions provided and received during the course of daily life: one obtains information directly from other living beings, from mass media, from electronic data banks, and from all sorts of observable phenomena in the surrounding environment. As a person uses such facts and opinions, he generates information of his own, some of which is communicated to others during discourse, by instructions, in letters and documents, and through other media. Information organized according to some logical relationships is referred to as a body of knowledge, to be acquired by systematic exposure or study. Application of knowledge (or skills) yields expertise, and additional analytical or experiential insights are said to constitute instances of wisdom. Use of the term information is not restricted exclusively to its communication via natural language. Information is also registered and communicated through art and by facial expressions and gestures or by such other physical responses as shivering. Moreover, every living entity is endowed with information in the form of a genetic code. These information phenomena permeate the physical and mental world, and their variety is such that it has defied so far all attempts at a unified definition of information.

Interest in information phenomena has increased dramatically in the 20th century, and today they are the objects of study in a number of disciplines, including philosophy, physics, biology, linguistics, information and computer science, electronic and communications engineering, management science, and the social sciences. On the commercial side, the information service industry has become one of the newer industries throughout the world. Almost all other industries—manufacturing and service—are increasingly concerned with information and its handling. The different, though often overlapping, viewpoints and phenomena of these fields lead to different (and sometimes conflicting) concepts and "definitions" of information.

This article touches on such concepts, particularly as they relate to information processing and information systems. In treating the basic elements of information processing, it distinguishes between information in analog and digital

form, and it describes their acquisition, recording, organization, retrieval, display, and dissemination. In treating information systems, the article discusses system analysis and design and provides a descriptive taxonomy of the main system types. Some attention is also given to the social impact of information systems and to the field of information science.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 735.

This article is divided into the following sections:

General considerations 553
  Basic concepts 553
  Information as a resource and commodity 553
Elements of information processing 554
  Acquisition and recording of information in analog form 554
  Acquisition and recording of information in digital form 555
    Recording media
    Recording techniques
  Inventory of recorded information 556
  Organization and retrieval of information 557
    Content analysis of analog-form records
    Content analysis of digital-form information
    Storage structures for digital-form information
    Query languages
  Information display 560
    Video
    Print
    Voice
  Dissemination of information 560
Information systems 561
  Impact of information technology 561
  Analysis and design of information systems 562
  Categories of information systems 563
    Management-oriented information systems
    Administration-oriented information systems
    Service-oriented information systems
    Public information utilities
  Impact of computer-based information systems on society 565
    Effects on the economy
    Effects on governance and management
    Effects on the individual
Information science 566
  The emergence of information science 566
  Subjects of research 567
  Professional information associations 567
Bibliography 567

General considerations

BASIC CONCEPTS

Interest in how information is communicated and how its carriers convey meaning has occupied, since the time of pre-Socratic philosophers, the field of inquiry called semiotics, the study of signs and sign phenomena. Signs are the irreducible elements of communication and the carriers of meaning. The American philosopher, mathematician, and physicist Charles S. Peirce is credited with having pointed out the three dimensions of signs, which are concerned with, respectively, the body or medium of the sign, the object that the sign designates, and the interpretant or interpretation of the sign. Peirce recognized that the fundamental relations of information are essentially triadic; in contrast, all relations of the physical sciences are reducible to dyadic (binary) relations. Another American philosopher, Charles W. Morris, designated these three sign dimensions syntactic, semantic, and pragmatic, the names by which they are known today.

Signs and their dimensions

Information processes are executed by information processors. For a given information processor, whether physical or biologic, a token is an object, devoid of meaning, that the processor recognizes as being totally different from other tokens. A group of such unique tokens recognized by a processor constitutes its basic "alphabet"; for example, the dot, dash, and space constitute the basic token alphabet of a Morse-code processor. Objects that carry meaning are represented by patterns of tokens called symbols. The latter combine to form symbolic expressions that constitute inputs to or outputs from information processes and are stored in the processor memory (as data and programs).

Information processors are components of an information system, which can be defined as a class of constructs. An abstract model of an information system features four basic elements: processor, memory, receptor, and effector (Figure 1). The processor has several functions: these are (1) to carry out elementary information processes on symbolic expressions, (2) to store temporarily in the processor's short-term memory the input and output expressions on which these processes operate and which they generate, (3) to schedule execution of these processes, and (4) to change this sequence of operations arbitrarily in accordance with the contents of the short-term memory. The memory serves to store symbolic expressions, including those that represent composite information processes called programs. The two other components, the receptor and the effector, are input and output mechanisms whose functions are, respectively, to receive symbolic expressions or stimuli from the external environment for manipulation by the processor and to emit the processed structures back to the environment. This abstract model of an information-processing system is representative of a broad variety of such systems, both natural and man-made.

Basic elements of information systems

The power of information processors is provided by their ability to carry out a small number of elementary information processes that operate on symbolic expressions. These processes are reading symbolic expressions (or generating representations of external stimuli by assigning them internal symbolic expressions in the processor memory), testing and comparing symbolic expressions by ascertaining whether tokens are part of the same symbol, creating and naming new, and modifying existing, symbolic expressions, copying and storing such expressions, and writing symbolic expressions (or transforming internal symbolic expressions into stimuli intended for the outside environment).

INFORMATION AS A RESOURCE AND COMMODITY

In the second half of the 20th century, information has acquired two major utilitarian connotations. On the one hand, it is considered an economic resource, somewhat on par with other resources such as labour, material, and capital. This view of information stems from the belief and evidence thereof that the possession, manipulation, and use of information can increase the cost-effectiveness of many physical and cognitive processes. The rise in information-processing activities in industrial manufacturing as well as in human problem solving has been remarkable. Analysis of one of the three traditional divisions of the economy, the service sector, shows a sharp increase in information-intensive activities since the beginning of the 20th century. By 1975 these activities accounted for half of the labour force of the United States (see Table), giving rise to the so-called information society.

As an individual and societal resource, information has some interesting characteristics that separate it from the traditional notions of economic resources. Unlike other resources, information is expansive with limits apparently imposed only by time and human cognitive capabilities.

A. Newell and H.A. Simon, *Human Problem Solving,* p. 20, © 1972, reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, N.J.

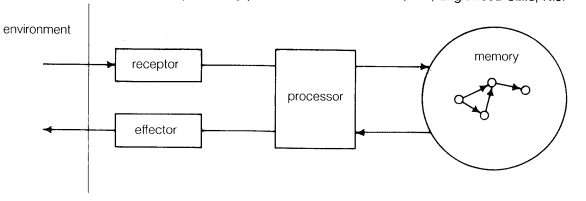


Figure 1: Structure of an information system.

Labour Distribution (%) in the United States, 1880–2000

	1880	1920	1955	1975	2000 (est.)
Agriculture and extractive	50	28	14	4	2
Manufacturing, commerce, industry	36	53	37	29	22
Information, knowledge, education	2	9	29	50	66
Other services	12	10	20	17	10

Source: Graham T.T. Molitor, "The Coming of the Information Society," *Communications Tomorrow.*

nforma-
ion as an
xpansive
esource

Its expansiveness is attributable to three reasons: (1) it is naturally diffusive; (2) it reproduces rather than is consumed through use; and (3) it can only be shared, not exchanged in transactions. At the same time, information is compressible, both syntactically and semantically. Coupled with its ability to be substituted for other economic resources, its transportability at very high speeds, and its ability to impart advantages to the beholder of information, these characteristics are at the base of such societal industries as research, education, publishing, marketing, and even politics. Societal concern with the husbanding of information resources has extended from the traditional domain of libraries and archives to encompass organizational, institutional, and governmental information under the umbrella of information resource management.

The second perception of information is that it is an economic commodity, which helps to stimulate the worldwide growth of a new segment of national economies—the information service sector. Taking advantage of the properties of information and building on the perception of its individual and societal utility and value, this sector provides a broad range of information products and services: it reduces information to manageable quantities, protects information and intellectual property rights, replaces labour with automata, replicates rare expertise through computer software, and provides consultancy in many walks of life.

mportance
of the in-
formation
service
sector

Elements of information processing

Humans receive information with their senses: sounds through hearing; images and text through sight; shape, temperature, and affection through touch; and odours through smell. To interpret the signals received from the senses, humans have developed and learned complex systems of languages consisting of "alphabets" of symbols and stimuli and the associated rules of usage. This has enabled them to recognize the objects they see, understand the messages they read or hear, and comprehend the signs received through the tactile and olfactory senses.

The carriers of information-conveying signs received by the senses are energy phenomena—audio waves, light waves, and chemical and electrochemical stimuli. In engineering parlance, humans are receptors of analog signals; and by a somewhat loose convention, the messages conveyed via these carriers are called analog-form information, or simply analog information. Until the development of the digital computer, cognitive information was stored and processed only in analog form, basically through the technologies of printing, photography, and telephony.

Although humans are very adept at processing information stored in their memories, analog information stored external to the mind is not processed easily. Modern information technology greatly facilitates the manipulation of externally stored information as a result of its representation as digital signals—*i.e.*, as the presence or absence of energy (electricity, light, or magnetism). Information represented digitally in two-state, or binary, form is often referred to as digital information. Modern information systems are characterized by extensive transformations between different types of analog and digital information.

Information in binary form

ACQUISITION AND RECORDING OF INFORMATION IN ANALOG FORM

The principal categories of information sources useful in modern information systems are text, video, and voice. One of the first ways in which prehistoric humans communicated was by sound; sounds represented concepts such as pleasure, anger, and fear, as well as objects of the surrounding environment, including food and tools. Sounds assumed their meaning by convention—namely, by the use to which they were consistently put. Combining parts of sound allowed representation of more complex concepts, gradually leading to the development of speech and eventually to spoken "natural" languages.

Development of writing systems

For information to be communicated broadly, it needs to be stored external to human memory; accumulation of human experience, knowledge, and learning would be severely limited without such storage. Storage of information external to memory made necessary the development of writing systems.

Civilization can be traced to the time when humans began to associate abstract shape with concepts and with the sounds of speech that represented them. Early recorded representations were those of visually perceived objects and events, as, for example, the animals and activities depicted in Paleolithic cave drawings. The evolution of writing systems proceeded through the early development of pictographic languages, in which a symbol would represent an entire concept. Such symbols would go through many metamorphoses of shape in which the resemblance between each symbol and the object it stood for gradually disappeared, but its semantic meaning would become more precise. As the conceptual world of humankind became larger, the symbols, called ideographs, grew in number. Modern Chinese, a present-day result of this evolutionary direction of a pictographic writing system, has upward of 50,000 ideographs.

At some point in the evolution of written languages the method of representation shifted from the pictographic to the phonetic: speech sounds began to be represented by an alphabet of graphic symbols. Combinations of a relatively small set of such symbols could stand for more complex concepts as words, phrases, and sentences. The invention of the written phonetic alphabet is thought to have taken place in Sumer, a civilization in the southern part of the Tigris-Euphrates Valley, during the 4th millennium BC.

The pragmatic advantages of alphabetic writing systems over the pictographic became apparent twice in the present millennium: after the invention of the movable-type printing press in the 15th century and again with the development of information processing by electronic means since the mid-1940s.

Recording media and techniques

From the time early man learned to represent concepts symbolically, he used whatever materials were readily available in nature for recording them. The Sumerian cuneiform, a wedge-shaped alphabet, was impressed by a stylus into soft clay tablets, which were subsequently hardened by drying in the sun or the oven. The earliest Chinese writing, dating to the 2nd millennium BC, is preserved on animal bone and shell, while early Indian writing was done on palm leaves and birch bark. Applications of technology yielded other materials for writing. The Chinese had recorded their pictographs on silk, using brushes made from animal hair, long before they invented paper. The Egyptians first wrote on cotton, but they began using papyrus sheets and rolls made from the fibrous lining of the papyrus plant during the 4th millennium BC. The reed brush and a palette of ink were the implements with which they wrote hieroglyphic script. Writing on parchment, a material which was superior to papyrus and was made from the prepared skins of animals, became commonplace around 200 BC, some 300 years after its first recorded use, and the quill pen replaced the reed brush. By the 4th century AD, parchment came to be the principal writing material in Europe.

Paper was invented in China at the beginning of the 2nd century AD, and for some 600 years its use was confined to the Far East. In AD 751 Arab and Chinese armies clashed at the Battle of Talas, near Samarkand; among the Chinese taken captive were some papermakers from whom the Arabs learned the techniques. From the 7th century on, paper became the dominant writing material of the Islāmic world. Papermaking finally reached Spain and Sicily in the 12th century—1,000 years after its in-

vention—and it took another three centuries before it was practiced in Germany.

With the invention of printing from movable type, typesetting became the standard method of creating copy. Typesetting was an entirely manual operation until the adoption of a typewriter-like keyboard in the 19th century. In fact, it was the typewriter that mechanized the process of recording original text. Although the typewriter was invented during the early 18th century in England, the first practical version, constructed by the American inventor Christopher Latham Sholes, did not appear until 1867. The mechanical typewriter finally found wide use after World War I. Today, its electronic variant, the computer video terminal, is used increasingly to record original text.

Development of photography

Recording of original non-textual (image) information was a manual process until the development of photography during the early decades of the 19th century; drawing and carving were the principal early means of recording graphics. Other techniques were developed alongside printing—for example, etching in stone and metal. The invention of film and the photographic process added a new dimension to information acquisition: for the first time, complex visual images of the real world could be captured accurately. Photography provided a method of storing information in less space and more accurately than was hitherto possible with narrative information.

During the 20th century, versatile electromagnetic media have opened up new possibilities for capturing original analog information. Magnetic audio tape is used to capture speech and music, and magnetic videotape provides a low-cost medium for recording analog voice and video signals directly and simultaneously. Such tape information can be transferred to an analog optical disc, a medium distinguished for its higher storage capacity and fidelity. Analog optical discs are employed in archival storage systems and other applications in which information is displayed without further processing.

Magnetic technology has other uses in the direct recording of analog information, including alphanumerics. Magnetic characters, bar codes, and special marks are printed on checks, labels, and forms for subsequent sensing by magnetic or optical readers and conversion to digital form. Banks, educational institutions, and the retail industry rely heavily on this technology. Nonetheless, paper and film continue to be the dominant media for direct storage of textual and visual information in analog form.

ACQUISITION AND RECORDING OF INFORMATION IN DIGITAL FORM

The versatility of modern information systems stems from their ability to represent information electronically as digital signals and to manipulate it automatically at exceedingly high speeds. Information is stored in binary devices, which are the basic components of digital technology. Because these devices exist only in one of two states, information is represented in them either as the absence or the presence of energy (electric pulse). The two states of binary devices are conveniently designated by the binary digits, or bits, zero (0) and one (1).

Coding systems

In this manner, alphabetic symbols of natural-language writing systems can be represented digitally as combinations of zeros (no pulse) and ones (pulse). Tables of equivalences of alphanumeric characters and strings of binary digits are called coding systems, the counterpart of writing systems. A combination of three binary digits can represent up to eight such characters; one comprising four digits, up to 16 characters; and so on. The choice of a particular coding system depends on the size of the character set to be represented. Among widely used systems are the Binary-Coded Decimal (BCD), a four-digit code representing decimal integers; the International Telegraph Code No. 2 (a modern version of the Baudot code, known as the International Telegraph Code No. 1), a five-digit combination representing the telegraph keyboard; the American Standard Code for Information Interchange (ASCII), a seven- or eight-bit code representing the English alphabet, numerals, and certain special characters of the standard computer keyboard; and the corresponding eight-bit Extended Binary Coded Decimal Interchange Code (EBCDIC), used for computers produced by IBM (International Business Machines) and most compatible systems. The digital representation of a character by eight bits is called a byte.

The seven-bit ASCII code is capable of representing up to 128 alphanumeric and special characters—sufficient to accommodate the writing systems of many phonetic scripts, including Latin and Cyrillic. Some alphabetic scripts require more than seven bits; for example, the Arabic alphabet, also used in the Urdu and Persian languages, has 28 phonetic characters (as well as a number of vowels and diacritical marks), but each of these may have four shapes, depending on its position in the word.

For digital representation of nonalphabetic writing systems, even the eight-bit code accommodating 256 characters is inadequate. Some writing systems that use Chinese characters, for example, have more than 50,000 ideographs (the minimal standard font for the Hanzi system in Chinese and the kanji system in Japanese has about 7,000 ideographs). Digital representation of such scripts can be accomplished in three ways. One approach is to develop a phonetic character set; the Chinese Pinyin, the Korean Han'gŭl, and the Japanese kana phonetic schemes all have alphabetic sets similar in number to the Latin alphabet. As the use of phonetic alphabets in Oriental cultures is not yet widespread, this approach is not entirely satisfactory. A second technique is to decompose ideographs into a small number of elementary signs called strokes, the sum of which constitutes a shape-oriented, nonphonetic alphabet. Hanzi uses a five-stroke or an eight-stroke scheme to decompose ideographs, supplemented with coding of polysyllabic ideographs (words and phrases). This technique permits the use of the ASCII and EBCDIC codes. The third approach is to use more than eight bits to encode the large numbers of ideographs; for instance, two bytes can represent uniquely more than 65,000 ideographs.

Recording media. Punched cards and perforated paper tape were once widely used to store data in binary form. Today, they have been supplanted by media based on electromagnetic and electrooptic technologies except in a few special applications.

The earliest form of magnetic mass-memory storage was magnetic tape. This medium, developed in 1951, consists of a ribbon of plastic film coated with iron oxide. Data are recorded on the tape through selective changes in magnetization of its surface. Depending on the recording density, a reel of magnetic tape 2,400 feet long can hold more than 100,000,000 characters. Special-purpose, high-density (50,000 bits per inch) cassette tapes have a capacity of 5,000,000,000 characters; they can accommodate the storage needs of imaging applications. Tape-storage devices suffer from one shortcoming: relatively slow retrieval of data. Because data are recorded serially, it may at times be necessary to traverse the entire tape to locate the information sought.

Magnetic disks

Another type of magnetic storage medium, the magnetic disk, provides rapid, random access to data. This device, developed in 1962, consists of either an aluminum or plastic platen coated with a metallic material. Information is recorded on a disk by turning the charge of the reading/writing disk head on and off, which produces magnetic "dots" representing binary digits in circular tracks. A block of data on a given track can be accessed without having to pass over a large portion of its contents sequentially, as in the case of tape. Data retrieval time is thus reduced dramatically. Disk packs, collections of 10 or more disks connected by a spindle, are often used on larger computer systems. A single pack of this kind can store from 10,000,000 to several hundred million characters. Some of the largest computer systems use disk cartridges, which are more compact than disk packs. Because of their high storage capacity as well as ability to access data directly, magnetic disk systems have become the dominant medium for recording and storing computer data.

During the 1970s the so-called floppy disk—a small, flexible disk—was introduced for use in personal computers and other microcomputer systems. Compared to the storage capacity of the conventional "hard" disk, that of such a "soft" diskette is low—only about 1,000,000 characters.

Accessing data on the latter is also slower, but the retrieval time is adequate for most microcomputer applications.

Optical discs

An entirely different kind of recording and storage medium, the optical disc, became available during the early 1980s. The optical disc makes use of laser technology: digital data are recorded by burning a series of microscopic holes, or pits, with a laser beam into thin metallic film on the surface of a 4¾-inch (12-centimetre) plastic disc. In this way, information from magnetic tape is encoded on a master disc; subsequently, the master is replicated by a process called stamping. In the read mode, low-intensity laser light is reflected off the disc surface and is "read" by light-sensitive diodes. The radiant energy received by the diodes varies according to the presence of the pits, and this input is digitized by the diode circuits. The digital signals are then converted to analog information on a video screen or in printout form.

Optical discs fall into two classes: (1) prerecorded, exemplified by the compact disc-read only memory (CD-ROM), from which data can be read but not modified; and (2) writable/erasable, on which data can be both recorded and overwritten. The latter variety is the full functional equivalent of the magnetic disk. It employs magneto-optic technology: the erasable medium is magnetic. Information is recorded or erased by reversing the polarity of the material. Reading is accomplished optically.

Optical storage represents a considerable improvement over magnetic disks in terms of superior recording capability (it accommodates storage of voice, text, graphics, and video), of storage capacity (2,000,000,000 characters on a 12-inch disc), of ruggedness (the disc is sealed between two plastic surfaces), of portability, and of low cost—in fact, the lowest of per-bit storage of any digital storage media. In their present stage of development, optical discs still have some shortcomings, particularly a relatively slow rate of data access. Technological advances promise improvements, however. For example, the development of a flexible storage medium, a dye polymer coated onto a polyester substrate, is expected to result in the production of flexible optical discs with storage capacities far exceeding those of the rigid discs currently in use.

During the 1990s new types of portable digital media are likely to find extensive application in information systems. Among the most promising are the digital audio tape, which has a storage capacity of more than 1,000,000,000 characters; and the so-called softstrip card, a paper strip with information encoded as highly condensed optical binary patterns and with a storage capacity of 45,000 characters per 8 × 11-inch page.

Recording techniques. Digitally stored information is commonly referred to as data, and its analog counterpart is called source data. The capture of source data for electronic storage and digital processing is accomplished by manual and automatic techniques. Manual data entry is the transcription, by means of a keyboard, of alphanumeric information from document manuscripts or forms. Keyboarding data is straightforward for alphabetic writing systems but unwieldy for pictographic data. Chinese, Japanese, Korean, and other languages that employ ideographs thus require the use of phonetic or shape-oriented alphabets for entering data. Manual data entry, a slow and error-prone process, is facilitated to a degree by special computer programs that include editing software with which to insert formatting commands, verify spelling, and make text changes and document-formatting software with which to arrange and rearrange text and graphics flexibly on the output page.

It is estimated that U.S. business deals annually with 400,000,000,000 paper documents, two-thirds of which cannot be digitized by keyboard transcription because they contain drawings, signatures, or images or because such transcription would be highly uneconomical. Such documents are digitized by automatic image-processing techniques. Image processing is the primary means by which large volumes of data, such as that handled by government agencies, is recorded in digital form.

Image processing

Image processing involves the use of high-resolution scanning cameras. To generate a digital representation of a document page, such image scanners divide it into minute areas called picture elements (or pixels), and produce an array of binary digits, each representing the brightness of a pixel. The resulting stream of bits is enhanced and compressed by a device called an image controller and is stored on a high-density storage medium such as optical disc or microfilm. (A large storage capacity is required, since it takes 45,000 bytes to store a typical document page of 2,500 characters.) A 12-inch optical disc can hold about 40,000 such pages. Compared with manual data entry, image scanning is rapid—up to 60 pages per minute. Aside from its application in electronic document filing, image scanning is used for the digital conversion of document pages for transmission via telecommunications systems and their reconversion into analog images (facsimile transmission), as well as in satellite photography.

An image scanner digitizes an entire document page for storage and display as an image and does not recognize characters and words of text. The stored material therefore cannot be linguistically manipulated by text processing and other software techniques. When such manipulation is desired, analog text is digitized by a method called optical character recognition. An optical character reader (OCR) converts each optically scanned character into an electric signal and compares it with internally stored representations of an alphabet of characters, so as to select from it the one that matches the scanned character most closely or to reject it as unidentifiable. OCR machines range from inexpensive, hand-held readers capable of recognizing dozens of standard fonts to sophisticated devices that can be taught to interpret the shape, size, and pitch of symbols, including handwriting.

Optical character recognition

In some applications, optical character recognition and image scanning are used jointly. The U.S. Patent and Trademark Office converts the text of patents into digital form by means of the former and digitizes drawings by means of the latter. The resulting digital data is stored on optical discs.

The digital recording of sound is important because speech is the most natural carrier of communicable information. This process is now commonplace. In the entertainment industry, analog audio signals are captured with microphones and transformed by analog-to-digital converters to produce recordings for the consumer market. In the communications industry, this same technique makes it possible to carry speech as digital data over voice communications systems, with attendant technical and economic benefits. When perfected, direct voice-to-printer technologies are expected to have a tremendous impact on methods of information acquisition and recording.

(V.Sl.)

INVENTORY OF RECORDED INFORMATION

The development of recording media and techniques enabled society to begin building a store of human knowledge and experiences. The idea of collecting and organizing written records is thought to have originated more or less simultaneously in two regions, Sumer and Egypt, about 5,000 years ago. Early collections of Sumerian and Egyptian writings, recorded in cuneiform on clay tablets and in hieroglyphic script on papyrus, contained information about legal and economic transactions—laws and treatises, contracts, property records, taxes, and wills. In these and other early document collections (*e.g.*, those of China produced during the Shang dynasty in the 2nd millennium BC and Buddhist collections in India dating to the 5th century BC), it is difficult to separate the concepts of the archive and the library.

Development of document collections

From the Middle East the concept of document collections penetrated the Greco-Roman world. Roman kings institutionalized the population and property census as early as the 6th century BC. The great Library of Alexandria, established in the 3rd century BC, is best known as a large collection of papyri containing inventories of property, taxes, and other payments by citizens to their rulers and to each other. It is, in short, the ancient equivalent of today's administrative information systems.

The scholarly splendour of the Islāmic world from the 8th to 13th centuries AD can in large part be attributed to the maintenance of public and private book libraries. The

Bayt al-Ḥikmah ("House of Wisdom"), founded in AD 830 in Baghdad, contained a public library with a large collection of materials on a wide range of subjects, and the 10th-century library of Caliph al-Ḥakam in Córdoba, Spain, boasted more than 400,000 books.

The late but rapid development of European libraries from the 16th century on followed the invention of printing from movable type, which spurred the growth of the printing and publishing industries. Since the beginning of the 17th century, literature has become the principal medium for disseminating knowledge in science and education. The phrase primary literature is used to designate original information in various printed formats: newspapers, monographs, conference proceedings, learned and trade journals, reports, patents, bulletins, and newsletters. The scholarly journal, the classic medium of scientific communication, first appeared in 1665. Three hundred years later the number of periodical titles published in the world was estimated at more than 60,000, reflecting not only growth in the number of practitioners of science and expansion of its body of knowledge through specialization but also a maturing of the system of rewards that motivates, and sometimes pressures, scientists to publish. To render this bulk of primary literature more manageable in storage and more affordable, a portion of it is also issued on microfilm. Other media containing scientific information in analog form—audio and video recording—abound as well.

Primary literature

The sheer quantity of printed information has for some time prevented any individual from fully absorbing even a minuscule fraction of it. Such devices as tables of contents, summaries, and indexes of various types, which aid in identifying and locating relevant information in primary literature, have been in use since the 16th century and led to the development of what is termed secondary literature during the 19th century. The purpose of secondary literature is to "filter" the primary information sources, usually by subject area, and provide the indicators to this literature in the form of reviews, abstracts, and indexes. Over the past 100 years there has evolved a system of disciplinary, national, and international abstracting and indexing services that acts as a gateway to several attributes of primary literature: authors, subjects, publishers, dates (and languages) of publication, and citations. The professional activity associated with these access-facilitating tools is called documentation.

Secondary literature

The quantity of printed materials also makes it impossible, as well as undesirable, for any institution to acquire and house more than a small portion of it. The husbanding of recorded information has become a matter of public policy, as many countries have established national libraries and archives to direct the orderly acquisition of analog-form documents and records. Since these institutions alone are not able to keep up with the output of such documents and records, new forms of cooperative planning and sharing recorded materials are evolving—namely, public and private, national and regional library networks and consortia.

The emergence of digital technology in the mid-20th century has affected humankind's inventory of recorded information dramatically. During the early 1960s computers were used to digitize text for the first time; the purpose was to reduce the cost and time required to publish two American abstracting journals, the *Index Medicus* of the National Library of Medicine and the *Scientific and Technical Aerospace Reports* of the National Aeronautics and Space Administration (NASA). By the late 1960s such bodies of digitized alphanumeric information, known as bibliographic and numeric data bases, constituted a new type of information resource. This resource is husbanded outside the traditional repositories of information (libraries and archives) by data-base "vendors." Advances in computer storage, telecommunications, software for computer sharing, and automated techniques of text indexing and searching fueled the development of an on-line data-base service industry. Meanwhile, electronic applications to bibliographic control in libraries and archives have led to the development of computerized catalogs and of union catalogs in library networks. They also have resulted in the introduction of comprehensive automation programs in these institutions.

Bibliographic and numeric data bases

The vast inventory of recorded information can be useful only if it is systematically organized and if mechanisms exist for locating in it items relevant to human needs. The main approaches for achieving such organization are reviewed in the following section, as are the tools used to retrieve desired information. (Y.-L.W./V.Sl.)

ORGANIZATION AND RETRIEVAL OF INFORMATION

In any collection, physical objects are related by order. The ordering may be random or by some characteristic called a key. Such characteristics may be intrinsic properties of the objects (*e.g.,* size, weight, shape, or colour), or they may be assigned from some agreed-upon set, such as object class or date of purchase. The values of the key are arranged in a sorting sequence that is dependent on the type of key involved: alphanumeric key values are usually sorted in alphabetic sequence, while other types may be sorted on the basis of similarity in class, such as books on a particular subject or flora of the same genus.

In most cases, order is imposed on a set of information objects for two reasons: to create their inventory and to facilitate locating specific objects in the set. There also exist other, secondary objectives for selecting a particular ordering, as, for example, conservation of space or economy of effort in fetching objects. Unless the objects in a collection are replicated, any ordering scheme is one-dimensional and invariably unable to meet all the functions of ordering with equal effectiveness. The main approach for overcoming some of the limitations of one-dimensional ordering of recorded information relies on extended description of its content and, for analog-form information, of some features of the physical items. This approach employs various tools of content analysis that subsequently facilitate accessing and searching recorded information.

Content analysis of analog-form records. The collections of libraries and archives, the primary repositories of analog-form information, constitute one-dimensional ordering of physical materials in print (documents), in image form (maps and photographs), or in audio-video format (recordings and video tapes). To break away from the confines of one-dimensional ordering, librarianship has developed an extensive set of attributes in terms of which it describes each item in the collection. The rules for assigning these attributes are called cataloging rules. Descriptive cataloging is the extraction of bibliographic elements (author names, title, publisher, date of publication, etc.) from each item; the assignment of subject categories or headings to such items is termed subject cataloging.

Descriptive and subject cataloging

Information on the same subject is located together in large libraries so as to minimize the physical effort of accessing such items and to make browsing possible. Consequently, librarianship has developed complex systems of classification by subject. The Dewey Decimal Classification, Universal Decimal Classification, and Library of Congress Classification are three widely used schemes for assigning every individual item a single classification number (which determines the location of the item on the shelves) and one or more subject headings (under which the item will appear in the library's subject catalog). The traditional card catalog is an alphabetic ordering of the items in a collection usually by three keys: author name, title, and subject matter.

Conceptually, the library catalog is a table or matrix in which each row describes a discrete physical item and each column provides values of the assigned key. When such a catalog is represented digitally in a computer, any attribute can serve as the ordering key. By sorting the catalog on different keys, it is possible to produce a variety of indexes as well as subject bibliographies. More importantly, any of the attributes of a computerized catalog becomes a search key (access point) to the collection, surpassing the utility of the card catalog.

The most useful access key to analog-form items is subject. The extensive lists of subject headings of library classification schemes provide, however, only a gross access tool to the content of the items. A technique called index-

Indexing and abstracting

ing provides a refinement over library subject headings. It consists of extracting from the item or assigning to it subject and other "descriptors"—words or phrases denoting significant concepts (topics, names) that occur in or characterize the content of the record. Indexing frequently accompanies abstracting, a technique for condensing the full text of a document into a short summary that contains its main ideas (but invariably incurs an information loss and often introduces a bias). Computer-printed, indexed abstracting journals provide a means of keeping informed of primary information sources.

Content analysis of digital-form information. Digitally stored data may represent alphanumeric, image, or audio information, each requiring different techniques for analyzing content. The objectives of content analysis of alphanumeric information are twofold. At the simple level the aim is to describe the record (document) in terms of some of its properties or characteristics so that the record can be located or assigned to a category of similar records. At a more complex level the objective is to represent the meaning or portent of the document so that its content can be "understood" and possibly further manipulated by machine in a cognitive sense.

The primary method for describing alphanumeric information is indexing. Although a document may be described in various ways (*e.g.,* indirectly by the citations of references contained in it), indexing typically provides pointers to the subject of documents. Whether carried out manually by humans or automatically by computers, indexing uses two main approaches: the assignment of subject descriptors from an unlimited vocabulary (free indexing), or their assignment from a list of authorized descriptors (controlled indexing). A collection of authorized descriptors is called an authority list or, if it also displays various relationships among descriptors such as hierarchy or synonymy, a thesaurus. The result of the indexing process is a computer file known as an inverted index, which is an alphabetic listing of descriptors and the addresses of their occurrence in the document body.

Full-text indexing, the use of every character string (word of a natural language) in the text as an index term, is an extreme case of free-text indexing: each word in the document (except function words such as articles and prepositions) becomes an access point to it. Used earlier for the generation of concordances in literary analysis and other computer applications in the humanities, full-text indexing placed great demands on computer storage because the resulting index is at least as large as the body of the text. With decreasing cost of mass storage, automatic full-text indexing capability has been incorporated routinely into state-of-the-art information-management software.

Machine indexing

Free, controlled, and full-text indexing affect not only the economics of information services but also to some extent the specificity and exhaustiveness of computerized information retrieval. This fact notwithstanding, information technology forces indexing toward complete automation and, in some applications, toward extinction. When computer hardware is very fast, as is the case with array processors, the digital information store may be searched serially in its entirety for each query without ever creating an index. Full-text searching methods for the new type of computer architecture called parallel processing may also dispense with the index. (Parallel processing, designed for faster and more efficient performance, involves the simultaneous execution of many operations by a large number of integrated circuits in which millions of logic, memory, and input/output circuits work together.) In contrast, information stored on slow-access devices such as the optical disc will continue to depend on the index as a means of retrieving portions of the data base relevant to the query. The index, however, has to be computer-generated. In either case, digital technology reveals the granularity of human information store to a finer degree than ever before.

Semantic content analysis

The analysis of digitally recorded natural-language information from the semantic viewpoint is a matter of considerable complexity, and it lies at the foundation of such potential applications as automatic question answering from a data base or retrieval by means of unrestricted natural-language queries (see below). The general approach has been that of computational linguistics: to derive representations of the syntactic and semantic relations among the linguistic elements of sentences and larger parts of the document. Syntactic relations are described by parsing (decomposing) the grammar of sentences (Figure 2). For semantic representation, three related formalisms dominate. In a so-called semantic network, conceptual entities such as objects, actions, or events are represented as a graph of linked nodes (Figure 3). "Frames" represent, in a similar graph network, physical or abstract attributes of objects and in a sense define the objects. In "scripts," events and actions rather than objects are defined in terms of their attributes.

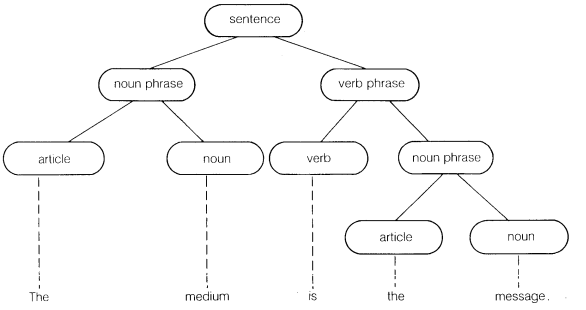


Figure 2: A parsing graph.

Image processing

The content analysis of images is accomplished by two primary methods: image processing and pattern recognition. Image processing is a set of computational techniques for analyzing, enhancing, compressing, and reconstructing images. Pattern recognition is an information-reduction process: the assignment of visual or logical patterns to classes based on the features of these patterns and their relationships. The stages in pattern recognition involve measurement of the object to identify distinguishing attributes, extraction of features for the defining attributes, and assignment of the object to a class based on these features. Both image processing and pattern recognition have extensive applications in various areas, including astronomy, medicine, industrial robotics, and remote sensing by satellites.

C.S. Shapiro, "Natural Language Processing," from A. Ralston and E.D. Reilly, Jr. (eds.), *Encyclopedia of Computer Science and Engineering,* Van Nostrand Reinhold, 1983

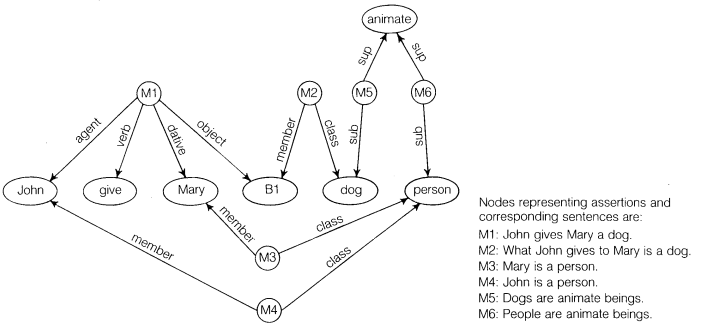


Figure 3: A semantic network representation.

Speech processing

The immediate objective of content analysis of digital speech is the conversion of discrete sound elements into their alphanumeric equivalents. Once so represented, speech can be subjected to the same techniques of content analysis as natural-language text—*i.e.,* indexing and linguistic analysis. Converting speech elements into their alphanumeric counterparts is an intriguing problem because the "shape" of speech sounds embodies a wide range of many acoustic characteristics and because the linguistic elements of speech are not clearly distinguishable from one another. The general technique used in speech processing is to classify the spectral representations of sound and to match the resulting digital spectrographs against prestored "templates" so as to identify the alphanumeric equivalent of the sound. (The obverse of this technique, the digital-to-analog conversion of such templates into sound, is a relatively straightforward approach to generating synthetic speech.)

Speech processing is complex as well as expensive in

terms of storage capacity and computational requirements. State-of-the-art speech recognition systems can identify limited vocabularies and parts of distinctly spoken speech and can be programmed to recognize tonal idiosyncracies of individual speakers. When more robust and reliable techniques become available and the process is made computationally tractable (as is expected with parallel computers), humans will be able to interact with computers via spoken commands and queries on a routine basis. In many situations this may make the keyboard obsolete as a data-entry device.

Storage structures for digital-form information. Digital information is stored in complex patterns that make it feasible to address and operate on even the smallest element of symbolic expression, as well as on larger strings such as words or sentences and on images and sound.

Data set

The sum of electronic representations, or data, in a modern information system is called its data set. Each data set typically consists of one or more data bases; the latter are cross-referenced sets of files, each comprising records consisting of fields. Data sets are organized in various types of data structures that facilitate the creation, accessing, and deletion of data and optimize the operating features of the storage media.

From the viewpoint of digital information storage, it is useful to distinguish between "structured" data, such as inventories of objects that can be represented by short symbol strings and numbers, and "unstructured" data, such as the natural-language text of documents or pictorial images. The principal objective of all storage structures is to facilitate the processing of data elements based on their relationships; the structures thus vary with the type of relationship they represent. The choice of a particular storage structure is governed by the relevance of the relationships it allows to be represented to the information-processing requirements of the task or system at hand.

In information systems whose store consists of unstructured data bases of natural-language records, the objective is to retrieve records (or portions thereof) based on the presence in the records of words or short phrases that constitute the query. Since there exists an index as a separate file that provides information about the locations of words and phrases in the data-base records, the relationships that are of interest (*e.g.,* word adjacency) can be calculated from the index. Consequently, the data-base text itself can be stored as a simple ordered sequential file of records. The majority of the computations use the index, and they access the text file only to pull out the records or those portions that satisfy the result of the computations. The sequential file structure is used with most commercial document-retrieval systems, including the ones containing digitized full text.

When relationships among data elements need to be represented as part of the records so as to make more efficient the desired operations on these records, two types

File structures

of "chained" structures are commonly used: hierarchical and network. In the hierarchical file structure, records are arranged in a scheme resembling a family tree, with records related to one another from top to bottom (Figure 4A). In the network file structure, records are arranged in groupings known as sets; these can be connected in any number of ways (Figure 4B), giving rise to considerable flexibility. In both hierarchical and network structures, the relationships are shown by means of "pointers" (*i.e.,* identifiers such as addresses or keys) that become part of the records. Large numbers of pointers complicate the storage structure but facilitate navigating through it to locate related records and fields. The efficiency of information retrieval and processing in hierarchical and network storage structures depends on the skill of the system designer to identify a priori the multiplicity of relationships among records and fields so that they can be represented by pointers. When a query assumes relationships that are not represented explicitly, retrieval becomes considerably more difficult—a matter of writing special programs.

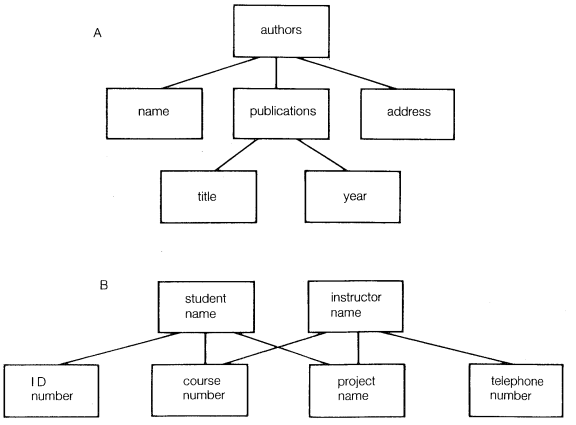


Figure 4: Examples of hierarchical (A) and network (B) file structures.

Another type of data-base storage structure, the relational structure, has become very popular since the 1970s. Its major advantage over the hierarchical and network structures is the ability to handle unanticipated data relationships without pointers. Relational storage structures are two-dimensional tables consisting of rows and columns, much like the conceptual library catalog mentioned above. The elegance of the relational model lies in its conceptual simplicity, availability of theoretical underpinnings (relational algebra), and the ability of its associated software to handle data relationships without the use of pointers. Like sequential files, relational data bases are supported by inverted indexes that improve their retrieval efficiency. The relational model is particularly suitable for data bases containing highly structured information.

The feasibility of storing large volumes of full text on an economic medium (the digital optical disc) has renewed interest in the study of storage structures that permit more powerful retrieval and processing techniques to operate on cognitive entities other than words, to facilitate more extensive semantic content and context analysis, and to organize text conceptually into logical units rather than those dictated by printing conventions.

Query languages. The uses of data bases are manifold. They provide a means of retrieving records or parts of records and performing various calculations before displaying the results. The interface by which such manipulations are specified is called the query language. Whereas early query languages were complex to a point that interacting with electronic data bases could only be done by specially trained individuals, recent interfaces allow "friendly" access to data-base information by casual users.

Types of query modes

The main types of popular query modes are the "menu," the "fill-in-the-blank" technique, and the structured query. Particularly suited for novices, the menu requires a person to choose from several alternatives displayed on the terminal video screen. The method is applicable for systems in which choices are arranged into hierarchies and is not suitable for key-word queries (though it may be combined with them). The fill-in-the-blank technique is one in which the user is prompted to enter key words as search statements. Employed in information-retrieval systems, it permits the user to expand or narrow down search statements by means of logical operators (conjunction and disjunction) and other tools. Effective use of these tools requires experience. To make the fill-in-the-blank technique more accessible to the casual user, it is often combined with friendly, menu-based interface software.

The structured query approach is effective with relational data bases. It has a formal, powerful syntax that is in fact a programming language, and it is able to accommodate logical operators. One implementation of this approach, the Structured Query Language (SQL), has the form

select [field Fa, Fb, . . . Fn] *from* [data base Da, Db, . . . Dn]
where [field Fa = abc] *and* [field Fb = def]

Structured query languages support data-base searching as well as other operations by using commands such as "find," "delete," "print," "sum," and so forth. The sentencelike structure of an SQL query resembles natural language except that its syntax is limited and fixed. Instead of using an SQL statement, it is possible to represent queries

in tabular form. The technique, referred to as query-by-example (or QBE), displays an empty tabular form and expects the searcher to enter the search specifications into appropriate columns. The program then constructs an SQL-type query from the table and executes it.

The most flexible query language is of course natural language. The use of natural-language sentences in a constrained form to search data bases is allowed by some commercial data-base management software. These programs parse the syntax of the query; recognize its action words and their synonyms; identify the names of files, records, and fields; and perform the logical operations required. Experimental systems that accept such natural-language queries in spoken voice have been developed; however, the ability to employ unrestricted natural language to query unstructured information will require further advances in machine understanding of natural language, particularly in techniques of representing the semantic and pragmatic context of ideas. The prospect of an intelligent conversation between humans and a large store of digitally encoded knowledge is not imminent.

INFORMATION DISPLAY

For humans to perceive and understand information, it must be presented as print and image on paper; as print and image on film or on a video terminal; as sound via radio or telephony; as print, sound, and video in motion pictures, on television broadcasts, or at lectures and conferences; or in face-to-face encounters. Except for live encounters and audio information, such displays emanate increasingly from digitally stored data, with the output media being video, print, and sound.

Video. Possibly the most widely used video display device, at least in the industrialized world, is the television set. Designed primarily for video and sound, its image resolution is inadequate for alphanumeric data except in relatively small amounts. Use of the television set in text-oriented information systems has been limited to menu-oriented applications such as videotex, in which information is selected from hierarchically arranged menus (with the aid of a numeric keyboard attachment) and displayed in fixed frames. The television, computer, and communications technologies are, however, converging in a high-resolution digital television set capable of receiving alphanumeric, video, and audio signals.

Computer video terminals

The computer video terminal is today's ubiquitous interface that transforms computer-stored data into analog form for human viewing. The two basic apparatuses used are the cathode-ray tube (CRT) and the more recent flat-panel display. In CRT displays an electron gun emits beams of electrons on a phosphorus-coated surface; the beams are deflected, forming visible patterns representative of data. Flat-panel displays use one of four different media for visual representation of data: liquid crystal, light-emitting diodes, plasma panels, and electroluminescence. Advanced video display systems enable the user to scroll, page, zoom (change the scale of the details of the display image for enhancement), divide the screen into multiple colours and "windows" (viewing areas), and in some cases even activate commands by touching the screen instead of using the keyboard.

The information capacity of the terminal screen depends on its resolution, which ranges from low (character-addressable) to high (bit-addressable). High resolution is indispensable for the display of graphic and video data in state-of-the-art workstations, such as those used in engineering or information systems design.

Print. Modern society continues to be dominated by printed information. The convenience and portability of print on paper make it difficult to imagine the paperless world that some have predicted. The generation of paper print has changed considerably, however. Although manual typesetting is still practiced for artwork, in special situations, and in some developing countries, electronic means of composing pages for subsequent reproduction by photoduplication and other methods has become commonplace.

Since the 1960s, volume publishing has become an automated process using large computers and high-speed printers to transfer digitally stored data on paper. The appearance of microcomputer-based publishing systems has proved to be another significant advance. Economical enough to allow even small organizations to become in-house publishers, these so-called desktop publishing systems are able to format text and graphics interactively on a high-resolution video screen with the aid of page-description command languages. Once a page has been formatted, the entire image is transferred to an electronic printing or photocomposition device.

Microcomputer-based publishing systems

Printers. Computer printers are commonly divided into two general classes according to the way they produce images on paper: impact and nonimpact. In the first type, images are formed by the print mechanism making contact with the paper through an ink-coated ribbon. The mechanism consists either of print hammers shaped like characters or of a print head containing a row of pins that produce a pattern of dots in the form of characters or other images.

Most nonimpact printers form images from a matrix of dots, but they employ different techniques for transferring images to paper. One major type, the laser printer, uses a beam of laser light and a system of optical components to etch images on a photoconductor drum from which they are carried via electrostatic photocopying to paper. Light-emitting diode (LED) printers resemble laser printers in operation but direct light from energized diodes rather than a laser onto a photoconductive surface. Ion-deposition printers make use of technology similar to that of photocopiers for producing electrostatic images. Another type of nonimpact printer, the ink-jet printer, sprays electrically charged drops of ink onto the print surface. An entirely different technique is used by thermal printers, which apply heat to a ribbon coated with a wax-based pigment in order to transfer the pigment to the printer paper.

Microfilm and microfiche. Alphanumeric and image information can be transferred from digital computer storage directly to film by the COM (Computer-Output-on-Microfilm) process. Reel microfilm, the first microform introduced before World War II, and microfiche, a flat sheet of film containing multiple microimages reduced from the original by a factor of 12 to 48, have become popular methods of document storage and reproduction. Ultrafilm, which reduces the image 200 times or more, is the highest density medium for analog-form information storage. Though portable and durable, microforms require optical reading equipment; the latter ranges from inexpensive personal viewers to computer-assisted microfilm retrieval systems capable of storing millions of pages.

Voice. In synthetic speech generation, digitally prestored sound elements are converted to analog sound signals and combined to form words and sentences. Digital-to-analog converters are available as inexpensive boards for microcomputers or as software for larger machines. Human speech is the most natural form of communication, and so applications of this technology are appearing in situations where there are numerous requests for specific information (*e.g.,* time, travel, and entertainment), where there is a need for repetitive instruction, in electronic voice mail (the counterpart of electronic text mail), and in toys.

Speech synthesis

DISSEMINATION OF INFORMATION

The process of recording information by handwriting was obviously laborious and required the dedication of the likes of Egyptian scribes or monks in monasteries around the world. It was only after mechanical means of reproducing writing were invented that information records could be duplicated more efficiently and economically.

The first practical method of reproducing writing mechanically was block printing; it was developed in China during the T'ang dynasty (618–907). Ideographic text and illustrations were engraved in wooden blocks, inked, and copied on paper. Used to produce books as well as cards, charms, and calendars, block printing spread to Korea and Japan but apparently not to the Islāmic or European Christian civilizations. European woodcuts and metal engravings date only to the 14th century.

Printing from movable type was also invented in China (in the mid-11th century AD). There and in the book-

making industry of Korea, where the method was applied more extensively during the 15th century, the ideographic type was made initially of baked clay and wood and later of metal. The large number of typefaces required for pictographic text composition continued to handicap printing in the Far East until the present time.

The invention of character-oriented printing from movable type (1440–50) is attributed to the German printer Johannes Gutenberg. Within 30 years of his invention, the movable-type printing press was in use throughout Europe. Character-type pieces were metallic and apparently cast from metallic molds; paper and vellum (calfskin parchment) were used to carry the impressions. Gutenberg's technique of assembling individual letters by hand was employed until 1886 when the German-born American printer Ottmar Mergenthaler developed the Linotype, a keyboard-driven device that cast lines of type automatically. Typesetting speed was further enhanced by the Monotype technique, in which a perforated paper ribbon, punched from a keyboard, was used to operate a typecasting machine. Mechanical methods of typesetting prevailed until the 1960s. Since that time they have been largely supplanted by the electronic and optical printing techniques described in the previous section.

Reproduction of graphics

Unlike the use of movable type for printing text, early graphics were reproduced from wood relief engravings in which the nonprinting portions of the image were cut away. Musical scores, on the other hand, were reproduced from etched stone plates. At the end of the 18th century the German printer Aloys Senefelder developed lithography, a planographic technique of transferring images from a specially prepared surface of stone. In offset lithography the image is transferred from zinc or aluminum plates instead of stone, and in photoengraving such plates are superimposed with film and then etched.

The first successful photographic process, the daguerreotype, was developed during the 1830s. The invention of photography, aside from providing a new medium for capturing still images and later video in analog form, was significant for two other reasons. First, recorded information (textual and graphic) could be easily reproduced from film and, second, the image could be enlarged or reduced. Document reproduction from film to film has been relatively unimportant because both printing and photocopying (see below) are cheaper. The ability to reduce images, however, has led to the development of the microform, the most economical method of disseminating analog-form information (see above).

Another technique of considerable commercial importance for the duplication of paper-based information is photocopying, or dry photography. Printing is most economical when large numbers of copies are required, but photocopying provides a fast and efficient means of duplicating records in small quantities for personal or local use. Of the several technologies that are in use, the most popular process, xerography, is based on electrostatics.

Methods and media of information dissemination continue to evolve. Today most forms of primary and secondary literature, from newspapers to scientific abstracting journals, are printed electronically. High-speed digital satellite communications facilitate electronic printing at remote sites; for example, the world's major newspapers transmit electronic page copies to different geographic locations for local printing and distribution. In business communications, documents and correspondence are generated via text-processing workstations and delivered electronically on an increasingly wider scale. Photocopying, facsimile transmission, and high-speed teleprinting are expected to be handled by a single device, a "universal" printer, in the near future.

Distribution of digital information by electronic publishing

While the volume of information issued in the form of printed matter continues unabated, the electronic publishing industry has begun to disseminate information in digital form. The development of the microcomputer has provided the main impetus. Not only is such a system almost as versatile and efficient as larger computers in information retrieval, but it carries out this operation relatively economically. Certain types of information media lend themselves particularly well to distribution in digital form. These include catalogs, handbooks, indexes, data bases, and reference materials designed to be consulted rather than read in toto. Computer software also is well suited for distribution via electronic publishing.

The digital optical disc (see above) has given electronic publishing its strongest boost to date for issuing large bodies of archival information: legislation, court and hospital records, patents, full-text data bases of periodicals, and libraries of computer software. The optical disc provides the mass production technology for publication in machine-readable form. Because of it, the prospect of having a vast electronic library in every school and at many professional workstations may very well be realized by the end of the 20th century.

The marriage of computers and digital telecommunications is also changing the modes of informal communication in society. Throughout history, these modes have consisted of various types of face-to-face meetings (conferences, seminars, workshops, exhibitions, and the like) and personal communication by telephone and mail. While direct encounters continue, they are supplemented and in some instances replaced by increasingly popular means of informal communication: electronic mail, a method of forwarding and storing messages between computer users; electronic bulletin, a technique of broadcasting newsworthy textual messages via computer to persons having similar interests; and electronic teleconferencing, a method of linking remote users in real time by either voice or voice-and-image communications. These technologies for informal communication are forging together virtual societal networks—communities of geographically dispersed individuals who have common professional or social interests.

Modes of informal communication

Information systems

The primary vehicles for the purposeful, orchestrated processing of information are information systems, constructs that collect, organize, store, process, and display information in all its forms (raw data, interpreted data, knowledge, and expertise) and formats (text, video, and voice). In principle, any record-keeping system—*e.g.*, an address book or a train schedule—may be regarded as an information system. What sets modern information systems apart is their electronic dimension, which permits extremely fast, automated manipulation of digitally stored data and their transformation from and to analog representation.

Evolution of electronic information systems

The evolution of modern information systems may be characterized by six stages. The first three—introduction, expansion, and control of computerized functions—reflect a technological emphasis. The last three stages—functional integration, allocation of responsibility over information, and system maturity—manifest a user-oriented focus. In industrialized countries most major information systems are presently in the fourth stage, integrating information horizontally into data bases. Some are entering the fifth stage in which organizations begin to treat information as a corporate resource and designate units and individuals as responsible data owners. The mature stage, characterized by the development of organizational planning models based on the information resource, may be attained by some organization by the end of the present century. Developing countries, by contrast, appear to be in transition from the second to the third stage.

IMPACT OF INFORMATION TECHNOLOGY

Electronic information systems are a phenomenon of the second half of the 20th century. Their evolution is closely tied with advances in the performance of four information technologies—namely, those involving digital processors, digital storage, communications, and software.

Since the invention of semiconductor devices known as transistors, improvement in the cost performance of digital processors has ranged from 15 to 25 percent annually, largely as a result of progress in integrated-circuit technology and its effects on the speed and reliability of digital devices. At one end, these advances produced the microcomputer; at the other end, they are leading to novel computer architectures that will enable machines to process data in a parallel fashion (see COMPUTERS: *Modern*

digital systems). The information-processing efficiencies so attained are expected to be higher than in today's largest and fastest computers.

Advances in digital storage technology came in three steps: the serial magnetic tape, the direct-access magnetic disk, and the digital optical disc (see above). In terms of both affordability and versatility, the optical disc surpasses the microform as the most economical analog-form storage medium, and it permits the consolidation of the digital representations of text, image, video, and sound on a single medium.

Advances in computer software and their ramifications

The widespread deployment of digital technology also is a result of parallel advances in software, which allows this technology to be used by persons other than computer experts. Early programming languages, called assembly and procedure-oriented languages, required programmers to specify in minute detail how data would be processed. So-called higher level programming languages subsequently developed have permitted programmers to stipulate only the tasks to be done without concern over how the programs are to be executed. Competent usage of these and more recent programming languages is still limited to practitioners trained in the art of programming. What has opened computer use to the general public is its interactive applications in the office and other professional environments. These applications have been made possible by the development of easy-to-use software products for the creation, maintenance, manipulation, and querying of files and records. The data-base has become a central organizing framework for information systems, taking advantage of the concept of data independence that allows data sharing among diverse applications. Many of today's information systems are tantamount to data-base systems—organized collections of data maintained and used with the aid of data-base management system (DBMS) software.

Since the late 1970s microcomputers and "friendly" application software have placed the power of digital technology at the disposal of millions. One consequence of this development was not fully anticipated: that by coupling the computer to the telephone communications network, the video terminal would become, in the short span of a decade, a window and a link to the vast resources of the digital information world. To facilitate the linkage, the telephone network for voice communications has been adapted to carry digital information, first as modulated analog signals over a twisted-wire carrier or coaxial cable and then as digital signals over high-speed microwave and satellite transmission channels. Two key innovations, the optical fibre (threadlike fibre of either glass or plastic) and the laser, used together can transmit significantly more information over greater distances than can conventional metal wire. For example, a pulse of laser light of 2,000,000,000 bits per second would transmit a 30-volume encyclopaedia across many kilometres in slightly more than one second. The optical fibre is replacing the coaxial cable in the integrated services digital network (ISDN), which is capable of carrying digital information in the form of voice, data, and video simultaneously. Common communication protocols (rules and procedures) permit computers of different manufacturers and sizes to communicate with each other in networks. Moreover, the networks themselves are interconnected by hardware devices and software called gateways that convert one protocol format to another.

Impact of fibre optics and laser technology

Such technological advances are shaping the nature of modern information systems and presage the trends of their evolution. State-of-the-art information systems process and retrieve information ever faster; they are more robust in terms of the volume of data that they can process and in terms of reliability; and their performance is more cost-effective. Even more significant effects of technological developments lie along other dimensions, however.

First, the community of users of information systems is changing dramatically, from computer programmers and professional information workers to "end users" who interact with these systems directly and personally. This change, in turn, has had a major impact on the design of information systems.

Second, the transportability of electronic information, coupled with the development of high-capacity, high-speed telecommunications facilities that permit the transmission of large volumes of digital information (such as that generated in image-processing applications), is at the base of the present trend toward teleprocessing information systems in which the processors and the data repositories are geographically dispersed. This change is conducive to data sharing, whether within a single building (as local-area networks) or more globally (as wide-area networks). The central component of information systems, the data base, may be "distributed" over a number of processors in different geographic locations, yet queries can be processed simultaneously against the composite data base. Users of teleprocessing information systems interact and collaborate within as well as across organizational and political boundaries.

Use of teleprocessing information systems

Third, the functions of information systems are constantly expanding as all information formats converted to digital form are manipulable by increasingly sophisticated techniques and programs. Evolving techniques of natural-language processing and understanding, as well as of knowledge representation and manipulation, are joining the repertoire of more traditional methods of content analysis and manipulation. Jointly these developments contribute to the gradual but continuing transformation from data-processing to cognition-aiding systems.

ANALYSIS AND DESIGN OF INFORMATION SYSTEMS

The building of information systems falls within the domain of engineering. As is true with other engineering disciplines, the nature and tools of information systems engineering are evolving due to both technological developments and better perceptions of societal needs for information services. Early information systems were designed to be operated by information professionals, and they frequently did not attain their stated social purpose. Modern information systems are increasingly used by persons who have little or no previous hands-on experience with information technology but who possess a much better perception about what this technology should accomplish in their professional and personal environments. A correct understanding of the requirements, preferences, and "information styles" of these end users is crucial to the design and success of today's information systems.

Methodology of constructing information systems

The methodology involved in building an information system consists of a set of iterative activities that are cumulatively referred to as the system's life cycle (Figure 5). The principal objective of the systems analysis phase is the specification of what the system is to do. In the

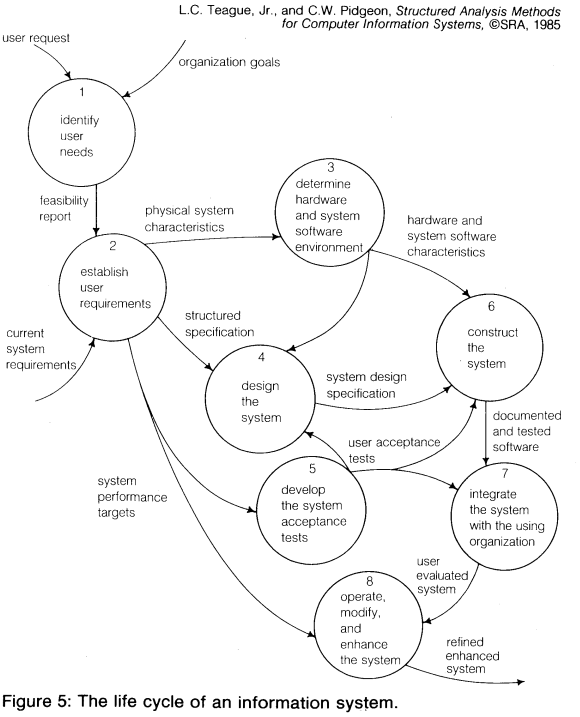


Figure 5: The life cycle of an information system.

systems design phase such specifications are converted to a hierarchy of increasingly detailed charts that define the data required and decompose the processes to be carried out on data to a level at which they can be expressed as instructions of a computer program. The systems development phase consists of writing and testing computer software and of developing data input and output forms and conventions. Systems implementation is the installation of a physical system and the activities it entails, such as the training of operators and users. Systems maintenance refers to the further evolution of the functions and structure of a system that results from changing requirements and technologies, experience with the system's use, and fine-tuning of its performance.

Many information systems are implemented with generic, "off-the-shelf" software rather than with custom-built programs; versatile data-base management software and its non-procedural programming languages fit the needs of small and large systems alike. The development of large systems that cannot use off-the-shelf software is an expensive, time-consuming, and complex undertaking. Some of the tools of computer-aided software engineering available to the systems analyst and designer verify the logic of systems design, automatically generate a program code from low-level specifications, and automatically produce software and system specifications. The eventual goal of information systems engineering is to develop software "factories" that use natural language and artificial intelligence techniques as part of an integrated set of tools to support the analysis and design of large information systems.

CATEGORIES OF INFORMATION SYSTEMS

A taxonomy of information systems is not easily developed because of their diversity and continuing evolution in structure and function. Earlier distinctions—manual versus automated, interactive versus off-line, real-time versus batch-processing—are no longer very useful. A more frequently made distinction is in terms of application: use in business offices, factories, hospitals, and so on. In the functional approach taken in this article, information systems may be divided into two categories: organizational systems and public information utilities. Information systems in formal organizations may be further distinguished according to their main purpose: support of managerial and administrative functions or support of operations and services. The former serve internal functions of the organizations, while the latter support the purposes for which these organizations exist.

Management-oriented information systems. The most important functions that top executives perform are setting policies, planning, and preparing budgets. At the strategic level, these decision-making functions are supported by executive information systems. The objective of these systems is to gather, analyze, and integrate internal (corporate) and external (public) data into dynamic profiles of key corporate indicators. Depending on the nature of the organization's business, such indicators may relate to the status of high-priority programs, health of the economy, inventory and cash levels, performance of financial markets, relevant efforts of competitors, utilization of manpower, legislative events, and so forth. The indicators are displayed as text, tables, graphics, or time series, and optional access is provided to more detailed data. The data emanate not only from within the organization's production and administrative departments but also from external information sources, such as public data bases. The structure of an executive information system contains a large mainframe computer that houses the data base and interfaces with facilities for importing internal data, accessing external data sources, and performing data analyses and modeling (Figure 6). The executive display device is a sophisticated high-resolution video terminal with which the manager interfaces by means of menus, a pointing device called a "mouse," and the keyboard.

Executive systems

In military organizations, the approximate equivalent of executive information systems is command-and-control systems. Their purpose is to maintain control over some domain and, if needed, initiate corrective action. Their key characteristic is the real-time nature of the monitoring and decision-making functions. A command-and-control system typically assumes that the environment exercises pressure on the domain of interest (say, a naval force); the system then monitors the environment (collects intelligence data), analyzes the data, compares it with the desired state of the domain, and suggests actions to be taken. Systems of this kind are used at both strategic and tactical levels.

Command-and-control systems

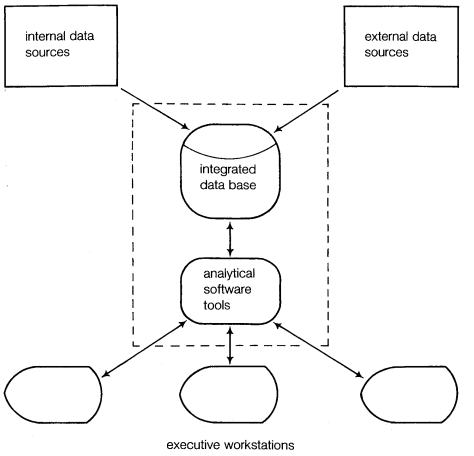


Figure 6: Structure of a typical executive information system.

Both executive and military command-and-control systems make use of computational aids for data classification, modeling, and simulation. These capabilities are characteristic of a decision-support system (DSS), a composite of computer techniques for supporting executive decision making in relatively unstructured problem situations. Decision-support software falls into one of two categories: (1) decision-aid programs, in which the decision maker assigns weighted values to every factor in the decision; and (2) decision-modeling programs, in which the user explores different strategies to arrive at the desired outcome.

Decision-support systems

Administration-oriented information systems. Administrative functions in formal organizations have as their objective the husbanding and optimization of corporate resources—namely, employees and their activities, inventories of materials and equipment, facilities, and finances. Administrative information systems support this objective. Commonly called management information systems (MIS), they focus primarily on resource administration and provide top management with reports of aggregate data. Executive information systems may be viewed as an evolution of administrative information systems in the direction of strategic tracking, modeling, and decision making.

Typically, administrative information systems consist of a number of modules, each supporting a particular function (Figure 7). The modules share a common data base whose contents may, however, be distributed over a number of machines and locations. Financial information systems have evolved from the initial applications of punched cards before World War II to integrated accounting and finance systems that cover general accounting, accounts receivable and payable, payroll, purchasing, inventory control, and financial statements such as balance sheets. Functionally close to payroll systems are personnel information systems, which support the administration of the organization's human resources. Job and salary histories, inventory of skills, performance reviews, and other types of personnel data are combined in the data base to assist personnel administration, explore potential effects of reorganization or new salary scales (or changes in benefits), and match job requirements with skills. Project management information systems concentrate on resource allocation and task completion of organized activities; they usually incorporate such scheduling methods as critical path method (CPM) or program evaluation and review technique (PERT).

Financial and personnel information systems

Since the advent of microcomputers, information pro-

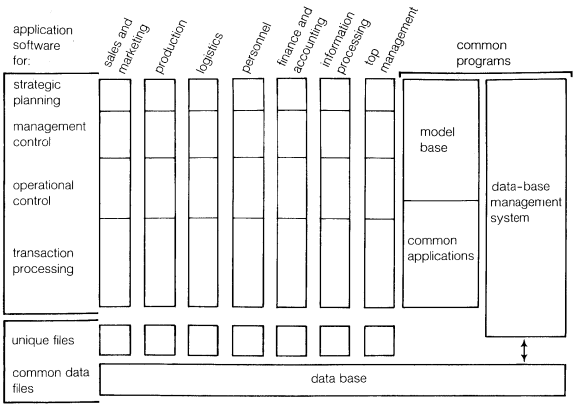


Figure 7: Structure of an administrative information system.
G.B. Davis and M.H. Olson, *Management Information Systems*, McGraw-Hill, 1985

cessing in organizations has become heavily supported by office automation tools. These involve six basic applications: text processing, data base, spreadsheet, graphics, communications, and networking. Administrative systems in smaller organizations are usually built as extensions of office automation tools; in large organizations these tools form an interface to custom software. The current trend in office automation is toward integrating the first five applications into a software utility, either delivered to each microprocessor workstation from a "server" on the corporate computer network or integrated into other applications software.

Administrative information systems abound in organizations in both the private and public sectors throughout the industrialized world. In the retail industry, point-of-sale terminals are linked into distributed administrative information systems that contain financial and inventory modules at the department, store, geographic area, and corporate chain levels, with modeling facilities that help to determine marketing strategies and optimize profits. Administrative information systems are indispensable to government; the agencies of virtually all U.S. municipalities with over 10,000 inhabitants use such systems. The systems are generally centred around a generic data-base management system and are increasingly supported by software modules and programs that permit data modeling—*i.e.,* they acquire management orientation.

Service-oriented information systems. Such information systems provide support for the operations or services that organizations perform for society. The systems are vertically oriented to specific sectors and industries (*e.g.,* manufacturing, financial services, publishing, education, health, and entertainment). Rather than addressing management and administrative functions, they support activities and processes that are the reason for an organization's existence—in most cases, some kind of manufacturing activity or the rendering of services. Systems of this kind vary greatly, but they tend to fall into three main types: manufacturing, transaction, and expert systems.

Computer-integrated manufacturing

The conceptual goal of modern factories is computer-integrated manufacturing (CIM). The phrase denotes data-driven automation that affects all components of the manufacturing enterprise: design and development engineering, manufacturing, marketing and sales, and field support and service. The development of the CIM concept has been gradual, from initial factory automation efforts aimed at saving labour costs to eventually addressing the gamut of activities of manufacturing organizations. The major components of the CIM concept are information systems for engineering design, for automation of production, and for the planning and control of these and other operations. Computer-aided design (CAD) systems were first applied in the electronics industry. Today they feature three-dimensional modeling techniques for drafting and manipulating solid objects on the screen and for deriving specifications for programs to drive numerical-control machines. Once a product is designed, its production process can be outlined using computer-aided process planning (CAPP) systems that help to select sequences of operations and machining conditions. Models of the manufacturing system can be simulated by computers before they are built. The basic manufacturing functions—machining, forming, joining, assembly, and inspection—are supported by computer-aided manufacturing (CAM) systems and automated materials-handling systems. Inventory control systems seek to maintain an optimal stock of parts and materials by tracking inventory movement, forecasting requirements, and initiating procurement orders.

The technological sophistication of manufacturing information systems is impressive, and it increasingly includes applications of artificial intelligence: robotics, computer vision, and expert systems (see COMPUTERS: *Artificial intelligence*). The core of the CIM concept is an integrated data base that supports the manufacturing enterprise and is linked with other administrative data bases.

In nonmanufacturing service organizations the prevalent type of information system supports transaction processing. Transactions are sets of discrete inputs, submitted by users at unpredictable intervals, which call for data-base searching, analysis, and modification. The processor evaluates the request and executes it immediately. Portions of the processing function may be carried out at the intelligent terminal that originated the request so as to distribute the computational load. Response time is an important characteristic of this type of real-time teleprocessing system. Large transaction-processing systems often incorporate private telecommunications networks.

Teleprocessing transaction systems

Teleprocessing transaction systems constitute the foundation of service industries such as banking, insurance, securities, transportation, and libraries. They are replacing the trading floor of the world's major stock exchanges, linking the latter via on-line telecommunications into a global financial market. Again, the core of a transaction system is its integrated data base. The focus of the system is the recipient of services rather than the system operator. This orientation represents a departure from earlier transaction systems. For example, data bases in the insurance industry were structured around insurance policy types (so that a client with three different types of policy would be represented in the data base three times); the current trend is to structure the data base around a single client record. Similar reorientation is taking place in systems used in banking, securities, and social services.

A relatively new category of service-oriented information systems is the expert system, so called because its data base stores a description of decision-making skills of human experts in some domain of performance, such as medical image interpretation, taxation, brickwork design, configuration of computer system hardware, troubleshooting malfunctioning equipment, or beer brewing. The motivation for constructing expert systems is the desire to replicate the scarce, unstructured, poorly documented empirical knowledge of a few so that it can be readily used by others.

Components of expert systems

Expert systems have three components: (1) a software interface through which the user formulates queries by which the expert system solicits further information from the user and by which it explains to the user the reasoning process employed to arrive at an answer; (2) a data base (called the knowledge base) consisting of axioms (facts) and rules for making inferences from these facts; and (3) a computer program (dubbed the inference engine) that executes the inference-making process.

The knowledge base is a linked structure of rules that the human expert applies, often intuitively, in problem solving. The process of acquiring such knowledge typically has three phases: a functional analysis of the environment, users, and tasks performed by the expert; identification of concepts of the domain of expertise and their classification according to various relationships; and an interview, either by human or automated techniques, of the expert (or experts) in action. The results of these steps are translated into so-called production rules (of the form "IF condition x exists, THEN action y follows") and stored in the knowledge base. Chains of production rules form the basis for the automated deductive capabilities of expert systems and for their ability to explain their actions to users.

Expert systems are a commercial variety of a class of computer programs called knowledge-based systems, whose

development forms one major goal of research in artificial intelligence (Figure 8). Knowledge in expert systems is highly unstructured (*i.e.,* the problem-solving process of the domain is not manifest), and it is stated explicitly in relationships or deductively inferred from the chaining of propositions. Since every condition that may be encountered must be described by a rule, expert systems cannot handle unanticipated events, but they can evolve with usage. Current expert systems are incapable of inductive inference; that is to say, they cannot generalize.

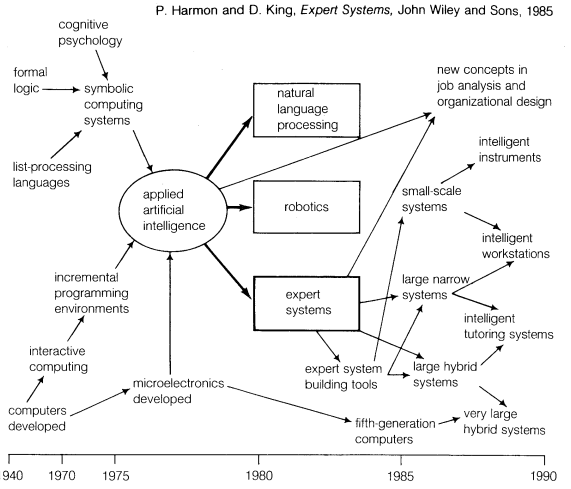


Figure 8: The domain and tools of artificial intelligence.

Public information utilities. Aside from the proliferation of organizational information systems, new types of teleprocessing systems became available for use by the public during the 1970s. With the proliferation of electronic data bases, the then new industry of "data-base vendors" began to make these resources available via on-line data-base search systems. Today, this industry operates for public access and uses hundreds of document data bases, some of them in full text; corporate and industry data and news; stock quotations; diverse statistics and time series; and catalogs of products and services.

On-line data-base search systems

In a growing number of countries, particularly in developing countries where the concepts of information resources and their management and use are relatively new, on-line data-base services are being coordinated in the form of planned "national" information systems, often in conjunction with public data networks and international telecommunications facilities. These efforts are accompanied by attempts to formulate national information policies for management and better exploitation of endogenous information resources and technologies.

In the United States, on-line data-base vendors constitute a kind of evolving electronic supermarket, a growing network of public information utilities. In addition to data bases, these services retail computer programs, acting as a giant multipurpose computational facility; software, ranging from products for business and scientific applications to computer games, may either be used on-line or be downloaded from the vendor (in the case of the latter, both programs and data—binary core images—are transferred from the vendor's computer to that of the user for execution). Recent services of public information utilities include transaction-processing systems: brokerage services to place on-line stock, bond, and options orders; home banking to pay bills and transfer funds; travel planning and reservations; and on-line catalog shopping. Some of these services combine on-line information retrieval (from, say, merchandise catalogs) and transaction processing (placing orders). Many include such functions as electronic mail and teleconferencing.

IMPACT OF COMPUTER-BASED INFORMATION SYSTEMS ON SOCIETY

Preoccupation with information and knowledge as an individual, organizational, and societal resource is stronger today than at any other time in history. The volume of books printed in 16th-century Europe is estimated to have doubled approximately every seven years. Interestingly, the same growth rate has been calculated for global scientific and technical literature in the 20th century and for business documents in the United States in the 1980s. If these estimates are reasonably correct, the growth of recorded information is a historical phenomenon, not peculiar to modern times. The present, however, has several new dimensions relative to the information resource: modern information systems collect and generate information automatically; they provide rapid, high-resolution access to the corpora of information; and they manipulate information with previously unattainable versatility and efficiency.

The proliferation of automatic data-logging devices in scientific laboratories, hospitals, transportation, and many other areas has created a huge body of primary data for subsequent analysis. Machines even generate new information: original musical scores are now produced by computers, as are graphics and video materials. Electronic professional workstations can be programmed to carry out any of a variety of functions. Some of those that handle word processing not only automatically look for spelling and punctuation errors but check grammar, diction, and style as well; they are able to suggest alternative word usage and rephrase sentences to improve their readability. Machines produce modified versions of recorded information and translate documents into other languages.

Modern information systems also bring new efficiency to the organization, retrieval, and dissemination of recorded information. Since the commercial development of digital computers in the late 1950s, the control of and access to the world's printed information have been truly revolutionized; electronic access techniques reveal the granularity of humankind's information store in hitherto unattainable detail. A similar historic event is in the offing with respect to information dissemination. The invention of printing from movable type made documents affordable to virtually all levels of society and thereby decentralized the possession of documents by individuals. In like manner, modern information technology is beginning to decentralize for personal possession collections equivalent to the holdings of entire libraries and archives.

Possibly the most significant new dimension of modern information systems lies in their ability to manipulate information automatically. This capability is the result of representing symbolic information in digital form. Computer-based information systems are able to perform calculations, analyses, classifications, and correlations at levels of complexity and efficiency far exceeding human capabilities. They can simulate the performance of logical and mathematical models of physical processes and situations under diverse conditions. Information systems also have begun to mimic human cognitive processes: deductive inference in expert systems, contextual analysis in natural-language processing, and analogical and intuitive reasoning in information retrieval. Powerful information-transforming technologies now under development—data/text to graphics, speech to printed text, one natural language to another—will broaden the availability of information and enhance human problem-solving capabilities in ways not yet foreseen.

Automatic information manipulation

Whether in manufacturing or in management and administration, information systems applications are motivated by a desire to augment the mental information-processing functions of humans or to find adequate substitutes for them. Their effects have already been felt prominently in three domains: the economy, the governance of society, and the milieu of individual existence.

Effects on the economy. Information systems are a major tool for improving the cost-effectiveness of societal investments. In the realm of the economy, they may be expected to lead to higher productivity, particularly in the industrial and service sectors—in the former through automation of manufacturing and related processes, in the latter through computer-aided decision making, problem solving, administration, and support of clerical functions. Awareness that possession of information is tantamount to a competitive edge is stimulating intensified efforts at gathering technical and economic intelligence at the cor-

porate and national levels. Similarly, concern is mounting over the safeguarding and husbanding of proprietary and strategic information within the confines of organizations as well as within national borders.

Changes in the distribution of labour

The growing number of information-processing applications is altering the distribution of labour in national economies. The deployment of information systems has resulted in the dislocation of labour, but it has not as yet had an appreciable long-term effect on unemployment because the implementation of such systems tends to engender a concomitant growth of other occupations, notably in the information service sector itself. Productivity increases induced by automation (particularly robotics) in industrially advanced countries, however, threaten to further lower the competitive potential of the labour-intensive economies of Third World nations, at least in the short run. It should be noted that most studies in this area have been controversial and that their findings remain inconclusive.

Effects on governance and management. That knowledge is power was understood by heads of state and military leaders long before the English essayist and statesman Francis Bacon (1561–1626) enunciated this maxim. In modern times the degrees of governmental control of information media serve as a testament to the power of information as well as a barometer of democracy.

Regardless of a country's political system, formal organizations tend to be better equipped than individuals to take advantage of information systems: their record-keeping practices are more mature and efficient; they have better facilities and skill to access and interpret information; and they possess larger financial resources to avail themselves of the powerful analytical tools that information systems provide. On the one hand, information systems harbour a potential for higher-quality governance and management but also entail possibilities for gross errors or for being misused. On the other, public information and communication utilities help to forge informal societal networks that counterbalance the power of formal organizations and increasingly divert their style of management in the direction of consultation with the well-informed.

An environment that encourages the use of information technology and systems fosters what might be termed high information maturity on the part of the populace, a prerequisite to participatory governance. Attempts to control information technology and resources may indicate budgetary exigencies but often signify a preference for autocratic forms of management.

Effects on the individual. Perhaps the most overt impact of modern information systems concerns the individual's standard and style of living. Information systems affect the scope and quality of health care, make social services more equitable, enhance personal comfort, provide a greater measure of safety and mobility, and extend the variety of leisure forms at one's disposal. More subtly but equally as important, they also affect the content and style of an individual's work and in so doing perturb the social and legal practices and conventions to which he is accustomed. At the same time, new kinds of information products and media necessitate a redefinition of the legal conventions regulating the ownership of products of the human intellect. Moreover, massive data-collecting systems bring into sharp focus the elusive borderline between the common good and personal privacy. Both the public and private sectors have recognized the need to safeguard stored data against accidental or illegal access. Unauthorized disclosure or misuse, including modification, of information of a confidential or personal nature (*e.g.*, data from income tax returns or family medical records) may result in serious damage or embarrassment at the least. In the United States, Canada, and many other countries, statutes have been enacted to impose legal requirements on government agencies to ensure the confidentiality of records in their possession. Other measures have been taken to promote data integrity and security, as, for example, the implementation of the Data Encryption Standard (DES), a cipher used in unclassified U.S. government applications for the protection of binary-coded data during transmission and storage in computer-based systems.

Concern over privacy and security

A more fundamental issue of paramount long-term significance for society has to do with the quality of the individual's spiritual existence in an increasingly knowledge-intensive environment. In such an environment, knowledge is the principal and perhaps most valuable currency. The growing volume and the rate of obsolescence of knowledge compel the individual to live in continuous presence of, and frequent interaction with, information resources and systems. Effective use of these resources and systems may be a modern definition of literacy, while the absence of such a skill may very well result in intellectual and possibly economic poverty. It has sometimes been suggested that automated machines, particularly "intelligent" systems, might eventually subjugate humankind. Yet, the real danger is not that such systems will take over but rather that humans, unwilling or incapable or not permitted to use them, may be relegated to an existence that falls short of the human potential.

Information science

The growing interest in information processing and the systems involved has given rise to an interdisciplinary field of study known as information science. This is an extremely broad field that draws on such disciplines as computer science, library science, communications, linguistics, and psychology.

THE EMERGENCE OF INFORMATION SCIENCE

The development of information science in the United States can be traced to events in the 1950s. One of these was the founding of the Center for Documentation and Communication as an adjunct to the library school at Western Reserve (now Case Western Reserve) University in Cleveland. The new centre led the way in the study of bibliographic information systems, particularly the computer-based variety. Efforts at the centre to define the territory of the new discipline through workshops and academic conferences led to the formation of several informal information study centres at other universities, including the University of California at Los Angeles (UCLA); American University, Washington, D.C.; University of Washington; and Georgia Institute of Technology.

The development of a discipline specializing in the study of information and information systems was given a major boost in 1958, when the U.S. National Defense Education Act directed the National Science Foundation to establish a service to "(1) provide . . . indexing, abstracting, translating, and other services leading to a more effective dissemination of scientific information, . . . and (2) undertake programs to develop new or improved methods, including mechanized systems, for making scientific information available." While the resulting Science Information Service focused on the handling of scientific and engineering bibliographies and texts, other areas received attention, including the study of communication processes in science and industry and the development of more efficient systems for organizing and storing scientific information.

The discipline of information science was formalized as a result of two conferences held at the Georgia Institute of Technology in 1961 and 1962. For the first time the interdisciplinary nature of the field was outlined, guidelines for research and study were established, and goals for a formal curriculum were defined. The 1961 conference offered the first comprehensive definition of information science as "the science that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing information for optimum accessibility and usability. The processes include the origination, dissemination, collection, organization, storage, retrieval, interpretation, and use of information. The field is derived from, and related to, mathematics, logic, linguistics, psychology, computer technology, operations research, the graphic arts, communications, library science, management, and other fields." This definition gave the new discipline great flexibility to grow with advances in related fields and also gave information scientists a mandate to search for the "grand, unifying relationships of human and human-machine communication."

Development of information science programs in the United States

Formal information science programs were quickly developed at colleges and universities throughout the United States. Some were established in separate departments or as independent programs, such as those at the Georgia Institute of Technology, Ohio State University, and Lehigh University, Bethlehem, Pa., while others, including those at the University of Chicago, Case Western Reserve University, and UCLA, were formed as integral parts of existing library science departments. Information science programs became entrenched in these and other schools during the mid- to late 1960s, and representatives from these schools used formal conferences and informal reports to convey their views and experiences in the new field to other academicians. In the years that followed, information science occasionally achieved departmental status; more often, however, it had elements of its subject matter integrated into the curricula of other departments. The general tendency has been for the theory- and technology-oriented subjects of information science to converge with computer science and engineering, while those pertaining to information management and services are taught as part of library science. Management information systems, on the other hand, have become a major concern of schools of business administration.

SUBJECTS OF RESEARCH

The initially applied activities of information science—automation of bibliographic control and nonnumeric, natural-language processing—provided the impetus for research in such areas as information description, coding, and organization. Operational information systems opened research opportunities in many other areas, including information economics, the study of information value and utility in different settings; and bibliometrics, the quantification of processes of written communication using mathematical methods to classify and evaluate scientific journals. Studies were also undertaken on models of the effectiveness of information storage and retrieval; modes of human–machine interaction; and the effect of form on the content and comprehension of information. Today, researchers in information science have three basic concerns: the search for an ever more general and useful definition and description of the nature of information; the study of the processes of information generation, transmission, transformation, storage, etc.; and the establishment of laws, theories, and other general principles that explain and predict information phenomena.

Interdisciplinary concerns

Because studies of information phenomena are extremely broad and diverse, more often than not they emanate from research communities whose domains of interest have allied foci: linguistics, the cognitive sciences, artificial intelligence, engineering, mathematics, the social and behavioral sciences, management, and of course computer science. Like the phenomena that it studies, information science is a composite, multidisciplinary field of inquiry whose structure continues to evolve.

PROFESSIONAL INFORMATION ASSOCIATIONS

There are hundreds of national, regional, and local professional associations concerned with information-related disciplines and topics. Some are large formal organizations that publish journals and newsletters and hold conventions; others are informal local groups of information professionals who meet to exchange ideas. Both are important sources of information for people working in this rapidly changing field, in which technological and institutional changes occur almost monthly.

Major national organizations

The largest U.S. organization of information processing professionals is the Association for Computing Machinery. Three important national organizations that are concerned with the general aspects of information and information science are the American Association for the Advancement of Science (Section T—Information, Computing and Communication), the American Library Association (Information Science and Automation Division), and the American Society for Information Science. Most information-related organizations are members of the American Federation of Information Processing Societies, which in turn is a member of the International Federation for Information Processing, in Geneva, Switz. The Association of Special Libraries and Information Bureaux is the main professional organization in Great Britain.

Other important associations are concerned with more limited or specific aspects of information science. Among these are the American Society of Indexers, Associated Information Managers, the Association of Information and Dissemination Centers, the Association for Systems Management, the Information Industry Association, and the Society for Management Information Systems. (V.Sl.)

BIBLIOGRAPHY

Concepts of information and information systems: A wide-ranging discussion by 39 scientists of the nature and goals of the information, computer, communication, and systems sciences appears in FRITZ MACHLUP and UNA MANSFIELD (eds.), *The Study of Information: Interdisciplinary Messages* (1983). A comprehensive, basic survey is offered in STEVEN L. MANDELL, *Computers and Information Processing: Concepts and Applications* (1987). Fundamental concepts of information representation and processes are dealt with, sometimes speculatively, in MARVIN MINSKY, *The Society of Mind* (1986); ROGER C. SCHANK, *Conceptual Information Processing* (1975); ALLEN NEWELL and HERBERT A. SIMON, *Human Problem Solving* (1972); and HERBERT A. SIMON, *The Sciences of the Artificial,* 2nd rev. ed. (1981). BÖRJE LANGEFORS, *Theoretical Analysis of Information Systems,* 4th ed. (1973), explores fundamental aspects of structure and design. The seminal publication А.И. МИХАЙЛОВ, А.И. ЧЕРНЫЙ, and Р.С. ГИЛЯРЕВСКЙ, *Основы информатики,* 2nd rev. ed. (1968), available also in German translation as A.I. MICHAILOW, A.I. CERNYI, and R.S. GILJAREVSKIJ, *Grundlagen der wissenschaftlichen Dokumentation und Information,* 2 vol. (1970), offers an early Soviet viewpoint on the subject. For independent study of the subject, some reference sources can be of use. See DENNIS LONGLEY and MICHAEL SHAIN, *Dictionary of Information Technology,* 2nd ed. (1986), for a professional reader; and A.J. MEADOWS, M. GORDON, and A. SINGLETON, *Dictionary of Computing and New Information Technology,* 2nd ed. (1984), for a nonexpert. G.G. WILKINSON and A.R. WINTERFLOOD (eds.), *Fundamentals of Information Technology* (1987), is a survey of information processing methods.

Information processing: JENS RASMUSSEN, *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering* (1986); and TERRY WINOGRAD and FERNANDO FLORES, *Understanding Computers and Cognition: A New Foundation for Design* (1986), address interface issues arising in computer processing. An in-depth exploration of natural-language processing, reasoning, vision, and learning is provided in EUGENE CHARNIAK and DREW MCDERMOTT, *Introduction to Artificial Intelligence* (1985). BARBARA J. GROSZ, KAREN SPARCK JONES, and BONNIE LYNN WEBBER (eds.), *Readings in Natural Language Processing* (1986), offers a compendium of contributions relating to information retrieval and artificial intelligence. GIO WIEDERHOLD, *File Organization for Database Design* (1987), treats digital information-storage structures systematically and quantitatively. A comprehensive discussion of text indexing and retrieval techniques is given in GERARD SALTON and MICHAEL J. MCGILL, *Introduction to Modern Information Retrieval* (1983); whereas C.J. VAN RIJSBERGEN, *Information Retrieval,* 2nd ed. (1979), offers a mathematical treatment. Standard earlier references on largely manual methods of content analysis of text include GERARD SALTON, *A Theory of Indexing* (1975); DAGOBERT SOERGEL, *Indexing Languages and Thesauri: Construction and Maintenance* (1974); B.C. VICKERY, *Classification and Indexing in Science,* 3rd ed. (1975); and HAROLD BORKO and CHARLES L. BERNIER, *Abstracting Concepts and Methods* (1975). Representative works on the processing of non-textual information include AZRIEL ROSENFELD and AVINASH C. KAK, *Digital Picture Processing,* 2nd ed., 2 vol. (1982); SATOSI WATANABE, *Pattern Recognition: Human and Mechanical* (1985); and STEVEN HARRINGTON, *Computer Graphics: A Programming Approach,* 2nd ed. (1987). THOMAS W. PARSONS, *Voice and Speech Processing* (1986), reviews appropriate techniques. An engrossing introduction to information processing for the artist is given in STEPHEN WILSON, *Using Computers to Create Art* (1986).

Organizational information systems: IVAN F. JACKSON, *Corporate Information Management* (1986), outlines pragmatic principles and problems of managing information in formal organizations; while PAUL A. STRASSMANN, *Information Payoff: The Transformation of Work in the Electronic Age* (1985), is concerned with justification of automation in such organizations. SITANSU S. MITTRA, *Decision Support Systems: Tools and Techniques* (1986), reviews tools of strategic decision making. Military decision making is the subject of A.M. WILLCOX, M.G. SLADE, and P.A. RAMSDALE, *Command, Control, & Communications (C3)* (1983). GORDON B. DAVIS and MARGRETHE H.

OLSON, *Management Information Systems: Conceptual Foundations, Structure, and Development,* 2nd ed. (1985), is a standard management-oriented text on information systems in business organizations. R.A. HIRSCHHEIM, *Office Automation: Social and Organizational Perspective* (1985), considers information systems in the office mainly from a cultural and anthropological point of view; a technological approach to the subject is found in WILLIAM M. NEWMAN, *Designing Integrated Systems for the Office Environment* (1987). ISAAC ASIMOV and KAREN A. FRENKEL, *Robots, Machines in Man's Image* (1985), is a popularly written introduction to the history, uses, and future of robots; scientific fundamentals of the developments are outlined in K.S. FU, R.C. GONZALEZ, and C.S.G. LEE, *Robotics: Control, Sensing, Vision, and Intelligence* (1987). Factory automation is studied in ULRICH REMBOLD, CHRISTIAN BLUME, and RUEDIGER DILLMANN, *Computer-Integrated Manufacturing Technology and Systems* (1985); and MARTIN J. HAIGH, *An Introduction to Computer-Aided Design and Manufacture* (1985). PAUL HARMON and DAVID KING, *Expert Systems: Artificial Intelligence in Business* (1985), provides a readable treatment of the subject.

Public information utilities: A broad view of the state-of-the-art library networking in the United States appears in SUSAN K. MARTIN, *Library Networks, 1986–87: Libraries in Partnership* (1986). B.C. VICKERY, RICHARD G. HESELTINE, and CAROLYN S. BROWN, *Information System Dynamics: The Impact of Interactive Information Networks on the UK Library and Information System* (1982), studies the British situation. ANTONE F. ALBER, *Videotex/Teletext: Principles and Practices* (1985), examines commercial aspects of information technology. J.S. QUATERMAN and J.C. HOSKINS, "Notable Computer Networks," *Communications of the ACM,* 29:932–971 (October 1986), reviews the multiplicity of global telecommunications networks and their services. STARR ROXANNE HILTZ and MURRAY TUROFF, *The Network Nation: Human Communication via Computer* (1978), deals with the early use of computer networking for informal communication in science. ROBERT JOHANSEN, *Teleconferencing and Beyond: Communications in the Office of the Future* (1984), explores the potential of electronic communication in business. Development of integrated services digital networks is outlined in ROBERT K. HELDMAN, *Telecommunications Management Planning: ISDN Networks, Products, and Services* (1987).

Impact of information systems: Seminal studies of the "information era" include FRITZ MACHLUP, *Knowledge, Its Creation, Distribution, and Economic Significance,* 3 vol. (1980–84); MARC URI PORAT, *The Information Economy,* 9 vol. (1977); DANIEL BELL, *The Coming of Post-Industrial Society: Venture in Social Forecasting* (1973, reprinted 1976); and JAMES R. BENIGER, *The Control Revolution: Technological and Economic Origins of the Information Society* (1986). The French view of the social effects of computers in telecommunications is presented in SIMON NORA and ALAIN MINC, *The Computerization of Society: A Report to the President of France* (1980; originally published in French, 1978); and Japanese research on the subject is summarized in Y. ITO and K. OGAWA, "Recent Trends in Johoka Shakai and Yohoka Policy Studies," *Keio Communication Review,* vol. 5 (1984). The importance of preserving individual freedoms in the information age is argued eloquently in ITHIEL DE SOLA POOL, *Technologies of Freedom* (1983). A number of future scenarios are examined in CLEMENT BEZOLD and ROBERT OLSON, *The Information Millennium: Alternative Futures* (1986), which draws on popular intuitive accounts of the early 1980s. DONALD MICHIE and RORY JOHNSTON, *The Knowledge Machine: Artificial Intelligence and the Future of Man* (1985), speculates on the impact of advanced computers on everyday life. Changes in the nature of work processes brought about by information technology are discussed in TOM FORESTER, *High-Tech Society: The Story of the Information Technology Revolution* (1987). ARTHUR R. MILLER and MICHAEL H. DAVIS, *Intellectual Property: Patents, Trademarks, and Copyright in a Nutshell* (1983), elaborates on issues of property rights in information; and CEES J. HAMELINK, *Cultural Autonomy in Global Communications: Planning National Information Policy* (1983), addresses transborder information flows. SIDNEY KARIN and NORRIS PARKER SMITH, *The Supercomputer Era* (1987), provides a readable survey of developments in computer capacity and examines its role in science and business.

Information science: PRANAS ZUNDE, *On Empirical Foundations of Information Science* (1981), explores fundamental philosophical and methodological aspects of information science. Interdisciplinary theoretical concepts integral to information theory are examined in LAURENCE B. HEILPRIN (ed.), *Toward Foundations of Information Science* (1985). The connection of information science to librarianship is discussed in JESSE H. SHERA, *Introduction to Library Science: Basic Elements of Library Service* (1976); B.C. VICKERY and A. VICKERY, *Information Science: Theory and Practice* (1984); *Encyclopedia of Library and Information Science,* edited by ALLEN KENT *et al.,* 35 vol. (1968–83), continued with supplemental volumes (irregular); DOROTHY B. LILLEY and ROSE MARIE BADOUGH, *Library and Information Science* (1982); and CHARLES H. DAVIS and JAMES E. RUSH, *Guide to Information Science* (1979).

Bibliographic sources: Useful series of printed reviews in information science include *Annual Review of Information Science and Technology; Advances in Information Systems Science* (irregular); *Advances in Computers* (irregular); *Advances in Robotics* (irregular); *Computer Integrated Manufacturing Review* (quarterly); *Computing Reviews* (monthly), a major abstracting journal, with an accompanying annual *ACM Guide to Computing Literature; Computer Literature Index* (quarterly), a major indexing source in the field; *Computer Abstracts* (monthly), a significant abstracting journal; and *Computer Book Review* (monthly), surveying new publications in a broad range of subjects on computing. Useful secondary sources that list publications dealing with information science and systems include *Information Science Abstracts* (monthly); and *Library & Information Science Abstracts* (monthly), containing abstracts of literature of librarianship and archives, documentation, publishing, dissemination of information, and mass communications.

(V.Sl.)

Information Theory

One of the most prominent features of 20th-century technology has been the development and exploitation of new communications media. Concurrent with the growth of devices for transmitting and processing information, a unifying theory known as information theory was developed and became the subject of intensive research. It is an example of a theory that was initiated primarily by one man, the U.S. electrical engineer Claude E. Shannon, whose initial ideas appeared in the article "The Mathematical Theory of Communication" in the *Bell System Technical Journal* (1948). In its broadest sense, information is interpreted to include the messages occurring in any of the standard communications media, such as telegraphy, radio, or television, and the signals involved in electronic computers, servomechanism systems, and other data-processing devices. The theory is even applied to the signals appearing in the nerve networks of humans and other animals. The signals or messages do not have to be meaningful in any ordinary sense.

The chief concern of information theory is to discover mathematical laws governing systems designed to communicate or manipulate information. It sets up quantitative measures of information and of the capacity of various systems to transmit, store, and otherwise process information.

Some of the problems treated are related to finding the best methods of using various available communication systems and the best methods for separating the wanted information, or signal, from the extraneous information, or noise. Another problem is the setting of upper bounds on what it is possible to achieve with a given information-carrying medium (often called an information channel). While the central results are chiefly of interest to communication engineers, some of the concepts have been adopted and found useful in such fields as psychology and linguistics. The boundaries of information theory are quite vague. The theory overlaps heavily with communication theory but is more oriented toward the fundamental limitations on the processing and communication of information and less oriented toward the detailed operation of the devices employed.

See the articles TELECOMMUNICATIONS SYSTEMS; INFOR-

MATION PROCESSING; INFORMATION SCIENCE; and COMPUTERS for a discussion of applications of information theory.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/23 and 735.

This article is divided into the following sections:

Central problems of information theory 569
  Elements of a general communication system
  Information as a physical quantity
Measurement of information 569
  Binary choices and bits
  Message sequence
Encoding information 570
Transmission of information 571
  Role of noise
  Channel capacity
Coding for error correction 571
  Parity checks
  Codes with high code rate and low probability of error
  Convolutional codes
Bandwidth-limited channels 572
Sources with distortion measures 573
Detection, estimation, and prediction 574
  Detection problems
  Estimation and prediction problems
Cryptographic, linguistic, and other applications 574
Bibliography 575

CENTRAL PROBLEMS OF INFORMATION THEORY

Elements of a general communication system. The type of communication system that has been most extensively investigated is shown in Figure 1. It consists of the following elements. (A) An information source that produces the raw information, or "message," to be transmitted. (B) A transmitter that transforms, or encodes, this information into a form suitable for the channel. This transformed message is called the signal. (C) The channel on which the encoded information, or signal, is transmitted to the receiving point. During transmission the signal may be changed or distorted. The static in radio reception and the snow in television reception are familiar examples of such changes. These disturbing effects are known generally as noise and are indicated schematically in Figure 1 by the noise source. (D) The receiver, which decodes, or translates, the received signal back into the original message or an approximation of it. (E) The destination or intended recipient of the information.

This system is sufficiently general to include a wide variety of communication problems if the various elements are suitably interpreted. In radio, for example, the information source may be a person speaking into a microphone. The message is then the sound that is produced, and the transmitter is the microphone and the associated electronic equipment that changes this sound into an electromagnetic wave, the signal. The channel is the space between the transmitting and receiving antennas, and any static or noise disturbing the signal corresponds to the noise source in the schematic diagram. The radio receiver converts the received signal into an audible output from a loudspeaker. The destination is a person listening to the message.

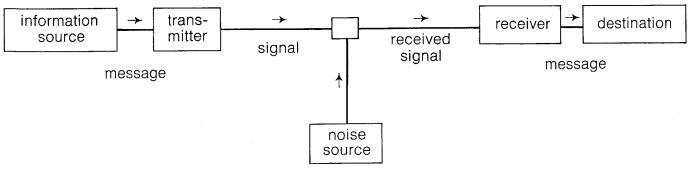


Figure 1: Elements of a general communication system.

Information as a physical quantity. A basic idea in information theory is that information can be treated very much like a physical quantity, such as mass or energy. A homely analogy may be drawn between the system in Figure 1 and a transportation system; for example, an information source (user who generates the message) is like a lumber mill producing lumber at a certain point. The channel in Figure 1 might correspond to a conveyor system for transporting the lumber to a second point. In such a situation there are two important quantities: the rate R (in cubic metres per second) at which lumber is produced at the mill and the capacity C (in cubic metres per second) of the conveyor. These two quantities determine whether or not the conveyor system will be adequate for the lumber mill. If the rate of production R is greater than the conveyor capacity C, it will certainly be impossible to transport the full output of the mill; there will not be sufficient space available. If R is less than or equal to C, it may or may not be possible, depending on whether the lumber can be packed efficiently in the conveyor. Suppose, however, that there is a sawmill at the source, which corresponds in the analogy to the encoder or transmitter. Then the lumber can be cut into small pieces in such a way as to fill the available capacity of the conveyor with 100-percent efficiency. In this case a carpenter would be provided at the receiving point to fasten the pieces back together in their original form before passing them on to the consumer.

If this analogy is sound, it should be possible to set up a measure R, in suitable units, giving the rate at which information is produced by a given information source, and a second measure C, determining the capacity of a channel for transmitting information. Furthermore, the analogy would suggest that by a suitable coding or modulation system the information can be transmitted over the channel if and only if the rate of production R is not greater than the capacity C. A key result of information theory is that it is indeed possible to set up measures R and C having this property.

MEASUREMENT OF INFORMATION

Before considering the question of how information is to be measured, it is necessary to clarify the precise meaning of "information." Often the messages to be transmitted have meaning: they describe or are related to real or conceivable events. This, however, is not always the case. In transmitting music, for example, the meaning, if there is any, is much subtler than in the case of a verbal message. In some situations the engineer is faced with transmitting a totally meaningless sequence of numbers or letters. In any case, meaning is quite irrelevant to the problem of transmitting the information.

Irrelevancy of meaning to the transmission of information

It is as difficult to transmit a series of nonsense syllables as it is to transmit straight English text. The significant aspect of information from the standpoint of transmission is the fact that one particular message is chosen from a set of possible messages. What must be transmitted is a specification of the particular message that was chosen by the information source. The original message can be reconstructed at the receiving point only if such an unambiguous specification is transmitted. Thus, in information theory, information is thought of as a choice of one message from a set of possible messages. Furthermore, these choices occur with certain probabilities; some messages are more frequent than others.

Binary choices and bits. The simplest type of choice is a binary choice, a choice from two equally likely possibilities; that is, each has a probability 1/2. This is the situation, for example, when a tossed coin is equally likely to come up heads or tails. It is convenient to use the amount of information produced by such a choice as the basic unit, and this basic unit is called a bit. The choice involved with one bit of information can be indicated schematically as shown in Figure 2 (left). Either the upper or the lower line may be chosen with probability 1/2 for each possibility.

If there are N possibilities, all equally likely, the amount of information is given by $\log_2 N$. The reason for this can be seen from Figure 2 (right), in which there are eight possibilities, each with probability 1/8. The choice can be imagined to occur in three stages, each involving one bit. The first bit corresponds to a choice of either the first four or the second four of the eight possibilities; the second bit corresponds to the first or second pair of the four chosen; and the final bit determines the first or second member

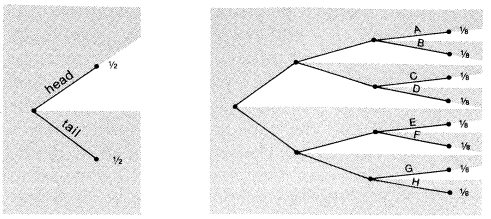


Figure 2: *Possible forms of choices.*
(Left) A binary choice, as in the flip of a coin, in which head and tail each has probability 1/2. (Right) Three successive binary choices, leading to eight possible outcomes, A through H, each with probability 1/8 (see text).

of the pair. The number of bits required is $\log_2 N$, in this case $\log_2 8 = 3$.

If the probabilities are not equal, the messages in the set have different amounts of information associated with them. If the probabilities of the possible messages are given by $p_1, p_2, \ldots, p_N$, then the amount of information associated with the first message is $\log_2 (1/p_1)$, that of the second is $\log_2 (1/p_2)$, and so forth. The expected value of these amounts of information (see Box, formula 1) is called the entropy, H, or the average information of the message set.

$$(1) \qquad H = p_1 \log_2 (1/p_1) + p_2 \log_2 (1/p_2) + \cdots + p_N \log_2 (1/p_N)$$

The entropy takes on its smallest value, zero, when one message is certain to occur (*i.e.,* has a probability of 1) and all other messages never occur (*i.e.,* have 0 probability). Intuitively, there is no information in a message that is a priori certain. Conversely, the entropy takes on its maximum value, $\log_2 N$, when the N messages in the set are equally likely. Individual messages in the set can have arbitrarily large amounts of information, but these messages with large information occur so rarely that the expected value remains bounded by $\log_2 N$.

The parlour game "Twenty Questions" illustrates some of these ideas. In this game, one person thinks of an object, and the other players attempt to determine what it is by asking not more than 20 questions that can be answered "yes" or "no." According to information theory, each question can, by its answer, yield anywhere from no information to $\log_2 2$, or one bit of information, on the average, depending upon whether the probabilities of "yes" and "no" answers are very unequal or approximately equal. To obtain the greatest amount of information, the players should ask questions that subdivide the set of possible objects as nearly as possible into two equally likely groups. For example, if they have established by previous questions that the object is a town in the United States, a good question would be, "Is it east of the Mississippi River?" This divides the possible towns into two roughly equal sets. The next question then might be "Is it north of the Mason and Dixon line?" If the players were able to choose questions that always had the effect of subdividing the set of possible objects into two equal groups, they would be able to isolate, in 20 questions, one object from 2^{20} (approximately 1,000,000) possibilities. This corresponds to 20 bits.

The mathematical formula for entropy in information theory is identical in form to that representing entropy in the field of statistical mechanics. There are a number of other connections between statistical mechanics and information theory, and some authors develop thermodynamics and statistical mechanics from the standpoint of information theory. These connections do not have to be considered, however, in the engineering and other applications of information theory.

Message sequence. Most information sources produce a message that consists not of a single choice but of a sequence of choices—for example, the letters of printed text or the elementary words or sounds of speech. The writing of English sentences can be thought of as a process of choice: choosing a first word from possible first words with various probabilities; then a second, with probabilities depending on the first; and so forth. This kind of statistical process is called a stochastic process; and information sources are thought of, in information theory, as stochastic processes. A more general formula for H can be given that determines the rate at which information is produced by a stochastic process or an information source.

Information sources as stochastic processes

Printed English is a type of information source that has been studied considerably. By playing a kind of "Twenty Questions" game, suitably modified, with subjects trying to guess the next letter in an English sentence, it can be shown that the information rate of ordinary written English is not more than about one bit per letter. This phenomenon is a result of the very unequal frequencies of occurrence of different letters (for example, *E, T,* and *A* are very common in English, while *Z, Q,* and *X* are infrequent), of pairs of letters (*TH* is very common and *QZ* rare), and of the existence of frequently recurring words, phrases, and so on. This body of statistical data related to a language is called the statistical structure of the language. If all 26 letters and the space in English had equal frequencies of occurrence (*i.e.,* each had probability 1/27) and the occurrence of each letter of text was independent of previous letters, the information rate would be $\log_2 27$, or about 4.76 bits per letter. Because only one bit actually is produced, English is said to be about 80 percent redundant.

The redundancy of English is exhibited by the fact that a great many letters can be deleted from a sentence without making it impossible for a reader to fill the gaps and determine the original meaning. For example, in the following sentence the vowels have been deleted:

MST PPL HV LTTL
DFFCLTY N RDNG THS SNTNC.

As might easily be deduced, redundancy in a language plays an important role in cryptography (see CRYPTOLOGY: *Cryptography*).

ENCODING INFORMATION

An important feature of the measure of information, H, is that it determines the saving in transmission time that is possible, by proper encoding, because of the statistics of the message source. To illustrate this, a model language is considered in which there are only four letters—*A, B, C,* and *D*—with the probabilities 1/2, 1/4, 1/8, and 1/8, respectively. In a long text in this language, *A* will occur one-half of the time, *B* one-quarter of the time, and *C* and *D* each one-eighth of the time. If this language is to be encoded into binary digits, 0 or 1, as for example in a pulse system with two types of pulse, then the most direct code is the following: $A = 00$; $B = 01$; $C = 10$; $D = 11$. This code requires two binary digits per letter of message. By proper use of the statistics, a better code can be constructed as follows: $A = 0$; $B = 10$; $C = 110$; $D = 111$. It is readily verified that the original message can be recovered from its encoded form. Furthermore, the number of binary digits used is smaller on the average. It will be, in fact, $(1)\frac{1}{2} + (2)\frac{1}{4} + (3)\frac{1}{8} + (3)\frac{1}{8} = 1\frac{3}{4}$, where the first term is due to the letter *A,* which occurs half the time and is one binary digit long, and similarly for the others. It may be found by a simple calculation that $1\frac{3}{4}$ is just the value of H, calculated for the probabilities 1/2, 1/4, 1/8, and 1/8. This basic idea, that of encoding common letters into short messages and uncommon letters into longer messages, is used in Morse Code.

The result verified for the special case above holds generally. If H is the entropy, in bits per letter, of an information source, then every binary encoding that represents the source uses at least H binary digits per source letter on the average. Conversely, binary encodings can be found using as close to H binary digits per source letter as desired. The difference in the general case is that the encoding becomes more complicated as the average number of binary digits per source letter approaches H. Rather than encoding single source letters into sequences of binary digits, it is required to map sequences of source letters into sequences of binary digits.

Sequences of binary digits

This important result in information theory gives a direct meaning to the entropy H for a source or a language. It

says, in fact, that H can be interpreted as the equivalent number of binary digits required when the language or source is encoded into 0 and 1 in the most efficient way. For instance, if the estimate of one bit per letter, mentioned above as the rate for printed English, is correct, then it is possible to encode printed English into binary digits using, on the average, one for each letter of text, and, furthermore, no encoding method would average less than this. The techniques used in practice to encode data sources with close to the minimum bit rate are called data-compression techniques. These techniques are often adaptive, automatically adjusting the encoding to the statistics of a given source.

TRANSMISSION OF INFORMATION

Role of noise. It is necessary to define information transmission when a message is transmitted over a channel. If there is no noise and the message is faithfully reproduced at the destination, the problem is simple. The information transmitted (measured in bits) is then the same as the information of the source message, and the average information transmitted is the same as the entropy of the source.

For a noisy channel, the situation is more complex. One simple example is called a binary erasure channel. If a source is binary, producing 0s and 1s with equal probability, and nine-tenths of the time the message gets through to the destination correctly and the other one-tenth of the time the noise is so bad that a special symbol called "erasure" is delivered to the destination, then it is concluded that nine-tenths of the time a bit of information is transmitted and one-tenth of the time no information is transmitted. This leads to an average information transmission of 9/10 bit per message.

Binary erasure and binary symmetric channels

Another simple example is known as a binary symmetric channel. When a 0 is transmitted, 0 is received with probability 9/10 and 1 is received with probability 1/10. Similarly, a transmitted 1 is received as a 1 with probability 9/10 and as a 0 with probability 1/10. Assuming the source to produce 0 and 1 with equal likelihood, the entropy, or average uncertainty, of a source message is 1 bit. After a 0, say, has been received at the destination, the a posteriori probability that a 0 came out of the source is 9/10, and that of a 1 is 1/10. Thus, the uncertainty of the source, conditioned on a received 0, is $9/10 \log_2 10/9 + 1/10 \log_2 10$, or about 0.47 bit. The same uncertainty arises if a 1 is received, so that the reception of a symbol reduces the uncertainty of the source message from 1 bit to 0.47 bit. It is therefore reasonable to conclude that 0.53 bit of information has been transmitted over the binary symmetric channel.

The difference between these two examples is striking. In both, 90 percent of the transmitted digits are correctly received, but the binary symmetric channel, in which it is not known what received digits are in doubt, conveys much less information than the binary erasure channel, in which dubious digits are overwritten with the erasure symbol. A similar example occurs in a "true–false" examination in which a student who gets 90 questions correct and leaves 10 unanswered appears to know more than a student who gets 90 correct and 10 wrong.

In general, the average transmitted information through a channel can be found in the same way as in the above examples. For each possible output symbol there is an uncertainty of the source, based on the a posteriori probabilities for that particular output.

The conditioned entropy of the channel input, given the channel output, is the average of the above uncertainties, weighted by the output symbol probabilities. The average transmitted information is the difference between the entropy of the input and the conditioned entropy of the input, given the output. In the special case when the input is uniquely specified by the output, the conditioned entropy is zero, and the transmitted information is the same as the original uncertainty of the input.

Channel capacity. The most important attribute of a communication channel is its capacity. Capacity is defined as the maximum (over the input probabilities) average transmitted information for a channel. The capacities of the binary erasure channel and binary symmetric channels previously discussed are 0.9 and 0.53 bits per symbol, respectively.

Channel capacity and source entropy are linked together by the noisy-channel coding theorem, which is the fundamental result of information theory. This result states that, if the source entropy (in bits per unit time) is less than the channel capacity (in bits per unit time), then encoders and decoders can be constructed such that the source output is re-created at the destination with as small an error probability as desired. Conversely, if the source entropy is greater than the capacity, then the error probability cannot be made arbitrarily small. As an example of this result, the binary erasure channel is considered, and it is assumed that a return channel is available so that, whenever an erasure occurs at the receiver, notification of it is sent back to the transmitter and the offending digit repeated. Because the transmitted digit is received correctly nine-tenths of the time and correct digits are never repeated, there are nine source digits for each 10 transmitted digits on the average. Thus, the source entropy and capacity are equal and no errors occur.

Noisy-channel coding theorem

The above example is misleading with respect to the ease with which channel noise can be combated. If the return channel is unavailable, or if the noise is not in known positions (*i.e.*, erasures), the coding and decoding required by the noisy-channel coding theorem are quite complex and become increasingly complex as the error probability is reduced toward zero.

CODING FOR ERROR CORRECTION

Parity checks. Most techniques for coding to reduce errors are based, at least indirectly, on the concept of parity checks. The parity of a set of binary digits is defined as 1 if the set contains an odd number of 1s and defined as 0 otherwise. Thus, the parity of 111 is 1, that of 110 is 0, that of 0001 is 1, etc. A change of any single digit in the set changes the parity of the set. Thus, if a set of digits is transmitted, followed by transmission of the parity of the set (called a parity-check digit), a single error in the transmission can be detected by the disagreement between the parity of the received set and the received parity-check digit. Of course, this disagreement does not indicate which received digit is in error, and no correction is possible.

The following example shows how parity checks can be used to correct errors. Four source digits (s_1, s_2, s_3, s_4) are transmitted, followed by three parity-check digits, p_1, p_2, p_3. The digit p_1 is chosen as the parity of the first three source digits; p_2 as the parity of the first, second, and fourth; and p_3 as the parity of the first, third, and fourth (see 2). Then, for example, the source digits 1001 would be encoded into 1001100. If the first source digit s_1 is received in error at the receiver but all other digits are received correctly, then each received parity check will disagree with the parity of the corresponding set of received source digits. Likewise, if the second source digit is received in error, only the first two parity checks will disagree with the corresponding received parities, because s_2 is not in the set used for the third parity check. Similarly, each source digit and parity-check digit has its own unique pattern of disagreements, allowing any single error in the seven digits to be corrected. Multiple errors unfortunately cannot be corrected in this example.

The code described above is an example of a block code, a code in which the source digits are segmented into sequences of a fixed length (4, for the example), and there is a rule for transforming each source sequence into a sequence of channel digits of fixed length (7, for the example). The block length of the code is defined as the length of the channel sequence, and the code rate (assuming a binary source) is the ratio of source sequence length to channel sequence length. Thus, the code rate for the example is 4/7.

Block length of a code

The coding theorem asserts that there are block codes with code rates arbitrarily close to channel capacity and probabilities of error arbitrarily close to zero. It can be shown, however, that the required block length increases both as the code rate approaches capacity and as the desired probability of error approaches zero.

Codes with high code rate and low probability of error. A great deal of mathematical research has been devoted to finding structures for block codes in which the code rate is relatively close to capacity, the probability of error is small, and the complexity of encoding and decoding is reasonable. There are now many practical encoding and decoding techniques that have been designed for a variety of different channels and that provide a wide range of complexity, block length, and code rate.

An example of coding for error correction can be found in compact-disc high-fidelity technology. First, the audio signal to be recorded is converted to digital data by sampling and quantizing. Successive blocks of the digital data are then encoded by a highly sophisticated code known as a Reed-Solomon code. The encoded signal is then recorded and reproduced on the compact discs. When the signal is played back in a compact-disc player, decoding is performed to reproduce the original signal without errors. In this system the compact-disc storage medium plays the role of the noisy channel.

Another common coding technique occurs in data networks. The data packets traveling through each channel of the network are encoded into a parity-check code as above. Rather than correcting errors at the opposite end of the channel, however, the receiver simply checks to see if errors have occurred. If errors are detected, the receiver requests retransmission of the encoded packet.

Convolutional codes. Most practical applications of coding to correct errors have used convolutional codes instead of block codes. For example, suppose that each source digit entering the encoder is transmitted and then followed by a parity-check digit so that the transmitted stream has the form $s_1\ p_1\ s_2\ p_2\ s_3\ p_3 \ldots$. Suppose that each parity-check digit is the parity of the two preceding source digits (see 3). Thus, if the first four source digits are 1011, the first eight transmitted digits will be *1*0*1*1*1*0 (the source digits are italicized).

$$(2)\quad \begin{cases} p_1 = \text{parity } (s_1, s_2, s_3) \\ p_2 = \text{parity } (s_1, s_2, s_4) \\ p_3 = \text{parity } (s_1, s_3, s_4) \end{cases}$$

$$(3)\quad p_n = \text{parity } (s_n, s_{n-1})$$

$$(4)\quad C = W \log_2 \left(1 + \frac{P}{NW}\right)$$

The effect of noise

If noise alters one of the source digits in transmission over the channel, then the parities will not check for the two succeeding parity-check digits. On the other hand, if noise alters a parity-check digit, only that one parity will fail to check. Thus, any single error (if preceded and followed by at least three correct digits) can be identified and corrected.

In order to attain very low probabilities of decoding error with convolutional codes, the parity-check constraints must extend over more digits. Rather than making p_n the parity of s_n and s_{n-1}, p_n would be the parity of a large set of previous source digits in a fixed set of positions relative to p_n.

Convolutional codes have been used very successfully in telemetering data from space probes. Coding is particularly useful in such applications since it is far less expensive than more traditional techniques for improving communication, such as increasing antenna size and increasing power. Convolutional codes are also widely used for transmitting digital data at very high speeds over the voice channels of a telephone network.

BANDWIDTH-LIMITED CHANNELS

The communication channels considered above were all discrete, involving sequences of symbols entering and leaving the channel. In most physical communication channels, however, the input and output are usually continuously varying signals on a wire, cable, or antenna terminal. The devices that convert the discrete sequences into continuous signals and vice versa are known as modems (digital data modulators and demodulators). The modulator converts the discrete sequence into a signal suitable for the channel, and the demodulator makes decisions on the received combination of signal and noise and produces a discrete sequence as its output.

Use of modems

Because of physical limitations of the transmission medium and because of the need to avoid interference with other channels, the transmitted signal on a channel is frequently restricted to lie within a given band of frequencies W hertz wide. For example, radio stations are restricted in this way, each station broadcasting within an area having one of a nonoverlapping set of frequency bands. A result known as the sampling theorem states that a signal of this type is specified by its values at a sequence of $2W$ appropriately placed sampling points per second. Thus, it may be said that the set of signals limited to a band W hertz wide has $2W$ degrees of freedom per second.

If there were no noise whatever on such a channel, it would be possible to distinguish an infinite number of different amplitude levels for each sample. Consequently, in principle, an infinite number of binary digits per second could be transmitted, and the capacity C would be infinite. In practice, there is always some noise; even so, if no limitations are placed on the transmitter power, P, the capacity will be infinite, because at each sample point an unlimited number of different amplitude levels may be distinguished. When noise is present and the transmitter power is limited, the capacity C becomes finite. This capacity depends on the statistical structure of the noise and the nature of the power limitation.

The simplest, and in some ways the most fundamental, type of noise is white Gaussian noise, such as that produced in an electrical resistor by thermal effects. It has the characteristic of being evenly spread out in frequency and is specified by N, its average power per hertz of frequency. Similarly, the simplest limitation on transmitter power is the assumption that the average transmitter signal power is no greater than P.

Gaussian noise

If the input signal of a channel has a power limitation P and a bandwidth limitation W and if the output signal is the sum of the input signal and white Gaussian noise of power per hertz N, then the capacity of the channel (measured in bits per second) can be shown to be a simple increasing function of W and of the signal-to-noise ratio P/N (see 4). The interpretation of C here is the same as before. By proper design of transmitter and receiver, it is possible to transmit up to C binary digits per second with as small a probability of error as desired, while it is not possible to transmit more than C binary digits per second with an arbitrarily small probability of error.

Many physical channels are subject to noise phenomena quite different from that described above, and the above expression for capacity does not apply to them. The above formula is inapplicable to high-frequency radio links and tropospheric scatter links, for example, but it is applicable to satellite communication and communication from space probes.

Figure 3 illustrates the relationship between capacity and bandwidth and between capacity and P/N. Figure 3 (top) shows that if W is much greater than P/N (called the power-limited region), then capacity is insensitive to changes in W. If W is much smaller than P/N (called the bandwidth-limited region), then capacity is relatively insensitive to changes in P/N (see Figure 3, bottom). Deep space channels are illustrative of the power-limited region, and telephone voice channels are illustrative of the bandwidth-limited region.

The relation between C, W, P, and N (see 4) can be regarded as an exchange relation between the bandwidth W and the signal-to-noise ratio P/N. Keeping the channel capacity fixed, the bandwidth can be decreased, provided the signal-to-noise ratio is sufficiently increased. Conversely, an increase in bandwidth allows a lower signal-to-noise ratio in the channel.

One method of exchanging bandwidth for signal-to-noise ratio is shown in Figure 4. The upper curve represents

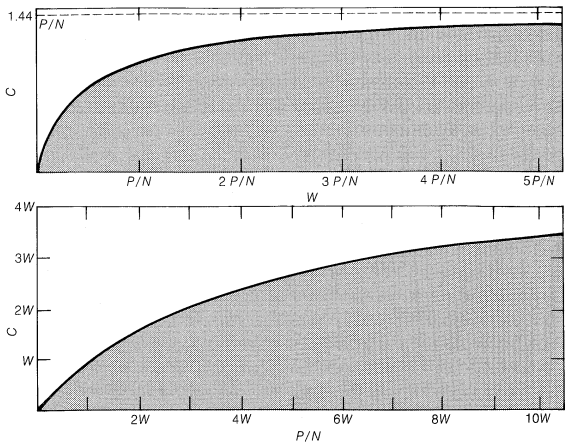


Figure 3: Relationship (top) between bandwidth and capacity and (bottom) between P/N and capacity.

a signal function with a bandwidth such that it can be specified by giving the samples shown. Each sample has four amplitude levels. The lower curve is obtained by combining pairs of samples from the first curve as shown. If the pair of samples from the upper curve have amplitudes x and y, a single amplitude z for the lower curve is computed by adding $4x$ to y. The four possible values of x combined with the four possible values of y produce 16 possible values of z, which must be distinguished. The samples, however, now occur only half as frequently; consequently, the bandwidth is reduced by half, at the cost of increasing the signal-to-noise ratio. Operating this in reverse doubles the bandwidth but reduces the required signal-to-noise ratio.

The simple pulse amplitude modulation technique of Figure 4 is somewhat wasteful of power if low probabilities of error are desired and error-correction coding is not used. For example, to obtain a probability of error of 10^{-5}, pulse amplitude modulation requires about 25 times more power than an ideal coded system in the power-limited region and about six times more power than a system in the bandwidth-limited region.

SOURCES WITH DISTORTION MEASURES

Many sources, such as voice, television, and the measurement of physical characteristics, are not appropriately regarded as being discrete. Serious problems arise in attempts to define the uncertainty or entropy of such sources. For example, if the source output is the speed of an automobile, which is a priori equally likely to be anywhere between 0 and 100 miles per hour (mph), it takes one bit of information to find out whether the speed lies in the interval 0 to 50 mph or 50 to 100 mph. Similarly, it takes two bits to find out in which 25-mph interval the speed lies and n bits to resolve the speed to an interval of 100 times 2^{-n} mph. If n gets increasingly larger, the conclusion is that there is an infinite amount of uncertainty in the speed of the automobile. A natural way out of the dilemma lies in the fact that the uncertainty of the speed depends upon the desired resolution.

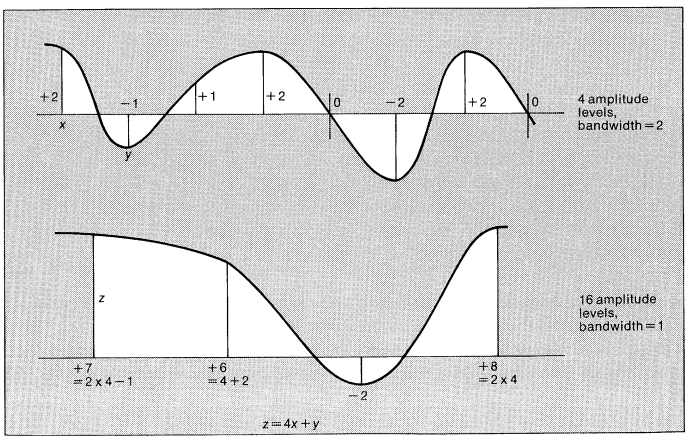


Figure 4: Exchange of bandwidth for signal-to-noise ratio. The amplitude z (in which $z = 4x + y$) on the lower curve is obtained by sampling the two levels x and y on the upper curve whose abscissas lie within the corresponding braces.

As a general approach to problems of this kind a distortion measure $D(u, v)$ is defined. For each possible source output u and each possible signal v used at the destination to represent a source output, $D(u, v)$ assigns a nonnegative number that indicates the cost of representing source output u by signal v. For the automobile example above, u would be the automobile speed and v the approximation to u at the destination; the distortion might reasonably be chosen as the square of the approximation error (see 5). If the source output is a continuously varying signal $u(t)$, and $v(t)$ is an approximation to $u(t)$ at the destination, then the distortion measure is a function of time and might be chosen as before (see 6).

$$D(u, v) = (u - v)^2 \tag{5}$$

$$D[u(t), v(t)] = [u(t) - v(t)]^2 \tag{6}$$

The distortion measure can be chosen to reflect the importance or meaning of the source output to the destination. For example, in speech transmission the ear is relatively insensitive to phase distortion of the speech signal, and thus an appropriate distortion measure would weight amplitude distortion more heavily than phase distortion. With speech the relative importance of intelligibility versus speaker recognition and voice quality can be weighed. The choice of an appropriate distortion measure for a source such as speech is very difficult and thus limits the applicability of information theory.

For any given source and distortion measure, it is possible to encode the source output into a binary sequence and, after storage or transmission of the sequence, to decode the sequence into an approximation of the source output. If, for example, the source output is limited to a band W hertz wide, the encoding could be the reverse of the procedure in Figure 4. The signal could be sampled $2W$ times per second, and the samples could be quantized to a given number of levels, each level having its own binary code word. The approximation to the source would be constructed by mapping the binary sequences back into sample levels and the sample levels back into a time-varying signal. The approximation would differ from the source output because of the error introduced in quantizing.

The rate R of a source relative to a distortion level D is defined as the smallest number of binary digits per unit time required to encode the source output, subject to the restriction that the average distortion between the source output and the final approximation be at most D. With the above definition, the rate of the source becomes a function, $R(D)$, of the allowable average distortion. The larger the allowable distortion D is, the smaller the required number of binary digits is.

Finding the best way to encode a source into a binary sequence for a given allowable distortion is usually at least as difficult as finding good error-correction techniques for noisy channels. Because of this the above definition is not very helpful in calculating the function $R(D)$. Instead $R(D)$ is found by the following procedure. A hypothetical channel is considered between the source and the destination. Such a channel is defined by the conditional probabilities of outputs given inputs, and corresponding to each such channel there is an average transmitted information per unit time and an average distortion per unit time. It can be shown that $R(D)$ is the smallest average transmitted information per unit time over all such channels for which the average distortion per unit time is at most D.

The fundamental theorem concerning sources and distortion measures can be stated loosely as follows. For a given source and distortion measure, $R(D)$ is the rate in

The fundamental theorem for sources and distortion measures

bits per second of the source relative to a distortion level D. The source output is to be transmitted over a channel of capacity C bits per second, and D_c is the distortion level for which $R(D_c) = C$. Then, independent of the transmitter and receiver, the average distortion between source output and destination must be at least D_c. Conversely, it is possible to construct transmitters and receivers so that the average distortion is as close to D_c as desired. Finally, the transmitter can first encode the source output into less than C binary digits per second and then encode the binary digits for transmission over the channel in such a way that the resulting average distortion is as close to D_c as desired.

Because the minimum average distortion with which a source can be transmitted depends only on the capacity of the channel and not on any other channel characteristics, this theorem gives considerable force to the notion that the capacity is indeed a fundamental parameter of a channel. The theorem also points out that the current trend toward using binary data as a common interface between sources and channels (thus separating the problem of source processing from channel error correction) does not lead to any fundamental degradation in performance.

The theorem does not show how to construct transmitters and receivers that achieve a distortion close to the theoretical minimum. Its major use has been as a standard against which to compare existing systems, thus indicating where major improvements are possible.

DETECTION, ESTIMATION, AND PREDICTION

Another type of problem that has been studied extensively in the field of information theory is that of determining the best devices for extracting a message from a noise-corrupted signal. If the message represents a selection from a discrete set of choices, then the process of extracting the message from a noisy signal is known as detection. If the message represents the specification of one or more parameters over a continuous range of values, then the extraction of the message from a noisy signal is called estimation. More generally, in estimation, the message might be a continuously varying function of time, or even a function of time and space. Prediction is an important special case of estimation in which the message of interest is the value of a random signal at some future time (see AUTOMATA THEORY). This future value is to be estimated from the portion of the random signal (perhaps noise-corrupted) that has occurred up to the present time.

Detection problems. A few of the many areas in which detection problems arise are as follows: receivers in digital communication systems, radar receivers, sonar receivers, seismic detectors, oil-field exploration, and radio astronomy. In many of these areas there are related estimation problems. If, for example, a person uses radar to determine whether there is an object in a particular region of the sky, he has a binary detection problem. If he also wants to know the location, size, and velocity of the object, it becomes an estimation problem, and a different type of radar system is needed.

Conceptually, detection problems are most naturally treated in what is called a Bayesian formulation. Here one assumes that there are n possible messages, numbered $0, 1, \ldots, n-1$, and that these messages occur with known probabilities, $p_0, p_1, \ldots, p_{n-1}$. Conditional on each message there is a known probability distribution on the received noise-corrupted signal. Finally, there is a cost C_{ij} associated with detecting message j if message i is actually present. Within this formulation the optimum detector can then reasonably be defined as the detector that minimizes the expected cost of incorrect decisions. Such an optimum detector can be implemented by just computing a set of $n-1$ numbers, called likelihood ratios, and then making a decision based on the relative sizes of the likelihood ratios. The ith likelihood ratio is defined as the ratio of the probability of the received noisy message conditional on message i to the probability of the received noisy message conditional on message 0. These likelihood ratios do not depend on the probabilities with which the messages occur or on the cost function C_{ij}. For this reason the likelihood ratios are also fundamental in other formulations of the detection problem where one is unwilling or unable to make assumptions about the probabilities of the messages or about the cost function.

Likelihood ratios

There is one important situation in which the likelihood ratios are quite easy to calculate. It can be assumed that there are n known signals, $S_0(t), S_1(t), \ldots, S_{n-1}(t)$, corresponding to the n messages. Further, the received noisy signal when message i occurs can be taken to be $S_i(t) + n(t)$, in which $n(t)$ is Gaussian noise. Then each likelihood ratio can be found by passing the received noisy signal through an appropriately chosen linear filter and sampling the output at a given time. More complex detection problems can be handled by extensions of this basic technique.

Estimation and prediction problems. A few of the many areas in which estimation and prediction problems arise are as follows: control, analog communication receivers, radar receivers, sonar receivers, seismic array processors, and radio astronomy. As with detection, the simplest conceptual formulation for estimation is the Bayesian formulation, in which one assumes complete probabilistic descriptions of both messages and received signals. If a cost function, such as the squared value of the error, is specified between message and message estimate, then an optimum estimator can be defined as the estimator that minimizes the expected cost.

A classic case of the above formulation is that in which the message and the noise are jointly Gaussian processes and the received signal is message plus noise. The optimum estimator, with an arbitrary delay or prediction interval, was found independently in the United States and the Soviet Union by the mathematicians Norbert Wiener and A.N. Kolmogorov and turned out to be an appropriately chosen linear filter. A later approach by the American mathematicians Rudolf E. Kalman and Richard S. Bucy turned out to have significant advantages in the many applications in which the processes are nonstationary. Curiously enough, the same solution arises if the message and noise processes are not assumed to be Gaussian but the estimator is restricted to be linear.

Aside from various extensions of the above case, most estimation problems cannot be solved exactly, and one must be content with approximate solutions and with choosing an estimator structure on intuitive grounds and then optimizing some parameters of the structure.

CRYPTOGRAPHIC, LINGUISTIC, AND OTHER APPLICATIONS

Some applications have been made in cryptography and linguistics. It is possible to formulate a theory of cryptography, or secrecy systems, in terms of the concepts occurring in information theory. When this is done, it appears that the information rate R of a language is intimately related to the possibility of solving cryptograms in that language. The smaller this rate, the easier such a solution becomes and the less material is necessary to render such a solution unique. Indeed, within limits, the theory becomes quantitative and predictive, giving means of calculating how much material must be intercepted in a given language and with a given cipher in order to ensure the existence of a unique solution.

Studies have been made of the distribution of lengths and frequency of occurrence of the different words in a language such as English. It has been found, for example, that the relative frequency of the nth most frequent word may be expressed quite closely by a formula of the type $P(n+m)^{-b}$, with suitable constants for P, m, and b. Experimental data of this type can be explained as consequences of the assumption that a language gradually evolves under continued use into an efficient communication code.

Psychologists have discovered interesting relationships between the amount of information in a stimulus and reaction time to the stimulus. For example, an experiment can be set up in which there are four lights and four associated push buttons. The lights go on in a random order and the subject is required to press the corresponding button as quickly as possible after a light goes on. It is found that the average time required for this reaction increases linearly with an increase in the amount of information conveyed by the lights. This experimental result holds true under a wide variety of changes in the experiment: the

number of lights, the probabilities of different lights, and even varying correlations between successive lights.

The results suggest that under certain conditions the human being, in manipulating information, may adopt codes and methods akin to those used in information theory.

BIBLIOGRAPHY. Shannon's original papers first appeared in book form in CLAUDE E. SHANNON and WARREN WEAVER, *The Mathematical Theory of Communication* (1949, reprinted 1975). The major papers in the development of information theory from 1948 to 1972, including Shannon's original work, are collected in DAVID SLEPIAN (ed.), *Key Papers in the Development of Information Theory* (1974). Other early influential works are NORBERT WIENER, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications* (1949, reprinted 1966); V.A. KOTELNIKOV, *The Theory of Optimum Noise Immunity,* trans. from Russian (1959); ALEKSANDR KHINCHIN, *Mathematical Foundations of Information Theory,* trans. from Russian (1957); AMIEL FEINSTEIN, *Foundations of Information Theory* (1958); ROBERT M. FANO, *Transmission of Information: A Statistical Theory of Communications* (1961); JACOB WOLFOWITZ, *Coding Theorems of Information Theory,* 3rd ed. (1978); and M.S. PINSKER, *Information and Information Stability of Random Variables and Processes,* trans. from Russian (1964).

Elementary texts on information theory include RICHARD W. HAMMING, *Coding and Information Theory,* 2nd ed. (1986); NORMAN ABRAMSON, *Information Theory and Coding* (1963); MASUD MANSURIPUR, *Introduction to Information Theory* (1987); and JOHN R. PIERCE, *An Introduction to Information Theory: Symbols, Signals & Noise,* 2nd rev. ed. (1980). More advanced and comprehensive general works include ROBERT G. GALLAGER, *Information Theory and Reliable Communication* (1968); ROBERT J. MCELIECE, *The Theory of Information and Coding* (1977, reissued 1984); IMRE CSISZÁR and JÁNOS KÖRNER, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (1981); and ANDREW J. VITERBI and JIM K. OMURA, *Principles of Digital Communication and Coding* (1979).

Coding and decoding for error correction are treated in the above-mentioned general texts and in greater depth in RICHARD E. BLAHUT, *Theory and Practice of Error Control Codes* (1983); SHU LIN and DANIEL J. COSTELLO, JR., *Error Control Coding: Fundamentals and Applications* (1983); ELWYN R. BERLEKAMP, *Algebraic Coding Theory,* rev. ed. (1984); GEORGE C. CLARK, JR., and J. BIBB CAIN, *Error-Correction Coding for Digital Communications* (1981); and W. WESLEY PETERSON and E.J. WELDON, JR., *Error-Correcting Codes,* 2nd ed. (1972). Sources with a distortion measure are covered in TOBY BERGER, *Rate Distortion Theory: A Mathematical Basis for Data Compression* (1971); TOBY BERGER and LEE D. DAVISSON, *Advances in Source Coding* (1975); and in the above-mentioned general texts. Treatments of detection, estimation, and prediction can be found in HARRY L. VAN TREES, *Detection, Estimation, and Modulation Theory,* 3 vol. (1968–71); JAMES L. MELSA and DAVID L. COHN, *Decision and Estimation Theory* (1978); and HARRY STARK and JOHN W. WOODS, *Probability, Random Processes, and Estimation Theory for Engineers* (1986).

The broader aspects of information theory and closely related fields are well presented in FRITZ MACHLUP and UNA MANSFIELD, *The Study of Information: Interdisciplinary Messages* (1983). Other works applying information theory to various fields include JACOB MARSCHAK, *Economic Information, Decision, and Prediction,* 3 vol. (1974); WENDELL R. GARNER, *The Processing of Information and Structure* (1974); WERNER HOLZMÜLLER, *Information in Biological Systems* (1984; originally published in German, 1981); SERAFIN FRAGA, K.M.S. SAXENA, and MANUEL TORRES, *Biomolecular Information Theory* (1978); and KAREL ECKSCHLAGER and VLADIMÍR ŠTĚPÁNEK, *Information Theory as Applied to Chemical Analysis* (1979), and *Analytical Measurement and Information: Advances in the Information Theoretic Approach to Chemical Analyses* (1985).

(R.G.G.)

Inheritance and Succession

Inheritance, in law, signifies the devolution of property upon the death of its owner. In modern society this devolution is regulated in minute detail by law. In the civil law of the continental European pattern the pertinent branch is generally called the law of succession. In Anglo-American common law it was customary to distinguish between descent of real estate and distribution of personal estate. The rules applicable to the two kinds of property have been fused, but no common, overall name is yet universally accepted. In England, books dealing with the subject are varyingly entitled *On Wills, On Probate, On Succession,* or *On Executors and Administrators.* In the United States, the term probate law is frequently, although inaccurately, applied to the field as a whole. Following the title of an important statute of the state of New York, another term, law of decedents' estates, has been gaining ground, as has the law of succession.

This article is divided into the following sections:

Inheritance and property rights 575
  Inheritance and individual ownership of property 575
  Critiques of inheritance 576
Prime issues in inheritance and succession 576
  Freedom of testation 576
  Divided or undivided inheritance 578
Intestate succession 579
  Historical development 579
  Intestacy in present law 580
Wills 582
  Formalities of wills 582
  Invalid wills 583
The machinery of transfer 583
  Transfer in civil law 583
  Transfer in common law 583
Bibliography 584

Inheritance and property rights

INHERITANCE AND INDIVIDUAL OWNERSHIP OF PROPERTY

Property in primitive and modern societies

Inheritance of property cannot occur unless goods are regarded as belonging to individuals rather than to groups and unless the goods are of such permanence that they continue to exist and to be useful beyond the death of the owner. Among primitive food-gatherers and hunters it has not been uncommon for such personal belongings as weapons or bowls to be destroyed after the death of the owner in order to protect the survivors from being molested by his spirit. Among the Papua of New Guinea and the Damara (Bergdama) of South Africa the hut of the dead man was abandoned or burned down so as to ban the magic of the disease of which the owner had died. Among the Herero of southwest Africa the dead man's goats were slaughtered and eaten; this custom seems to have been connected with the fear that they were affected by his magic but also with the belief that the spirits of the slaughtered goats would follow the dead owner into the realm of spirits, where he would need them. Belief in providing for the needs of the dead seems to have been the root of the widespread custom of burying with the body or burning victuals, utensils, treasure, slaves, or wives. Tombs have yielded rich evidence of such practices in the cultures of the Stone and Bronze ages as well as in the high civilizations of ancient Egypt or pre-Columbian Mexico. Another way of disposing of a dead man's effects was to distribute them among remote relatives and friends, as in American Indian tribes such as the Delaware and the

Iroquois; distribution of this sort, in the absence of rules of inheritance, could easily lead to quarrel and violence, as was often the case among the Comanche Indians.

The view of some Marxist writers that common ownership of all goods, or at least of land, was once universal among mankind can be neither proved nor disproved. Group ownership has been widespread but by no means universal among primitive and archaic agriculturalists. It has, indeed, persisted into modern times in India and parts of Africa and Asia, and it played a considerable role in the development of the Teutonic and Slavic peoples of Europe. In parts of Yugoslavia, ownership of the land by *zadrugas*—*i.e.*, large groups of progeny of a common ancestor—continued into the 20th century. In western Europe the common ownership of pastures and woods, which grew out of the former system of common ownership of the land of a village, can still be found, especially in the Alpine regions of Switzerland and Austria. While in earlier times colonization of new land tended to be carried on by groups—for instance, the German settlement of the regions east of the Elbe in the 10th to 13th centuries—the Europeans who settled in North America, Australia, South Africa, and other parts of the world in the 18th and 19th centuries regarded individual ownership of land as most favourable to efficient use. In the 20th century, Socialist ideas, combined with large-scale mechanization, have resulted in new forms of land ownership in common: the *kolkhozy* of the Soviet Union, the communes of Communist China, and the *kibbutzim* of Israel. Wherever land is held in common, the death of a member of the group results not in inheritance but in a rearrangement of duties and of rights of participation in the produce of the land or rights of temporary usage of the land itself.

CRITIQUES OF INHERITANCE

The institution of inheritance has been criticized because it renders possible the acquisition of wealth without work and because it is regarded as a principal source of economic inequality. Such attacks have come not only from radicals to whom complete equality of income appeals as a social ideal but also from more moderate thinkers to whom great differences in the distribution of wealth appear to be incompatible with modern views of the dignity of man. In response to their criticisms, inheritance has been defended on economic as well as on moral grounds.

Inheritance defended

Inheritance has been said to be necessary within the framework of an economy of individual property to guarantee the continuity of enterprise without which long-range economic activity could not flourish. This argument has lost much of its force as large-scale enterprise has come to be carried on in corporate form and thus to be directed not by owners but by specialists in management who succeed each other in the manner of officeholders. There is, however, still force in the argument that without the incentive of handing on the fruits of one's work, competition, and consequently the functioning of the total economy, would be hampered.

It is possible to conceive of a social system in which property rights would end with the owner's death. If the assets left behind were not reassigned to some other individual, the eventual result would be complete ownership of all wealth by the community, and the system of individual property would end. A new individual owner could be determined in one of four ways: ownership by the first taker, a practice that would produce strife and disorder; reassignment by a governmental agency, which would constitute an exercise of power regarded as dangerous in a free society; reallotment in accordance with settled rules generally fixed for all; or reallotment in accordance with the wishes of the decedent. The last two are the ways in which the modern systems of inheritance work: the estate is reallotted according to the rules of intestacy law or according to the will of the decedent.

The only debatable issues within a system of private ownership are: who are to be the takers in intestate succession; and whether or not and within what limits freedom of testation shall be permitted. In all societies, inheritance has developed as an incident of kinship. Even in a society in which ownership of property is regarded as belonging to individuals rather than kinship groups, the feeling of belonging to one's group is still so strong, especially between parents and children, that a person's sense of freedom would not be complete unless he knew that he could pass on his possessions to his children. The question arises, however, whether inheritance shall extend beyond the circle of those persons with whom the decedent was connected by ties of affection or about whose well-being he or she was, or should have been, concerned. In the urbanized, mobile population of highly industrialized nations, the family as a felt unit has tended to shrink to the small circle of husband, wife, and children. Ties of relationship tend to be weak even among first cousins.

Inheritance law as an instrument of social policy

In an age of expanding demands on government, there has been an inclination to let the estate of a person dying intestate pass to the public treasury rather than go to enrich distant relatives. In England, the circle of intestate takers has been limited by the Administration of Estates Act of 1925 to relatives no more remote than the grandparents, uncles, and aunts of the deceased. Even more restrictive than those of England are the intestacy laws of the Soviet Union and other Communist countries. Under the law of the U.S.S.R., intestate succession does not extend beyond descendants, the surviving spouse, grandparents, brothers, sisters, and incapacitated persons who were dependent upon the decedent for at least one year prior to his death.

Another way of limiting the rights of remote relatives for the benefit of the public treasury consists in increasing the rates of inheritance taxes in proportion to the remoteness of the relationship between the takers and the decedent. In the United States, although the federal tax on succession depends solely on the size of the estate, the additional inheritance taxes levied by the states are widely patterned upon the closeness of relationship. This method is also employed in numerous other countries but not, since 1949, in England.

Inheritance law is also used to reduce inequalities in the distribution of wealth. This may be done by compulsory partitions, as under the laws of the French and the German pattern, or by means of progressive inheritance taxation, as in the United Kingdom or the United States, or by a combination of both. The law of inheritance and inheritance taxation is thus an instrument of social policy.

An impressive illustration of the way in which the law of inheritance serves social policy is the inheritance legislation of the Soviet Union. In the early stage of the Bolshevik Revolution, inheritance was limited to the descent of a modest amount of property to close relatives or to the surviving spouse, provided they were in need. The limit upon the amount of the property was lifted in 1926; and the requirement of need was also abolished for inheritance by the surviving spouse, descendants, parents, grandparents, brothers, and sisters. Under the civil code of 1922, the power of testation was limited to increasing or reducing the share of particular intestate successors. After 1961 property could be left to any person. Private property cannot exist in the means of production, and therefore inheritance is limited to goods of use or consumption and to savings accounts. Within the limits stated, inheritance and freedom of testation are regarded as constituting useful incentives to productivity without constituting a danger to the Socialist system. Inheritance of private property was thus listed in the constitution of 1936 and reaffirmed in the constitution of 1977 as one of the rights of citizens.

Prime issues in inheritance and succession

In a society in which inheritance exists, two issues are of prime importance for the distribution of wealth and for the social and political structure of the society: the issue of the extent to which owners of property shall have the power by their own decision to determine the course of inheritance, and the issue of whether or not estates shall be allowed or even required to pass undivided to one single heir.

FREEDOM OF TESTATION

The power of an owner of property to determine who is to have it upon his death is thought to stimulate economic

The question of the property owner's power

activity: it is also considered desirable that a property owner be allowed to modify the rigid rules of the intestacy laws so as to adapt them to the particular situation of his family by preferring, for instance, a disabled child over one of proven capacity. The freedom to disinherit a child may be used to induce filial obedience, but freedom of testation also implies the freedom of making provision for charity. The possibility of abuse for ends of spite, arbitrariness, or whimsy is the price society has to pay for such power. Freedom of testation developed slowly, and nowhere does it exist without limitations. The questions of what the limits shall be, especially to what extent an owner of property shall be free to disinherit close members of his family and to what extent he shall have the power to tie up property from beyond the grave, have been answered in widely diverse ways.

Historical development. In a primitive or archaic society in which property is owned by the kinship or neighbourhood group rather than by individuals, freedom of testation cannot exist. Transition from group to individual ownership has rarely if ever occurred in one single step. As to land, even when its use was regarded as rightfully belonging to an individual, its free alienation by sale or gift, and even more so by will, was for long periods hedged in by superior rights of the kinship group, the village, or the feudal lord. Transition to free alienation has often been achieved by means of subterfuge, such as the adoption of the "purchaser" or "devisee" as a son, or, once free alienation had become possible inter vivos (between living persons) but not yet upon death, by fictitious sale or gift to a middleman who would promise to let the grantor keep the property as long as he should live and upon his death to deal with it as directed by the grantor. Such use of adoption occurred in ancient Babylonia, China, Japan, India, and other societies of an archaic patriarchic order. In ancient Greece effects similar to those of a will were achieved by gift, to take effect upon the death of the donor or, where the only child of the family was a daughter, by giving her in marriage together with the estate. Transfer by use of a middleman became possible among the Germanic peoples following the decline of the Roman Empire.

In ancient Rome the institution of the will appeared at an early stage of cultural development, but there, too, it seems to have been preceded by a stage in which its effects could be achieved only by indirection. The so-called will made in assembly (*testamentum comitiis calatis*) seems to have been the approval by the assembly of the adoption of a son by the childless chief of an aristocratic house so that the house and the worship of its deities would be perpetuated.

By the 5th century BC the head of a Roman family seems to have been able during his lifetime to achieve the purposes of a testamentary transaction by fictitious sale to a middleman, *familiae emptor* (purchaser of the family property). In the period of the early principate (1st century AD), the testament was fully recognized in its proper sense. In the mature form in which it is dealt with in the Corpus Juris (6th century AD), it became in the late Middle Ages the model for continental Europe.

Among the Anglo-Saxons and other Germanic peoples, land was subject to ties of the kinship group and, later, of feudalism, so that there was no place for disposition by will. Chattels were more freely alienable. In establishing freedom of testation a prominent role was played by the church, which desired thereby to obtain funds for its activities, which included the bulk of medieval education, charity, and cultivation of the arts. In England the church succeeded shortly after the Norman Conquest in establishing the jurisdiction of its courts for matters concerning succession upon death to personal property. Through the church the will of the Roman pattern became firmly institutionalized, but a testator still had to leave a "reasonable part of the estate" (ordinarily at least one-third) to his wife and children.

Once the alienation of real property had again become possible by gift or sale, there grew up all over Europe that same practice of indirectly achieving the effects of a will by fictitious grant to a middleman (German *Salmann*, "sale man"; English feoffee to uses) that, in analogous circumstances, had grown up at other times and places. On the Continent, the will as such became again available when Roman law was rediscovered and "received," which occurred from the 11th century onward, first in Italy and then north of the Alps. In France and Germany the will of the Roman pattern was fully recognized in the late 15th century. Just about that time, however, the enfeoffment to uses, which had been popular in England, was abolished by Henry VIII's Statute of Uses in 1535. The King wished to restore to the crown its prospects of escheat and of certain feudal duties, which could be evaded by the alienation to uses. Public indignation was so strong, however, that five years later the King found it advisable, by the enactment of the Statute of Wills, to open the way for true testamentary disposition of land. Restrictions limiting devises of those lands of which ownership was connected with the duty of rendering military service were abolished at the time of the Restoration by the Military Tenures Act of 1662. In Scotland, testamentary disposition of land remained precarious until the enactment of the Titles to Land Consolidation Act in 1868.

Limits on freedom of testation. Freedom of testation has never been absolutely unlimited. Nowhere is a testamentary provision valid if its enforcement would be shocking to public morals. When a testamentary gift is conditioned upon an act of the beneficiary that in good morals should not be so conditioned, as, for instance, a gift conditioned upon the beneficiary's changing his religion, the gift is either invalid or valid unconditionally. Generally, property given by testament cannot be tied up by the testator for an indefinite future. Under the rule against perpetuities, as developed in England and commonly applied in the United States, a testator may leave property to a person for life and upon the first taker's death to some other person; but the last "remainder" must "vest" not later than, roughly speaking, one generation after the testator's death or, in England, since the Perpetuities and Accumulation Act of 1964, a fixed period of years up to 80. In the civil-law countries of the German system, the freedom to provide for substitutions is limited in similar ways, but in those of the French system it is limited much more strictly.

A testator's freedom to disinherit a surviving spouse, children, or other heirs has been more extensive in ancient Roman and modern Anglo-American law than in the modern civil-law countries, but it has always had limits. In republican Rome a testator had the power to disinherit a spouse and children, but if he wished to do this he had to say so expressly in the will. In the period of the principate (27 BC–AD 284), it became necessary to state the reasons, because a will disinheriting a close member of the family without reasonable and honest cause was in danger of being declared invalid. In the late Roman Empire the descendants—and if there were no descendants, the ascendants (*e.g.*, parents)—were given the right to a share in the estate (*pars legitima*), of which none of them could be deprived except upon serious cause stated in the will. When, after the fall of the Roman Empire, testamentary disposition came to be recognized again in the later Middle Ages, custom generally required that some minimum share, frequently one-third, be left to the surviving spouse, or the descendants, or both. Upon the revival of Roman law on the European continent and in Scotland, these customs were in various ways combined with the rules of the Corpus Juris.

Protection of the rights of heirs

In the modern civil law, two systems are used to provide protection against disinheritance. Under the French system, a testator who is survived by descendants, parents, or (in some countries) brothers, sisters, or even other close relatives, cannot dispose at all of the "reserved portion" of his estate, the size of which depends upon the number and the degree of nearness of relationship of the surviving "forced heirs." Under the civil code of France, for instance, donations inter vivos or by last will cannot exceed one-half of the property of the disposer, if he leaves at his decease one child; one-third, if he leaves two children; and one-fourth, if he leaves three or a greater number. The indisposable share is one-half of the property if the disposer, having no children, leaves ascendants of both

his father's and his mother's lines; and three-quarters if he leaves ascendants in only one line. Under the German pattern, the surviving spouse, a descendant or, if there are no descendants, a parent can claim to be paid in money one-half the value of the share that would have been his in the case of intestate succession.

In England those customs that required a minimum share in the personal property to be left to the surviving spouse and descendants disappeared in the 17th and 18th centuries. The interest of dower, which guaranteed a life estate to the widow in one-third of each parcel of the real estate of the predeceased husband, lost its protective effect in 1833. At the turn of the 20th century, freedom of disinheritance was complete in England as well as in the dominions but not in Scotland. There, in the movable estate, the legitim (bairn's part) is still reserved to the children, the *ius relicti* to the widower, and the *ius relictae* to the widow. Until 1964 (in immovables) the widower was entitled to curtesy, a life rent in his wife's heritage (*i.e.*, immovable) property, and the widow had the right of terce—*i.e.*, a life rent out of one-third of her husband's inheritable estate. In England, freedom of testation, while unlimited by law, was kept within narrow limits by the custom among wealthy families of preventing the splitting up or alienation of the family wealth by means of a so-called strict settlement. In each generation, the head of the family would settle the estate upon the eldest son in such a way that it would descend to him undivided but subject to a generous life estate for the widow and to provisions for the daughters, younger sons, and other needy relatives.

In the different social climate of New Zealand, a new device for protecting needy family members against disinheritance was invented with the enactment, in 1900, of a statute that empowers the court to order adequate provision for the maintenance of a spouse or a needy child out of the estate of any testator who has not made such provision. Family provision acts of this kind have since been enacted in Australia, Canada, and England.

Under the English Inheritance (Family Provision) Act of 1938, as amended in a series of enactments, the court, if it found the decedent had failed to make reasonable financial provision for the applicant, was empowered to order maintenance from the estate to the surviving spouse, an unmarried daughter, a minor son, any incapacitated child, or an unmarried former spouse of the decedent. The scope of this system of discretionary financial provision was extended by the Inheritance (Provision for Family and Dependents) Act of 1975. Under that act, the standard for provision for a surviving spouse is no longer limited to maintenance but is a reasonable share of the deceased's estate. The class of applicants has been widened to include any person treated by the deceased as a child of the family and any person who was being wholly or partly maintained by the deceased immediately before his death.

In the United States the surviving spouse is protected against complete disinheritance in every state through one or more of the following devices: dower, indefeasible share, community property, homestead, or family allowances. The most widespread is the indefeasible share, which guarantees to the surviving spouse a certain portion, usually expressed in terms of a fixed dollar amount plus a fraction, or, under older statutes, as just a fraction, of the decedent's estate. The weakness of this system, however, is that the indefeasible share can be diminished or wiped out if the decedent has given away most or all of his property before his death. A number of states have tried to remedy this difficulty by permitting the surviving spouse's rights to be asserted against certain inter vivos transfers.

In many states the indefeasible share system exists alongside a modernized version of the old common-law estates of dower and curtesy, which have now been generally assimilated to each other under the single heading of dower. Under some statutes each spouse's dower rights attach upon marriage to any real estate owned by the other spouse and upon acquisition to any real estate acquired by the other spouse during the marriage. These rights cannot be affected during the marriage by any transaction of the owner-spouse without the other's consent. Upon the death of the owner-spouse, dower entitles the surviving spouse to a life estate in all or part of the real estate of the predeceasing spouse. Dower has long ceased to be the major device for protecting a surviving spouse against disinheritance because of the fact that it applies only to real estate and thus offers no security at all in the situation where the wealth of the predeceasing spouse was only or mainly composed of personal property such as savings or shares of stock. A further reason for the decline of dower is that a system of marital rights in real estate that cannot be defeated by sale, gift, mortgage, or will of the owner-spouse came to be seen as a clog on marketability and a threat to the security of titles. Thus, several states have followed the example of England and have abolished dower altogether, while a number of others have redefined dower as an interest that attaches only to whatever real estate is left upon the death of the predeceasing spouse.

Community-property systems

In those U.S. jurisdictions that have adopted the so-called community-property system, an indefeasible share in the family wealth is secured to the surviving spouse by his or her being entitled to one-half of the community property, which generally consists of the property acquired during the marriage by the gainful activities of either spouse. Varying systems of community property also exist in numerous European and Latin-American countries. In the countries of the French system, community-property law applies unless it has been expressly contracted out by the parties to the marriage. Under the Scandinavian and West German systems the assets of husband and wife remain separate during marriage but upon the termination of the marriage are distributed between them. Protection of the surviving spouse can, furthermore, be achieved through homestead laws and family allowance laws that guarantee to the widow or the widower an award of income payable out of the estate for a few months immediately following the death of the other spouse.

The only jurisdictions in the United States that protect descendants against disinheritance by giving them indefeasible shares are Louisiana and Puerto Rico, whose legal systems are not derived from the common law. In the other states the descendants are protected either not at all or only indirectly and incompletely by (1) "pretermitted heir" statutes, which, like early Roman law, require the testator to state the disinheritance of a descendant expressly in the will; or (2) "afterborn heir" statutes, under which a child born after the making of the will receives his intestate share unless a contrary intention is stated in the will; or (3) "charity begins at home statutes," under which no more than a certain fraction (*e.g.*, one-half) of the estate may be given to charity by a testator who is survived by certain close relatives; or (4) "hellfire statutes," which declare ineffective a testamentary provision for charitable purpose made by the testator upon his deathbed, in his last illness, or within a fixed period immediately preceding his death.

In the Soviet Union a compulsory share of one-third of his intestate share is guaranteed to each minor child of the decedent and to any of the following who are unable to work: the decedent's children, spouse, parents, and those who were dependent on him.

DIVIDED OR UNDIVIDED INHERITANCE

Like the problem of whether and to what extent freedom of testation shall be permitted, the question of whether a person's estate may pass undivided to one person or whether it should be divided among several takers has significant political implications. The issue has been especially important in the history of Anglo-American law, where it is usually referred to as the problem of primogeniture. The term is too narrow, however, because the sole heir need not necessarily be the first-born son (*primogenitus*). Under the system of ultimogeniture, which existed in parts of England as the custom of Borough English, and also under the German National Socialist law of 1933, the person favoured was the youngest son; under systems of seniorate or juniorate it is the oldest or youngest member of the family; under that of majorate or minorate, it is the oldest or the youngest person standing in equal degree of consanguinity to the decedent. There have also been cases

Primogeniture

where certain lands have been reserved to the second-born son and his line (*secundogeniture*) or the third-born and his line (*tertiogeniture*), etc.

In England, undivided inheritance was applied to real but not to personal property. The distinction between the two kinds of property was important in the struggle for power between church and state. In medieval England the organization of society in general and of the army and the public offices in particular was based upon the distribution of the ownership of the land, over all of which the king was lord paramount. The church, on the other hand, concerned itself with divine worship, the care of the sick and poor, and the cultivation of learning and the arts. After the Norman Conquest a compromise was worked out between the king and the church under which the royal courts exercised jurisdiction over real property, while succession to personal property was to be the concern of the ecclesiastical courts. Until 1926 descent to real property thus was subject to rules different from those applying to the distribution of personal property. For the former, the common-law courts developed a system that tended to maintain the existing military and social order through unpartitioned descent of land to one heir rather than division among several coheirs and, for a long time, by reluctance in admitting freedom of testation.

As to personal property, however, the ecclesiastical courts favoured a freedom of testation that allowed a decedent to leave part of his property to the church for the promotion of its manifold activities. In case of intestacy, the church favoured distribution among family members of equal nearness to the decedent. It applied rules, similar to those laid down in the 6th century by the Roman emperor Justinian (see below), that, in 1670, were fixed in the Statute of Distribution. The problem about which the two sets of courts differed—namely, whether an inheritance should be split up among several coheirs or pass undivided to a single heir—has, of course, not been limited to England. Unpartitioned inheritance has occurred in the most diverse civilizations—among the Khoikhoin of southwest Africa, the Maori of New Zealand, the inhabitants of the Tonga Island, in parts of China and Siberia, and in western Europe.

Why and how primogeniture became the common-law system of inheritance of freehold real estate is not clear. Primogeniture obviously served the needs of feudalism, in which the ownership of a parcel of land tended to be connected with a public office or with military duties that could not be well divided among several people. Partition was also likely to result in confusion regarding the services the peasants were bound to render to the landlord. At the peasant level, primogeniture prevented holdings from being split up until they were too small to allow a family to make a living. Attempts to avoid physical partition by selling the land and dividing the proceeds were impracticable in a society in which it was considered important to preserve family ownership of the farm and in which money was not readily available.

In spite of these circumstances, undivided descent of land to one heir never was the exclusive system among the European peasantry. It became, however, the almost universal system among the nobility, who were anxious to preserve intact the family wealth. In order to achieve this purpose it became necessary, after alienability of land and freedom of testation had developed, not only to establish unpartitioned descent as the rule of intestate succession but also to "entail" the land—*i.e.*, to prevent the owner from selling, giving away, or encumbering the land as well as from disposing of it by will. In England varying legal devices were used from the 13th century on. After the 17th century the so-called strict family settlement became the principal device, while on the Continent the *fideicommissum* of late Roman law was adapted to serve the purpose. The political power secured in this way to the nobility and gentry enabled it, as the necessary counterpart of primogeniture, to secure for the younger sons the lucrative positions in the church, the army, and, on the Continent, the expanding bureaucracy. In the 18th century this system was attacked both by the supporters of democratic ideals and by the economists of the classical school. To the latter, entails were objectionable because they not only stood in the way of mortgaging the land for purposes of improvement but also because inalienability prevented its coming into the hands of the most efficient cultivator. The system was first destroyed in the British colonies in New England, and in the course of the American Revolution it was swept away in the other states. In Europe, it collapsed during the French Revolution, and in the Napoleonic Code care was taken to prevent its reestablishment. Not only were all descendants or other relatives of equal degree to take part equally in intestate succession but also, by giving each child the right to a minimum share, a testator was prevented from giving all to one child.

Entailing the land

In England the main object of the economists' attack upon settlements was removed when, through a series of statutes, settled lands were gradually restored to the market; life tenants were thereby given the power under certain circumstances to mortgage or sell the land. Stocks and bonds, which had become a form of wealth more important than land, could, and still can be, tied up by means of a trust; but through the impact of heavy death duties this power has now been restricted. Primogeniture as the rule of intestacy was finally abolished by the Administration of Estates Act in 1925.

On the Continent, equal division among descendants and other relatives in equal degree became the general rule in the codes of the 19th century, but in certain countries, especially Austria and Germany, the possibility of entail lingered on until World War I. A new argument came to be used, however, in favour of unpartitioned inheritance of land in the 19th century. First in France and then in central Europe and Scandinavia, the argument was put forward that agricultural holdings were being reduced to less than the size necessary to provide a living for a family and that the old peasantry was thus in danger of being driven from the land. This led to the enactment of special laws on farm inheritance in sections of Austria, Germany, and Scandinavia. These laws, while providing unpartitioned inheritance in the case of intestacy, most often left unimpaired the power to provide for multiple succession by will and the power of alienation by sale. A more radical farm-inheritance law was enacted in Germany in 1933 by the National Socialists. It provided not only for undivided inheritance but also forbade partition by will and even the sale of the farm or its encumbrance by mortgage. The peasantry was to be secured as a social class living on the soil, removed from the vicissitudes and temptations of a market economy, although the law allowed the state to remove an unproductive holder. The law was repealed after World War II, but statutes attempting in milder ways to counteract the partitioning of farmsteads have been enacted. Even in France, the civil code was amended by a chain of laws beginning in 1922 so as to postpone, at least temporarily, the physical partition of a farmstead and certain other small holdings.

Intestate succession

Insofar as the course of succession is not determined by will, it is regulated by the laws of intestate succession. The legal systems of the world present a bewildering variety of intestacy laws, but they all have one feature in common: the intestate takers of the estate of a decedent are universally persons standing to him in a relation of kinship. Consequently, the composition of the group of successors in a society in which kinship is organized matrilineally is different from that of a society of patrilineal or, in modern society, bilineal kinship organization. Whether or not a surviving spouse belongs to the kin group of the decedent depends again on the way in which kinship is organized in the society in question. In modern laws, the surviving spouse is universally given some place in the table of successors, even though he or she may not be regarded as kin, or a relative, of the decedent.

Succession and kinship

HISTORICAL DEVELOPMENT

In preliterate society the order of succession seems to be basically determined by the kinship structure. But in both

archaic and developed societies the laws of intestacy have often been distorted by traditionalism, so that features once well adapted to the structure of the family were preserved into periods in which that structure had assumed new shape. The formalism that is characteristic of archaic legal systems (and often occurs in developed ones) tends to generalize rules that have originated in connection with special situations into applications beyond their initial scope. Intestacy laws have thus frequently looked obsolete, confused, or arbitrary. Even the Roman law of the Twelve Tables (*c.* 450 BC) seems not to have fully accorded with the social needs of its day.

Roman law. The basic unit of society in ancient Rome was the "house," the extended family ruled by its head, the paterfamilias, to whom his wife, his slaves, and possibly several generations of his descendants were subject and in whom title to all property was vested, so that a son or any other member of the house, even as an adult, did not own anything until he had been released from membership by emancipation. The paterfamilias was responsible for all liabilities incurred by any member. The Roman house of those early times resembled the system that prevailed in Japan until very recently. But whereas in Japan the leader of the house had just one successor, under the system of the Twelve Tables the Roman paterfamilias was succeeded by as many new ones as there were *sui heredes; i.e.,* persons who by the death of the chief were freed from his power and thus became persons *sui iuris.* If a house chief died without being survived by *sui heredes,* the law of the Twelve Tables provided that the estate (*familia*) could be acquired by the nearest agnatic relative; *i.e.,* the person related to the decedent by male descent who would be closest to him. If there was no such person, the estate could be had by the Gentiles, who seem to have been the clanlike group—composed of all descendants of a real or mythical ancestor—that apparently had ceased to play a significant role in Roman society even at the time of the Twelve Tables. This arrangement for succession seems to have been so unsatisfactory that it became customary—and even a moral, religious, and political duty—to eliminate its coming into play by the execution of a testament. The very name "intestate succession" (*successio ab intestato*) indicates that dying without having made a will constituted an exceptional situation.

As Rome grew into an empire, the system of the Twelve Tables became less and less satisfactory. The house of olden times receded in significance; relationship through females came to play as much a role in the consciousness of the people as that through males; and wives mostly ceased to be subject to the power of their husbands or their husbands' house chiefs. Adaptation of the law to the new structure of the family was made, first by the heads of the judicial system, the praetors, and then by imperial legislation. But the changes were unsystematic and half-hearted. In its final stage, the intestacy law became such a patchwork that in AD 543 and 548 the emperor Justinian found it necessary to make an entirely new beginning. By Novels (Novellae Constitutiones post Codicem, part of the Corpus Juris Civilis), a new order of intestacy was established. Relatives of a decedent were divided into four classes: (1) the descendants of the decedent; (2) the ascendants of the decedent, his brothers and sisters of the full blood, and the children of brothers and sisters of the full blood; (3) the decedent's brothers and sisters of the half blood and the children of such brothers and sisters; and (4) the other collaterals of the decedent related to him in the nearest grade of consanguinity. No person in a more remote class was to succeed as long as the decedent was survived by a member of a prior class. The surviving spouse stood outside the four classes of relatives. He or she was to succeed only if there was no relative at all. As long as any relative, no matter how remote, could be found, the family wealth was not to be diverted from the bloodline. But a widow's needs were ordinarily taken care of by the dowry, which, given to the husband, usually by her family, at the time of the marriage, was to be hers after the husband's death. For the exceptional case of a "poor widow"; *i.e.,* a widow without dowry, a share in the estate was provided. Distribution among members of the same class was not in all respects clearly regulated by Justinian's text, and so several points remained controversial.

The system of Justinian

Common law. Justinian's scheme was influential in the practice of the English ecclesiastical courts in their dealings with personal property. But in England the surviving spouse was treated much more generously. A surviving husband had no need to succeed to his wife's personal property upon her death. With the sole exception of choses in action not reduced to possession (*i.e.,* liabilities due to the wife not yet paid), he already owned all of her personal property by virtue of the marriage. But English custom gave the widow one-third of her predeceased husband's estate if he was survived by descendants and one-half if he was survived by other relatives. English ecclesiastical court practice also clarified some of the points that had been left open in Justinian's codification, abandoning the distinction between the siblings of the full blood and those of the half blood (although under the Canons of Descent to real property applied by the secular courts, the latter remained excluded). The English ecclesiastical practice was codified in the Statute of Distribution in 1670. This in turn became the model for state legislation in the United States, although the state laws show considerable variation in many respects.

English ecclesiastical court practice

Civil law. While the Anglo-American system has largely adopted the Roman order of succession, many of the civil-law countries appear to have followed the Germanic system of parentelic order.

Germanic system

The first parentela, or order, consists of the descendants of the decedent; the second, of his parents and their descendants collateral to the decedent; the third, of his grandparents and their descendants collateral to the decedent, etc. As long as there is any person standing in a nearer order, no person standing in a more remote one can succeed. In each parentela, persons of a lower grade exclude those of a higher grade. Variations exist in several respects. A person who stood in the first two orders but who predeceased the intestate is generally represented by his descendants; as to the more remote orders, the legal systems differ, as they do also with respect to the question of whether, insofar as representation takes place, the distribution is by roots or by heads. (When distribution is by roots—per stirpes—the estate is divided into as many parts as there are living and dead, but represented, persons standing in the same grade; when distribution is by heads—per capita—the estate is divided into as many parts as there are living persons entitled to sharing.) Considerable differences also exist in the treatment of the surviving spouse of the decedent.

INTESTACY IN PRESENT LAW

In recent times intestacy has been the subject of much legislation. Since the purpose of intestacy is to fill the void left where a decedent has not effectively disposed by will, lawmakers tend to create rules corresponding to those, which, in their opinion, a reasonable testator would have made. But they may also be inclined to lay down rules that they think testators ought to follow in the interests of social policy.

Modern tendencies. Among the most conspicuous trends of modern legislation is the vanishing concern about keeping property within the bloodline through which it came to the decedent. This traditional idea, which was particularly strong with respect to land, had in the field of intestacy resulted in the so-called rule of ancestral estate. In Anglo-American law the doctrine of ancestral estate was part of the Canons of Descent of real estate. It meant that if an intestate died without descendants, property that had come to him from ancestors should revert to the line whence it had come. In England the principle was abolished in 1925, and in the United States, too, it has practically disappeared. In civil-law countries, where it was once known as *jus recadentiae,* the principle has disappeared, except in the Spanish province of Aragon. But France has preserved the related ideas of the *fente* and the *droit de retour.* Under the former, the estate is divided equally between the paternal and the maternal lines (and under the *refente* between the various lines of grandparents). Under the *droit de retour,* assets that were received as a gift by an intestate who dies without descen-

Decline of the principle of ancestral estate

dants return to the donor. The once widespread idea that, among collaterals, relatives of the full blood occupy better positions than those of the half blood survives in France, in several Latin American countries, and, in several different forms, in some jurisdictions of the United States.

Changing attitudes about sex and marriage, but also about the equality and dignity of human beings, are widely finding expression in laws assimilating the legal situation of persons born out of wedlock to that of legitimate children. The equation of an illegitimate child with a foundling long ago disappeared from common-law systems, but it still persists in the laws of France and of countries following the French pattern, where a child born out of wedlock lacks legal relationship not only to his father but also to his mother until the relationship is formally recognized or established. Where parentage is established, a 1972 French law gives a child born out of wedlock nearly the same inheritance rights as a legitimate child, including the right to an indefeasible share of the parent's estate. An exception remains, however, for children conceived while their father or mother was married to a third person. In such cases, the child is not barred from inheritance altogether, but his or her rights are less than those of a surviving spouse or legitimate children.

In contrast, rights of inheritance between the illegitimate child and his mother have long been recognized in other jurisdictions, including those of the common law and of the civil law other than the French type. Impetus to the recognition of inheritance rights between an illegitimate child and his father was given by Norway's Castberg Law of 1915. Similar legislation now exists in the other Scandinavian countries, in the Federal Republic of Germany, in several countries of Latin America, in the countries of eastern Europe, in England under the Family Law Reform Act of 1969 and the Legitimacy Act of 1976, in Scotland under the Law Reform Act of 1968, and in several states of the United States. The statutes vary greatly with respect to such details as the mode in and the time at which paternity must be ascertained, the extent of the child's share in the estate, the question of the extent, if any, to which illegitimate relationship creates rights of intestate succession from or through an illegitimate, and the illegitimate's rights to intestate succession to kindred of his father or mother.

Rights of spouses. There is also a widespread trend toward improvement of the successoral position of the surviving spouse, often even favouring the spouse above the decedent's blood relatives. Benefits for a surviving spouse can, of course, be achieved by devices other than rights of inheritance. A method of great antiquity is the giving of a dowry, meant to sustain a woman after the death of her husband. In societies in which dowries are customary, the "poor widow" who lacks a dowry can then be helped by an exceptional right to a share in the estate, as was provided in Justinian's reform of the Roman law; this provision still exists in the state of Louisiana for the widow or the widower. A widow may be given a claim for support out of the estate, as in Austria (and in France between 1891 and 1925). Such support may even be provided for a wider circle of persons dependent upon the decedent, as in the family provision laws of England, New Zealand, and other Commonwealth countries.

Broadening of benefits for spouses

Benefits for a surviving spouse can also be provided through some system of community property, as found in numerous civil-law countries and in some states of the United States. The community fund may consist of the acquests made during marriage through the exertions of either spouse, or, additionally, of the movable assets owned by either spouse at the initiation of the marriage, or even of all property owned by the spouses. Upon one spouse's death, the fund is split into halves. One half constitutes the survivor's share in the community and thus belongs to him, whereas the estate of the predeceasing spouse consists of the other half of the community, along with such assets as the predeceasing spouse may have owned as his separate fund. The law may or may not then grant the surviving spouse an intestate share of the estate. Still another way of providing benefits for a surviving spouse is to give him or her a life estate in certain assets of the predeceasing spouse, as is done in the common-law institutions of dower and curtesy. French law, under certain circumstances, gives the surviving spouse a share in the estate or in a fraction of it.

Of great and increasing practical importance are the rights to pensions, social security benefits, and damage claims arising from the death of a married person, which are now universally available to a surviving spouse. Improvements in the right of a surviving spouse to share in the married couple's capital have been brought about in France, England, and numerous U.S. jurisdictions by giving him a preferred position in the scheme of intestate succession; or, as in Scandinavia, by giving the surviving spouse a share in the acquests made during marriage by the exertions of both spouses; or, as in the Federal Republic of Germany, by a combination of both devices. This revalorization of the surviving spouse is the consequence of several factors, including the weakening of family ties, the decreasing importance of inherited wealth, and the diminishing willingness of children and aged parents to share the same household. In addition, the feeling has grown that a wife who stays at home to run the household and bring up the children, instead of going out and earning a living of her own, enables her husband to act as the breadwinner and is therefore entitled to a share in his accumulated earnings.

Examples of existing laws. Intestacy laws vary widely in detail. The principal features of the intestacy rules of England, New York State, the American Uniform Probate Code, France, and the Russian Soviet Federated Socialist Republic are presented below.

England. The complex provisions of the Administration of Estates Act of 1925, as amended by the Intestates' Estates Act of 1952 and the family provision legislation (see above), are based on the following scheme:

Recent English legislation

1. The relatives of the decedent are grouped in seven classes. No member of a class takes in intestacy as long as any member of a preceding class has survived the decedent. The classes are: (a) descendants per stirpes; (b) parents; (c) brothers and sisters of the full blood, a deceased brother or sister being represented by his descendants per stirpes; (d) brothers and sisters of the half blood, such a deceased brother or sister being represented by his descendants per stirpes; (e) grandparents; (f) parents' brothers and sisters of the full blood; and (g) parents' brothers and sisters of the half blood.

2. A surviving spouse takes, if the decedent is survived by descendants: the "personal chattels," *i.e.,* the assets of the household, £40,000, and a life estate in one-half of the remaining part of the estate. If the decedent is not survived by descendants but is survived by parents, or brothers and sisters of the full blood, or descendants of such brothers and sisters, the surviving spouse takes the personal chattels, £85,000, and one-half of the remaining part of the estate. If the decedent is not survived by any of the above, the surviving spouse takes all.

If the intestate share of the surviving spouse, or of any person enumerated in the Inheritance Act of 1975 (see above), is insufficient to provide reasonable maintenance, the court may order that provision for such persons be made out of the estate.

New York State. Under the New York Estates, Powers and Trusts Law of 1966, as amended, relatives, grouped under the parentelic system, take by intestacy up to, but not beyond, the parentela of the grandparents. In the first and second parentelas, distribution is per stirpes; in the third it is per capita among persons standing in the same grade. If the decedent is survived by one child or the issue of one child, the surviving spouse takes $4,000 in money or intangible personal property and one-half of the residue; if more than one child or their issue survive, the spouse takes $4,000 in money or intangible personal property and one-third of the residue; if only parents or one parent survive, the spouse takes $25,000 and one-half of the residue; if there is no issue and no parent, the spouse takes all.

Recent U.S. approaches to intestacy

In addition, the surviving spouse or, if there is none, surviving minor children, are entitled to an allowance of $5,000 worth of household chattels, certain books of

the family library, $10,000 worth of farm equipment, and $1,000 in money or other personal property insofar as this amount is not needed to pay the reasonable family expenses.

Uniform Probate Code (U.S.). The latest state of U.S. thinking is expressed in the *Uniform Probate Code,* approved in 1969 and amended in 1975 and 1982 by the National Conference of Commissioners on Uniform State Laws and the American Bar Association. It has been adopted in several states and has significantly influenced law reform in others. Its provisions on intestacy are as follows:

> The intestate share of the surviving spouse is:
> (1) if there is no surviving issue or parent of the decedent, the entire intestate estate;
> (2) if there is no surviving issue but the decedent is survived by a parent or parents, the first [$50,000], plus one-half of the balance of the intestate estate;
> (3) if there are surviving issue all of whom are issue of the surviving spouse also, the first [$50,000], plus one-half of the balance of the intestate estate;
> (4) if there are surviving issue one or more of whom are not issue of the surviving spouse, one-half of the intestate estate.
>
> (For states with community-property laws, an alternative provision on the intestate rights of the surviving spouse exists.)
>
> The part of the intestate estate not passing to the surviving spouse . . . , or the entire intestate estate if there is no surviving spouse, passes as follows:
> (1) to the issue of the decedent; if they are all of the same degree of kinship to the decedent they take equally, but if of unequal degree, then those of more remote degree take by representation;
> (2) if there is no surviving issue, to his parent or parents equally;
> (3) if there is no surviving issue or parent, to the brothers and sisters and the issue of each deceased brother or sister by representation; if there is no surviving brother or sister, the issue of brothers and sisters take equally if they are all of the same degree of kinship to the decedent, but if of unequal degree then those of more remote degree take by representation;
> (4) If there is no surviving issue, parent or issue of a parent, but the decedent is survived by one or more grandparents or issue of grandparents, half of the estate passes to the paternal grandparents if both survive, or to the surviving paternal grandparent, or to the issue of the paternal grandparents if both are deceased, the issue taking equally if they are all of the same degree of kinship to the decedent, but if of unequal degree those of more remote degree take by representation; and the other half passes to the maternal relatives in the same manner; but if there be no surviving grandparent or issue of grandparent on either the paternal or the maternal side, the entire estate passes to the relatives on the other side in the same manner as the half.

Intestate succession in France

France. The French Civil Code was enacted in 1804, and its provisions of intestate succession have been changed many times. With respect to the surviving spouse, one must take into account the one-half share in the marital acquests that belongs to the surviving spouse unless some other arrangement was agreed upon at the time of the marriage.

The relatives are grouped in four classes, and no member of a more remote class succeeds as long as there is one of a prior class. The four classes are: (1) descendants per stirpes; (2) parents, brothers, sisters, and children of brothers and sisters; (3) ascendants other than parents; and (4) collaterals other than group 2, up to and including the 6th grade of consanguinity (*i.e.,* first cousins and grandchildren of great-grandparents).

The surviving spouse, in addition to his or her one-half share in the marital community fund, has the following rights in intestate succession: (1) a life interest in one-fourth of the estate if the decedent is survived by descendants; (2) a life interest in one-half the estate if the decedent is survived by ascendants in both the maternal and paternal lines or by certain collaterals (brothers, sisters, or their descendants); (3) one-half the estate outright if the decedent is survived by an ascendant in one line only; and (4) the entire estate if the decedent leaves no descendants, ascendants, or any of the above-mentioned collaterals.

The R.S.F.S.R. The civil code of the Russian Soviet Federated Socialist Republic (1964) provides the following order of intestate succession: (1) children, spouse, and parents of the decedent, in equal shares, a deceased child being represented by his child or children, and a deceased grandchild by his child or children; (2) brothers and sisters of the decedent and his paternal and maternal grandfathers and grandmothers, in equal shares.

In Soviet law

Intestate takers in the first group also include those persons who, unable to work, were dependent upon the decedent for not less than one year prior to his death.

Ordinary household furnishings and articles pass to those intestate takers who have lived with the decedent for not less than one year prior to his death, without regard to their class or statutory shares.

Wills

A will, or "testament," is the legal transaction by which an owner of property disposes of his assets for the event of his death. The terms are also applied to the written instrument in which the testator's dispositions are expressed. While in modern usage the terms will and testament are interchangeable, in traditional Anglo-American law "will" referred to the disposition of real property and "testament" to that of personal property.

FORMALITIES OF WILLS

A will must be declared in the form of an instrument in writing. A nuncupative (orally declared) will is exceptionally admitted in some jurisdictions in emergency situations, such as those of the soldier on active war duty, the sailor on board ship, or a person finding himself in immediate danger of death.

Kinds of wills

In their rules establishing the requirements for the execution of a regular testamentary instrument, the legal systems of the modern world usually follow one or more of three forms: (1) the witnessed will as developed in England, especially through the Statute of Frauds of 1677; (2) the unwitnessed holographic will as developed in French customary law; (3) the notarial will as developed in the late Roman Empire. Under the system of the witnessed will, which prevails throughout the United States and in all common-law parts of the British Commonwealth, the instrument, which may be typed or printed or written by anyone, must be subscribed by the testator, and his signature must be attested to by two (in some states, three) witnesses, who must also sign their names to the instrument. Under the system of the holographic will, which is available not only in most civil-law countries but also in numerous states of the South and West in the United States, the entire instrument, generally including the date and the indication of the place of execution, must be exclusively in the testator's own handwriting and must also be signed by him; witnesses are not required. The notarial will, which is also available in most civil-law countries, is executed so that the testator either dictates its provisions to the notary or hands him an instrument declaring that it contains his will. (In civil-law countries, a notary is not a layman but a respected member of the legal profession who is experienced in matters of drafting wills, estate planning, and conveyancing.)

The proper drafting of a will can be difficult. In the United States it is complicated not only by the diversity of the law from state to state but also by the fact that, unless different provisions have been expressly stated in the will, rules that are in many respects obsolete apply to such questions as how to apportion the burden of death taxes among the beneficiaries; in which order creditors ought to be paid; what assets are to be used for the payment of debts; which legacies are to be abated in case of insufficiency of the estate to pay them all in full; and what to do when a beneficiary has predeceased the testator. Unless the testator has given special powers to his executor, it may be necessary for the latter to observe cumbersome and expensive formalities in administration. In the United States it is therefore unwise to draft a will without expert legal advice; and it is advisable for a testator to have his will periodically checked by a lawyer in order to keep up not only with the changing circumstances of the testator's family circle and of his property but also to keep abreast

of frequent changes in the tax laws in order to avoid unnecessary taxes.

INVALID WILLS

Breaking and revoking wills

A testamentary disposition is not valid if at the time of its execution the testator was mentally incompetent or if he acted under "undue influence," *i.e.,* coercion, or under fraud. It is difficult, however, to break a will upon such grounds. The courts, especially those of Anglo-American systems, demand strict proof that the testator, when he made the provision, was mentally unable to know what he owned or who were his relatives or was unable to form a reasonable plan for the disposition of his property. The mere fact that the testator laboured under some insane delusion will not affect the validity of his will unless it is proved that this governed the disposition made by him. Coaxing and persuasion are generally not held to constitute undue influence in the absence of actual threats. A testator must not be pushed, but he may be led. Undue influence may be held to exist, however, where a testamentary disposition was brought about by a person upon whom the testator was dependent or whom he was likely to obey blindly.

The statutory formalities prescribed for the execution of a will must be observed meticulously. An unwitnessed holographic will may fail because the instrument contains a printed letterhead or some other words, figures, or signs in print, a rubber stamp, or another person's handwriting. A witnessed will may fail because a witness signed outside the testator's line of sight, or because the witnesses were not told that the instrument was the testator's will, or because a blank space was left between the end of the text and the signature of the testator. This strict compliance doctrine has come under increasing scholarly attack, and a few places now permit judges to uphold wills containing formal defects if the proponent of the will can show the defect in question was harmless to the purposes of the will.

The witnesses are supposed to be absolutely disinterested; *i.e.,* persons who derive no direct or indirect benefit from any of the provisions of the instrument. A witness may be held to be benefitted indirectly if his spouse is appointed in the will as executor and thus given the opportunity to earn the fees of that office. Ordinarily, attestation of a will by such a disqualified witness will not result in the invalidity of the entire instrument but only of the provision from which the witness would have benefitted.

A will is ambulatory; that is, it is of no effect until the testator's death, and it can be revoked or changed by him at any time. Revocation is effected either by the testator's physically destroying the instrument or by his executing a new testamentary instrument, the provisions of which are incompatible with those of the earlier one or in which it is simply declared that the will is revoked. In many states of the United States a will is also revoked automatically if the testator marries after its execution. Divorce often revokes any provision for the spouse in a preexisting will. In England a will is revoked by marriage unless it is stated to be made in contemplation of marriage. Attempts by contractual promise to limit one's freedom of changing or revoking one's will are without any effect in those legal systems that follow the pattern of the French civil code. But under the system of the German civil code, a disposition is irrevocable if it is expressed in a hereditary pact (*Erbvertrag*) made with a beneficiary or even with a third person. In Anglo-American law the will remains revocable even if the testator has promised that he will not revoke it; but if he does, his estate will be treated as if the testator had lived up to his promise. In practical effect, a testator may thus bind himself to make and not to revoke a will favouring a person who has promised to take care of him in old age. A husband and wife may promise each other that upon the death of one of them his property shall be enjoyed by the survivor and that upon the latter's death it will go to the children or to certain relatives or charities.

The machinery of transfer

Rules of intestate succession and of wills do not by themselves bring about the actual transfer of the decedent's assets to the new owners. A society with an economic system based on credit cannot operate on the once accepted principle that a person's debts die with him. Modern law must provide techniques for making sure that the debts left are properly paid. For the purpose of orderly transfer of assets and the proper payment of the debts of a decedent, two different techniques have been developed—one in the civil law and one in the common law.

TRANSFER IN CIVIL LAW

Development of the universal successor in civil law

The civil-law technique goes back to ancient Rome. When the head of a house, the paterfamilias, died, his position of headship devolved upon his heir or heirs. The heir, or heres, not only acquired all the ancestor's property but also his duties. The heres became liable for the debts, which meant that he had to dig into his own pocket if the assets of the estate did not suffice. This harsh rule was mitigated by the possibility given to the heir to abstain from, or to decline, the accession to the heirship. Then the option of accepting or declining devolved upon the person or persons next in line under the will or the rules of intestacy. If all declined, the succession ultimately came to the state, which was never liable beyond the value of the assets of the estate. Refusal to accept heirship to a father could appear as a violation of the duty of filial piety. Also, at the time the choice was to be made it might not always be apparent whether or not the estate was solvent. So another protective device was invented by Justinian: if, within a certain period of time, the heir fully and correctly inventoried the assets of the estate, his liability would be limited to the assets of the estate or to their value.

The Roman system is still basically that of the civil-law countries. There are, of course, many variations in detail, especially in the treatment of the situation of succession by a plurality of coheirs and in the treatment of the period of uncertainty as to who will ultimately accept the succession. There must be a person or a plurality of persons who, like the Roman heres, succeeds to the universality of the decedent's estate; *i.e.,* to the assets as well as to the debts. He or they, as the case may be, is or are determined by the decedent's testament, or by the law of intestacy, or by a combination of both. By his testament the decedent may charge the universal successor or successors with the duty to carry out legacies—*i.e.,* to hand over certain assets of the estate to third persons, or to pay to them certain amounts of money. Any person called to be universal successor is free to accept or to decline the position. If he chooses to accept, he may limit his liability to the assets of the estate either, as under the French system, by declaring his acceptance to be under the benefit of the inventory and by then making the inventory fully and correctly or, as under the German system, by handing over the estate to a judicially appointed administrator.

TRANSFER IN COMMON LAW

Development of probate in Anglo-American law

The Anglo-American system developed along quite different lines. Until the 19th century, liability for the debts of a decedent was limited to the assets of his personal estate. Real property was not liable unless it had been specifically mortgaged, in which case the mortgagee had his remedy of foreclosure. Thus, the title to the real property descended, like title to all assets under the civil law, directly to the heir, who acquired it immediately upon the death of the ancestor. But in order to guarantee the liability of the personal property for the debts of the decedent, as well as its proper distribution among the plurality of distributees, the ecclesiastical courts, which had the jurisdiction to deal with succession to personal property, worked out an original technique. Title was treated as passing from the decedent to the bishop or, later, to his substitute (surrogate) and ultimately to a middleman, on whom it was incumbent to pay the debts of the decedent and other claims that might exist against the estate and then to distribute the surplus remaining among the persons entitled thereto under the will or under the rules of intestacy. The provisions of a testament, however, would not be considered until the instrument had been admitted to probate, which means that it had been found to be properly executed and valid by the ecclesiastical court.

Probate. The modern laws of the Anglo-American countries have been developed upon this historical pattern with its peculiar features of probate and administration. In England the jurisdiction of the ecclesiastical courts was continuously narrowed by the royal courts. In the court reform of the 1870s, the new Probate, Divorce and Admiralty Division was established in the High Court. It took over from the ecclesiastical courts the narrow jurisdiction left to them, that of scrutinizing instruments purporting to be testaments; but simultaneously its jurisdiction was extended to wills; *i.e.,* instruments purporting to dispose of real property. Administrators are appointed by the Probate Division, but executors derive their powers directly from the will, so that they can act as soon as it is admitted to probate. If the personal representative wishes to obtain authoritative instructions on a problem occurring in the course of the administration, he can turn to the Chancery Division of the High Court. But as a general rule the personal representative is free to act on his own and under his own responsibility, much as does a civil-law heir.

In the U.S. branch of the common law a somewhat different machinery came into being. Ecclesiastical courts on the English pattern had not existed in the American colonies. The tasks of probate and of appointing administrators had thus to be performed by other agencies. In some places the job was for some time performed by the governor or some other officer of the executive. But the institution that was peculiarly developed in America was the Probate Court, which in most states is called by that name but in New York is known as Surrogate's Court and in Pennsylvania as Orphan's Court.

The tasks of a U.S. Probate Court are much more extensive than those of its English counterpart. United States Probate Courts usually are concerned with the administration of estates not only of decedents but also of minors and mentally incompetent persons. All of the persons entrusted with these tasks are treated as "officers of the court," who must all be appointed by the court, must give bond, and will be closely supervised by the court. In contrast to the practice in England, an American executor, or administrator, is not permitted to take any step without previous approval of the Probate Court, which normally requires that all interested parties be formally notified. Strictly speaking, no piece of furniture may be sold, no claim be settled, without the consent of the court.

If, as it is frequently done, the strict rules are not meticulously followed, troublesome complications may arise. In any case, time is needed and expenses have to be incurred. Devices have therefore been developed enabling the parties to dispense with administration or to avoid both probate and administration by resorting to transactions inter vivos that permit a person to give away his property while he is alive but under conditions allowing him to retain for himself not only the income and enjoyment during his lifetime but also the power of management, disposition, and revocation. Through such devices as revocable inter vivos trusts, joint tenancies, or "tentative trusts" of bank accounts (so-called Totten trusts), one can achieve the practical effects of a will without probate and without administration. One can also to some extent escape those safeguards that have been established for the protection of creditors and forced heirs and for the prevention of tax avoidance. While both the English and the civil-law systems maintain the principle that it is not possible to give and to keep, the ingenuity of U.S. lawyers and the indulgence of U.S. judges often have made it possible for an owner to "eat his cake and have it too." Reform legislation in a great many American states has made administration simpler and speedier, especially for small estates.

Transactions inter vivos

Administration. The Uniform Probate Code published by the National Conference of Commissioners on Uniform State Laws (1969) provides a choice among several systems of administration: simple and inexpensive ones for simple cases, and administration supervised by the court and containing elaborate safeguards for estates that are insolvent, under dispute, or that present other difficulties. The scheme of the new code thus assimilates the American law of winding up decedents' estates to those of England and of the civil-law countries, where the simple estate is treated as the normal, where no executor is needed unless he is expressly provided for in the testament, and where judicial administration is limited to cases of exceptional risk or complexity.

Under both the English and the American systems, the winding up of an estate follows essentially the same pattern. Anything looking like a will is filed with the Probate Court. The person named executor by the will, or some other interested person, petitions for the admission of the instrument to probate. The procedure varies in detail but, following English tradition, generally provides the possibility of simple, quick proceedings.

The executor

As executor a testator may choose any person or any corporation engaging in the business of administering trusts and estates. But the law seeks to have the position of administrator filled by that person likely to be the one primarily interested in the estate. The statutes thus contain lists of the persons who have a right to be appointed. The first place is usually given to the surviving spouse, followed by relatives in the order in which they are intestate takers. If there is no surviving spouse or relative, the right to administer devolves to the creditors of the decedent and, in the last line, to a public officer.

Under the laws that prevailed in the United States into the mid-20th century, real property and personal property were not handled in the same way. Unless a testator disposed otherwise, the payment of claims was from the personal estate. The real estate was not to be touched until the personal estate had been exhausted. Accordingly, title to only the personal estate passed to the executor or administrator. Title to the real estate descended directly from the decedent to the heir or heirs, or to the devisee or devisees named in the will. If the personal estate was insufficient to pay all the claims the executor or administrator would be authorized by the Probate Court to sell or mortgage as much of the real estate as was necessary. In many, although not all, U.S. jurisdictions this different treatment of real estate and personal estate has been abolished, making both liable for claims on an equal footing and equally subject to administration.

Under the English system an executor derives his position directly from the will; under the U.S. system an executor, like an administrator, must be appointed by the Probate Court, and no appointment will be made if the person is found to be disqualified or until he has taken the oath of office and posted bond.

An executor or administrator has to collect the assets of the estate, ascertain and pay the taxes and debts, and distribute the surplus to the legatees or intestate takers. In the civil-law system proper performance of these functions has its sanction in the personal liability of the heir for the debts of the decedent; if he takes steps to free himself of this liability, he has then to manage the estate like a common-law administrator or to hand it over to an administrator appointed by the court.

The office of executor or administrator may involve complex responsibilities. Unless one can be certain that there are no unknown debts, that the estate is solvent, that there will be no dispute about the distribution of the estate, that there are none but the simplest tax problems, and that there are no minors or mentally incompetent persons among the possible takers, one should not act as executor or administrator without the assistance of a lawyer.

BIBLIOGRAPHY

The historical background of modern laws of inheritance: MAX KASER, *Roman Private Law,* 3rd ed. (1980; originally published in German, 10th ed., 1977); W.W. BUCKLAND, *A Text-Book of Roman Law from Augustus to Justinian,* 3rd ed. rev. by PETER STEIN (1963, reprinted 1975); RUDOLF HÜBNER, *History of Germanic Private Law* (1918, reprinted 1968; originally published in German, 2nd ed., 1913); FREDERICK POLLOCK and FREDERICK W. MAITLAND, *History of English Law Before the Time of Edward I,* vol. 2, 2nd ed. (1959, reissued 1968); and SIR WILLIAM HOLDSWORTH, *History of English Law,* 7th rev. ed. by A.L. GOODHART and H.G. HANBURY, 17 vol. (1956–72).

Discussions and critiques of inheritance: JOHN STUART MILL, *Principles of Political Economy,* vol. 1 (1848), available also in many later editions; LUJO BRENTANO, *Alte und neue Feudalität: Gesammelte Aufsätze zur Erbrechtspolitik,* 2nd ed. (1924); JOSIAH WEDGWOOD, *The Economics of Inheritance* (1929,

reprinted 1971); SIR JOSIAH STAMP, *Some Economic Factors in Modern Life* (1929); FRIEDRICH ENGELS, *The Origin of the Family, Private Property and the State* (1942; originally published in German, 1884); and KARL RENNER, *The Institutions of Private Law and Their Social Functions* (1949, reissued 1976; originally published in German, 1929). MARVIN B. SUSSMAN, JUDITH N. CATES, and DAVID T. SMITH, *The Family and Inheritance* (1970), is a case study culminating in conclusions about testamentary patterns, practices, and effects of inheritance in American society.

Inheritance in socialist societies: J.N. HAZARD, *Communists and Their Law* (1969), and *Law and Social Change in the U.S.S.R.* (1953).

The current law of inheritance: Monographic treatises on contemporary inheritance laws are collected in MURAD FERID and KARL FIRSCHING (eds.), *Internationales Erbrecht* (1955–), a multivolume publication, loose-leaf for updates and additions, covering Argentina, Australia, Belgium, Bulgaria, China, Denmark, the Federal Republic of Germany, the German Democratic Republic, Finland, France, Greece, Great Britain, Ireland, Israel, Italy, Japan, Yugoslavia, Luxembourg, Mexico, New Zealand, The Netherlands, Austria, Poland, Sweden, Switzerland, Spain, South Africa, Czechoslovakia, Turkey, the Soviet Union, Hungary, the United States, and Vatican City. Concise digests of the succession laws of all jurisdictions of the United States, Canada, the United Kingdom, and a large number of other countries are published annually in *Martindale-Hubbell Law Directory*. For the United States, see WILLIAM HERBERT PAGE, *Page on the Law of Wills*, rev. ed. by WILLIAM J. BOWE and DOUGLAS H. PARKER, 8 vol. in 9 (1960–65), kept up-to-date by annual supplements; MAX RHEINSTEIN and MARY ANN GLENDON, *Law of Decedents' Estates* (1971), an introduction to the field. For England, see D.H. PARRY, *Parry and Clark on the Law of Succession*, 8th ed. by J.B. CLARK (1983); W.J. WILLIAMS, *Williams' Law Relating to Wills*, 5th ed. by C.H. SHERRIN, R.F.D. BARLOW, and R.A. WALLINGTON, 2 vol. (1980); and *Williams, Mortimer, and Sunnucks on Executors, Administrators, and Probate: Being the 16th Ed. of Williams on Executors and the 4th Ed. of Mortimer on Probate*, by J.H.G. SUNNUCKS, J.G. ROSS MARTYN, and K.M. GARNETT (1982). For Scotland, see DAVID M. WALKER, *Principles of Scottish Private Law*, 3rd ed., 4 vol. (1982–83).

(M.Rh./M.A.Gl.)

Insects

Insects constitute the class Insecta (or Hexapoda), the largest class of the largest animal phylum, Arthropoda. Like all arthropods, the insects have segmented bodies, jointed legs, and, when present, external skeletons (exoskeletons). Insects are distinguished from other arthropods by their body, which is divided into three major regions: (1) the head, which bears the mouthparts, eyes, and a pair of antennae; (2) the thorax, which usually has three pairs of legs (hence "Hexapoda," the name that was formerly given to this class) and usually one or two pairs of wings; and (3) the many-segmented abdomen, which contains the digestive, excretory, and reproductive organs.

In a popular sense, "insect" usually means the familiar pests or disease carriers, from bedbugs, houseflies, and clothes moths to Japanese beetles and aphids; the annoyers, such as mosquitoes, fleas, horseflies, and hornets; and the conspicuous butterflies and moths. Many insects, however, are beneficial from man's viewpoint; they pollinate plants, produce useful substances, control pest insects, act as scavengers, and serve as food for other animals (see below *Importance*). Furthermore, insects are valuable objects of study in elucidating many aspects of biology. Much of our knowledge of genetics has been gained from fruit fly experiments, and of population biology from flour beetle studies. Insects are often used in investigations of hormonal action, nerve and sense organ function, and many other physiological processes as well.

This article is divided into the following sections:

Insects: the class Insecta 585
- General features 585
  - Appearance and habits
  - Distribution and abundance
- Importance 586
  - Role in nature
  - Commercial significance
  - Agricultural significance
  - Medical significance
  - Control of insect damage
- Natural history 588
  - Life cycle
  - Sensory perception and reception
  - Behaviour
  - Ecology
- Form and function 591
  - External features
  - Internal features
- Evolution and paleontology 595
  - Origin of insects
  - Insect fossil record
  - Insect phylogeny
  - Evolution
- Classification 597
  - Distinguishing taxonomic features
  - Annotated classification
  - Critical appraisal

Major insect groups 598
- Apterygote (springtails, bristletails, and allies) 598
- Ephemeroptera (mayflies) 601
- Odonata (dragonflies) 603
- Orthopteran (cockroaches, mantids, stick and leaf insects, grasshoppers, crickets) 604
- Isoptera (termites) 612
- Phthiraptera (lice) 616
- Thysanoptera (thrips) 619
- Homoptera (cicadas, hoppers, whiteflies, aphids, scale insects) 621
- Heteroptera (true bugs) 628
- Neuropteran (lacewings, snakeflies, dobsonflies, alderflies) 635
- Trichoptera (caddisflies) 638
- Lepidoptera (butterflies, moths, skippers) 640
- Coleoptera (beetles, weevils) 651
- Hymenoptera (ants, bees, sawflies, wasps, and allies) 659
- Diptera (true flies) 667
- Siphonaptera (fleas) 673

Bibliography 675

INSECTS: THE CLASS INSECTA

GENERAL FEATURES

In numbers of species and individuals and in adaptability and wide distribution, insects are perhaps the most eminently successful of all animals. They dominate the present-day land fauna; almost 1,000,000 species have been described, representing about five-sixths of known animal life—the actual number of living species could range from 2,000,000 to 5,000,000, entomologists estimate. The orders that contain the greatest numbers of species are Coleoptera (beetles), Lepidoptera (butterflies and moths), Hymenoptera (ants, bees, wasps), and Diptera (true flies).

Appearance and habits. The majority of insects are

small, usually less than six millimetres long. The range in size, however, is wide. Some are almost microscopic, as certain of the feather-winged beetles and parasitic wasps; while some tropical forms attain considerable size, up to 16 centimetres in the hercules beetles, African goliath beetles, certain Australian stick insects, and some Asian and South American moths.

Diversity

In many species the difference in body structure between the sexes is pronounced, and knowledge of one sex may give few clues to the appearance of the other sex. In some, as the twisted-wing insects (Strepsiptera), the female is a mere inactive bag of eggs, and the winged male is one of the most active insects known. Modes of reproduction are diverse, and reproductive capacity is generally high. Some insects, as the mayflies, feed only in the immature or larval stage and go without food as adults. Among the social insects the queen ants and queen termites live for many years (15 to 50); some adult mayflies live less than two hours.

Some insects advertise their presence to the other sex by flashing lights, and many imitate other insects in colour and form and thus avoid or minimize attack by predators that feed by day and find their prey visually, as do birds, lizards, and other insects.

Behaviour is diverse, from the almost inert parasitic forms whose larvae lie in the nutrient blood streams of their hosts and feed by absorption, to dragonflies that pursue victims in the air, tiger beetles that outrun prey on land, and dytiscid beetles that outswim prey in water.

In some cases the adult insects make elaborate preparations for the young; in others the mother alone defends or feeds her young; and in still others the young are supported by complex insect societies, some of which (tropical termites and ants) may reach populations of millions of inhabitants.

Distribution and abundance. No scientist familiar with insects has attempted to estimate individual numbers beyond areas of a few acres or a few square miles in extent. Figures soon become so large as to be incomprehensible. The large populations and great variety of insects are related to their small size, high rates of reproduction, and abundance of suitable food supplies. Insects abound in the tropics, both in numbers of different kinds and in numbers of individuals.

If the insects (including the young and adults of all forms from microscopic young springtails [Collembola] to all large adults) are counted on a square yard of rich moist surface soil, 500 are found easily and 2,000 are not unusual in soil samples in the north temperate zone. This amounts to roughly 4,000,000 insects on one moist acre. In such an area only an occasional butterfly, bumblebee, or large beetle—supergiants among insects—probably would be noticed. Only a few thousand species, those that attack man's crops, herds, and products, and those that carry disease, interfere with human life seriously enough to require control measures. When they do appear as enemies of man, however, they can build up enormous populations with speed.

Adaptability

Insects are adapted to every land and freshwater habitat where food is available, from deserts to jungles, from glacial fields and cold mountain streams to stagnant, lowland ponds and hot springs. Many live in brackish water up to 1/10 the salinity of seawater, a few live in seawater, and some fly larvae can live in pools of crude petroleum, where they eat other insects that fall in.

IMPORTANCE

Role in nature. Insects play many important roles in the economy of nature. They aid bacteria, fungi, and other organisms in the decomposition of organic matter and in soil formation. The decay of carrion, for example, brought about mainly by bacteria, is accelerated by the maggots of flesh flies and blowflies. The activities of these larvae, which distribute and consume bacteria, are followed by those of moths and beetles, which break down hair and feathers. Insects and flowers have evolved together. Many plants depend on insects for pollination. Some insects are predators of others.

Commercial significance. Certain insects provide sources of commercially important products, for example, honey, silk, wax, dyes, pigments; therefore, insects can be of direct benefit to man. Because they feed on many types of organic matter, however, insects cause considerable agricultural damage. Insect pests devour crops of food or timber, either in the field or in storage, and convey infective micro-organisms to crops, farm animals, and human beings. The technology for combatting such pests of man constitutes the applied sciences of agricultural and forest entomology, stored product entomology, and medical and veterinary entomology.

Insects as a source of raw materials. Among primitive peoples who gathered their food (*e.g.*, Australian aborigines), insects were a significant food source. Grasshopper plagues, termite swarms, and large palm weevil grubs still serve as sources of protein in some countries. The dry scaly excreta of coccids (Homoptera) on tamarisk or larch trees is still the source of manna in the Sinai Desert. Coccids were once the source of the crimson dye kermes; and the cochineal, or carmine, from *Dactylopius* (Homoptera) on Mexican cacti, used for dying cloth by the Aztecs, is used also in lipsticks. Several insect waxes are used commercially, especially beeswax and lac wax. And the resinous product of the lac insect *Tachardia* (Homoptera), which is cultured for this purpose, is the source of commercial shellac.

Food, dyes, waxes, and silk

The most important domesticated insects are the silkworm (Lepidoptera) and the honeybee (Hymenoptera). Some coarse silks are derived from the cocoons of large wild species. Most commercial silks, however, come from the silkworm *Bombyx mori*. This insect is unknown in the wild state, and exists only in culture: it was domesticated in China thousands of years ago. Selective breeding, notably in China and Japan, has produced many specialized strains. The honeybee is a close relative of existing wild bees. In the Middle Ages honey was Europe's most important sweetener, and both beeswax and honey are still articles of commerce. The major importance of honeybees lies in their pollination of fruit trees and other crops.

Insect damage to commercial products. When insects that break down dead timber invade structural timbers in buildings, they become pests. This is true also among the latecomers to carcasses (*e.g.*, dermestid beetles and various tineid moths); they are capable of breaking down the keratin in hair and feathers. When these insects invade skins, furs, and wool garments or carpets, they cease to be useful members of the economy and become enemies of man.

In many hot, dry climates, as in North Africa or the plains of India, ripened grain in the fields is invaded by beetles. When the grain is harvested, these insects thrive in the grain stores. Such beetles (and certain moths with the same habits), carried throughout the world in commerce, have become universal pests of stored grain, dried fruit, tobacco, and other products. Quarantine and disinfestation methods are used to control importation of such insects from grain-exporting countries.

Agricultural significance. *Ecological factors.* Many insects are plant feeders; when the plants are of agricultural importance, man finds himself in competition with these insect pests. Populations of insects are limited by such factors as unfavourable weather, predators and parasites, and viral, bacterial, and fungal diseases; many other factors operate to make insect populations stable. Agricultural methods that encourage the planting of ever larger areas to single crops, however, have enhanced the rate of population growth of insects that attack those crops, thus increasing the probability of great infestations of certain insect pests. Many natural forests, which form similar giant monocultures, always seem to have been subject to periodic outbreaks of destructive insects.

Monocultures

In agricultural monocultures, an alien insect pest has often been introduced along with a crop, but without also bringing along its full range of enemies. This has occurred in the United States with the oyster scale (*Lecanium*) of apple, the cottony cushion scale (*Icerya*) of citrus, the European corn borer (*Pyrausta*), and many more. The Colorado potato beetle (*Leptinotarsa*), which caused appalling destruction a century ago, was a local insect, a native of semidesert country; that beetle, which fed on the

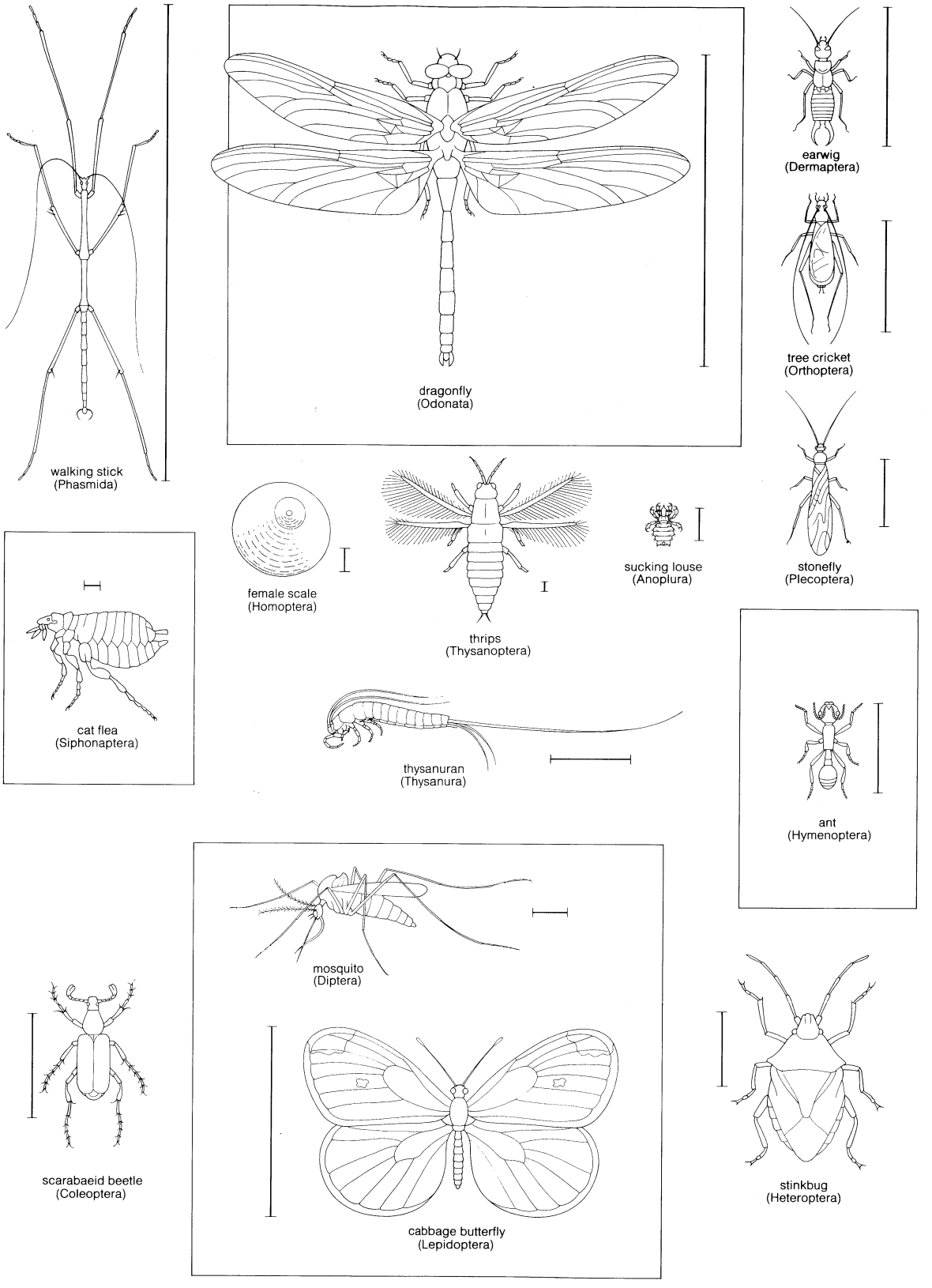


Figure 1: Diversity among adult insects. (Line scales indicate approximate size of each insect.)

buffalo burr plant, adapted itself to a newly introduced and abundant diet, potatoes, and thus escaped from all previous controlling factors. Similar situations often have been controlled by determining the major predators or parasites of an alien insect pest in its country of origin and introducing them as control agents. A classic example is the cottony cushion scale, which threatened the survival of the California citrus industry in 1886. The predatory ladybird beetle, or vedalia beetle (*Rodolia cardinalis*), was introduced from Australia, and within a year or two the scale insect had virtually disappeared. The success was repeated in every country where the scale insect became established without its predators. In eastern Canada in in the early 1940s the European spruce sawfly (*Gilpinia*), which

had caused immense damage, was completely controlled by the spontaneous appearance of a viral disease, perhaps unknowingly introduced from Europe. This event led to increased interest in plant diseases as potential means of pest control.

Damage to growing crops. Insects are responsible for two major kinds of damage to growing crops. First is direct injury done to the plant by the feeding insect, which eats leaves or burrows in stems, fruit, or roots. There are hundreds of pest species of this type, both in larvae and adults, among orthopterans, homopterans, heteropterans, coleopterans, lepidopterans, and dipterans. The second type is indirect damage, in which the insect itself does little or no harm but transmits bacterial or, more frequently, viral infection into a crop. Examples include the viral diseases of sugar beets and potatoes, carried from plant to plant by aphids.

Although most insects grow and multiply in the crop they damage, certain grasshoppers are well-known exceptions. They can exist in a relatively harmless solitary phase for a number of years after which their numbers may increase. They enter a gregarious phase, forming gigantic migratory swarms, which are transported by winds or flight for hundreds of miles. These swarms may completely destroy the crops in an invaded region. The desert locust (*Schistocerca gregaria*) and the migratory locust (*Locusta migratoria*) are examples.

Stings and bites

Medical significance. Insect damage to man and livestock also may be direct or indirect. Direct injury to man by insect stings and bites is of relatively minor importance, although swarms of biting flies and mosquitoes often make life almost intolerable, as do biting midges (sand flies) and salt-marsh mosquitoes. Persistent irritation by biting flies can cause deterioration in the health of cattle. Some blowflies, in addition to depositing their eggs in carcasses, also invade the tissue of living animals and man, a condition known as myiasis; the screwworm fly (*Cochliomyia*) of the southern U.S. and Central America is a specific example. In many parts of the world various blowflies infest the fleece and skin of living sheep; this infestation, called sheep-strike, causes severe economic damage.

Transmission of diseases

Most of the major fevers of man are produced by microorganisms that are conveyed by insects, which, therefore, are vectors of the fevers. Malaria is caused by the protozoan *Plasmodium,* which spends part of its developmental cycle in *Anopheles* mosquitoes. Epidemic relapsing fever, caused by spirochetes, is transmitted to man by the human louse *Pediculus.* Leishmaniasis, caused by the protozoan *Leishmania,* is carried by the sand fly *Phlebotomus.* Sleeping sickness in man and a group of cattle diseases that are widespread in Africa and known as nagana are caused by protozoan trypanosomes transmitted by the bites of tsetse flies (*Glossina*). Under nonsanitary conditions the common housefly *Musca* can play an incidental role in the spread of human intestinal infections (*e.g.,* typhoid, bacillary and amebic dysentery) by contamination of human food. The tularemia bacillus can be spread by deer fly bites, the bubonic plague bacillus by fleas, and the epidemic typhus rickettsia by the louse *Pediculus.* Various mosquitoes spread viral diseases (*e.g.,* equine encephalitis; dengue and yellow fever in man and other animals).

The relationships among the various organisms are complex. Malaria, for example, has a different epidemiology in almost every country in which it occurs, with different *Anopheles* species responsible for its spread. These same complexities affect the spread of sleeping sickness. The relationships between man and some diseases are indirect; plague, a disease of rodents transmitted by flea bites, is dangerous to man only when heavy mortality among domestic rats forces their infected fleas to attack man, thereby causing an outbreak of bubonic plague. Typhus, tularemia, and yellow fever also are maintained in animal reservoirs and spread occasionally to man.

Control of insect damage. The objective of the entomologist is to develop an applied ecology, that is, to introduce modifications into the environment in such ways that diseases will not be spread by insects, and crops will not be damaged. This objective has been achieved in some cases: in many cities flies no longer play a major role in spreading intestinal infections; improved land drainage and improved housing of man and animals has eliminated malaria in many parts of the world.

Massive outbreaks of the Colorado potato beetle in the 1860s led to the first large-scale use of insecticides in agriculture. These highly poisonous chemicals (*e.g.,* Paris green, lead arsenate, concentrated nicotine) were used in large quantities. The continued search for effective synthetic compounds led in the early 1940s to the production of DDT, a remarkable compound that is highly toxic to most insects, nontoxic to man in small quantities (although cumulative effects may be severe), and long lasting in effect. Widely used in agriculture for many years, DDT was not the perfect insecticide. It often killed parasites more effectively than the insects themselves, creating ecological imbalances that permitted new pests to develop large populations. Furthermore, resistant strains of pests appeared. Similar difficulties were encountered with many successors to DDT, such as Dieldrin and Endrin.

Biological control methods

Biological methods of control have become increasingly important as the use of undesirable insecticides decreases. Biological methods include introducing pest strains that carry lethal genes; or flooding an area with sterile males (as was successfully done for the control of the screwworm fly); or perhaps preparing a new kind of insecticide based on modifications of insects' growth hormones. The sugar industry in Hawaii and the California citrus industry rely on biological control methods. Although these methods are not consistently effective, they are considered to be less harmful to the environment than are some chemicals.

NATURAL HISTORY

Life cycle. *Egg.* Most insects begin their independent lives as fertilized eggs. The chorion, or eggshell, is commonly pierced by respiratory openings that lead to an air-filled meshwork inside the shell. For some insects (*e.g.,* cockroaches) a batch of eggs is cemented together to form an egg packet or ootheca. Insects may pass unfavourable seasons in the egg stage. Eggs of the lucerne flea *Sminthu-*

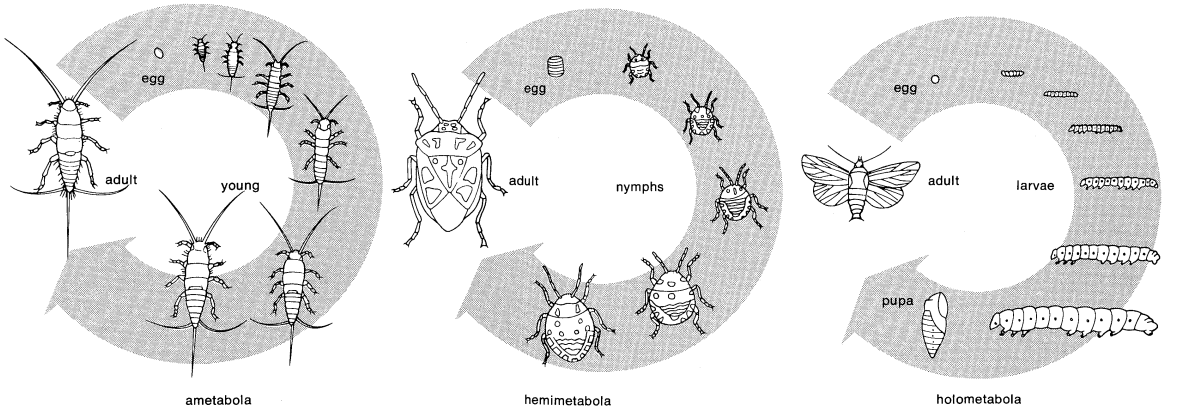


Figure 2: Types of insect development.

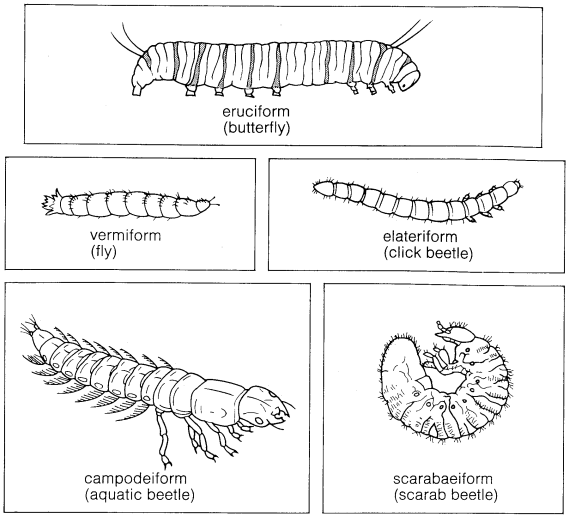


Figure 3: Insect larvae.

rus (Collembola) and of some grasshoppers (Orthoptera) pass summer droughts in a dry shrivelled state and resume development when moistened. Most eggs, however, retain their water although they may pass the winter in a state of arrested development, or diapause, usually at some early stage in embryonic development. Dried eggs of *Aedes* mosquitoes enter a state of dormancy after development is complete; they quickly hatch when placed in water.

The hatching of young larvae is achieved in several ways. Some, such as caterpillars, bite their way out of the egg. Many, such as the flea, have hatching spines with which they cut a slit in the shell; others force off a preformed egg cap. In order to exert this force, the young larva swallows air; after hatching, it continues to distend itself in this way until the cuticle hardens.

Growth and metamorphosis

Once formed, the insect cuticle cannot grow. Growth can occur only by a series of molts (ecdyses) during which new and larger cuticles form and old cuticles are shed. Molting makes possible large changes in body form.

Types of metamorphosis. In the most primitive wingless insects (apterygotes) such as the silverfish *Lepisma,* there is almost no change in form throughout growth to the adult. These are known as ametabolous insects. Among insects such as grasshoppers (Orthoptera), true bugs (Heteroptera), and homopterans (*e.g.,* aphids, scale insects), the general form is constant until the final molt, when the larva undergoes substantial changes in body form to become a winged adult with fully developed genitalia. These insects, termed hemimetabolous, are said to undergo incomplete metamorphosis. The higher orders of insects—*i.e.,* Lepidoptera (butterflies and moths), Coleoptera (beetles), Hymenoptera (ants, wasps, and bees), and Diptera (true flies)—are termed holometabolous because larvae are totally unlike adults. These larvae undergo a series of molts with little change in form before they enter into complete metamorphosis, which includes molting first into pupae and then into fully winged adults.

Types of larvae. Larvae, which vary considerably in shape, are classified in five forms: eruciform (caterpillar-like), scarabaeiform (grublike), campodeiform (elongated, flattened, and active), elateriform (wireworm-like), and vermiform (maggot-like). The three types of pupae are obtect, with appendages more or less glued to the body; exarate, with the appendages free and not glued to the body; and coarctate, essentially exarate but remaining covered by the cast skins (exuviae) of the next to the last larval instar (name given to the form of an insect between molts).

Molting hormone and juvenile hormone

Role of hormones. Both molting and metamorphosis are controlled by hormones. Molting is initiated by a hormone from neurosecretory cells in the brain. The hormone acts upon a prothoracic gland, an endocrine gland in the prothorax; this gland, in turn, secretes the molting hormone, a steroid known as ecdysone, which, by its action on the epidermis, stimulates growth and cuticle formation. Metamorphosis likewise is controlled by a hormone. Throughout the young larval stages a small gland behind the brain, called the corpus allatum, secretes the juvenile hormone (also known as neotenin). So long as this hormone is present in the blood the molting epidermal cells lay down a larval cuticle. In the last larval stage, juvenile hormone is no longer produced, and the insect undergoes metamorphosis into an adult. Among holometabolous insects the pupa develops in the presence of a very small amount of juvenile hormone.

Although a state of arrested development may occur during any stage, diapause occurs most commonly in pupae. In temperate latitudes many insects overwinter in the pupal stage (*e.g.,* cocoons). The immediate cause of diapause, failure to secrete the growth and molting hormones, usually is induced by a decrease in daylength as summer wanes.

In addition to the changes in form during development, many insects exhibit polymorphism as adults. For example, the worker and reproductive castes in ants and bees may be different; termites have a soldier caste as well as reproductives and persistent larvae; adult aphids (Homoptera) may be winged or wingless; and some butterflies show striking seasonal dimorphism. The general interpretation of all such differences is that, although the capacity to develop different forms is present in the genes of every member of a given species, particular lines of development are evoked by environmental stimuli. Hormones, including perhaps the juvenile hormone, may be agents for the control of such changes. For a discussion of the biochemistry of insect hormones, see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Hormones.*

Reproduction. The life of the adult insect is geared primarily to reproduction. Since reproduction is sexual in almost all insects, mating must be followed by impregnation of the female and fertilization of eggs. Usually the male seeks out the female. In butterflies in which vision is important, the colour of the female in flight can attract a male of the same species. In mayflies (Ephemeroptera) and certain midges (Diptera) the males dance in swarms to provide a visual attraction for females. In certain beetles (*e.g.,* fireflies and glowworms) parts of the fat body in the female have become modified to form a luminous organ that attracts the male. Male crickets and grasshoppers attract females by their chirping songs, and the male mosquito is lured by the sound emitted by the female in flight. The most important element in mating, however, is odour. Most female insects secrete odorous substances called pheromones that serve as specific attractants and excitants for males. The male likewise may produce scents that excite the female. Certain scales (androconia) on the wings of many male butterflies function in this way. Assembling scents, active in small quantities, are well known in female gypsy moths and silkworms as male attractants. The queen substance in the honeybee serves the same purpose.

Pheromones

Mating and egg production require appropriate temperatures and adequate nutrition. The need for protein is particularly important, and in insects such as Lepidoptera (butterflies and moths), which take only sugar and water in the adult stage, necessary protein is derived from larval reserves. Temperature and nutrition often influence hormone secretion. Juvenile hormone or hormones from the neurosecretory cells commonly are needed for egg production. In the absence of these hormones reproduction is arrested, and the insect enters a reproductive diapause. This phenomenon occurs in the potato beetle *Leptinotarsa* during the winter.

A few insects (*e.g.,* the stick insect *Carausius*) rarely produce males; the eggs develop without fertilization in a process known as parthenogenesis. During summer months in temperate latitudes, aphids occur only as parthenogenetic females in which embryos develop within the mother (viviparity). In certain gall midges (Diptera) oocytes start developing parthenogenetically in the ovaries of the larvae; the young larvae escape by destroying the body of their mother in a process called paedogenesis.

Sensory perception and reception. *Touch.* Insects have an elaborate system of sense organs. Tactile hairs, con-

centrated on the antennae, palps, legs, and tarsi, cover the entire body surface. The hairs serve to inform the insect about its surroundings and its body position (a phenomenon known as proprioception). For example, contact between the hairs on the feet and the ground inhibits movement and may lead to a state of sleep in some insects. Modified mechanical sense organs in the cuticle called campaniform organs detect bending strains in the integument. Such organs exist in the wings and enable the insect to control its movements. Campaniform organs, well developed in small clublike halteres (the modified hindwings of dipterans), serve as strain gauges and enable the fly to control its equilibrium.

Sound. Exceedingly sensitive organs called sensilla are concentrated in organs of hearing; *e.g.,* bushy antennae of the male mosquito or tympanal organs in the front legs of crickets or in abdominal pits of grasshoppers and many moths. In moths these sensitive organs can perceive the high-pitched sounds emitted by bats as they hunt by echolocation. Insects complement organs of sound reception with sound-producing organs, which usually are (as in crickets) wing membranes that vibrate in response to movement of a stiff rod across a row of stout teeth. Sometimes (as in cicadas) a tymbal membrane in the wall of the thorax is set in vibration by a rapidly contracting muscle attached to it.

Chemicals. Chemical perceptions by the thin-walled sensilla may be comparable to man's organs of taste or his sense of smell. Many insect chemoreceptors are specialized according to specific behaviour patterns. For example, although approximately equivalent to man in its perception of flower odours and sugar sweetness, the honeybee is exceedingly sensitive to the queen substance, which is scentless to man. And the male silkworm moth is excited by infinitesimal traces of the female sex pheromone, even in the presence of odours that are intensely strong to man.

Sight. Although the insect eye provides very poor form perception, insects by using a process of scanning (*i.e.,* moving the eye rapidly across a field of view) probably can form adequate visual impressions of their surroundings. Insects have good colour vision; colour perception commonly extends (as in ants and bees) into the ultraviolet, although it often fails to extend into the deep red. Many flowers have patterns of ultraviolet reflection invisible to the human eye, but visible to the insect eye.

Behaviour. *Instincts.* The insect orients itself by making orientation responses to the stimuli it receives. Formerly, insect behaviour was described as a series of forced movements in response to stimuli. That hypothesis has been supplanted by one that holds that the insect has a central nervous system with built-in patterns of behaviour or instincts that can be called forth by environmental stimuli; these instincts are modified by the insect's internal state, which has been affected by preceding stimuli. Searching for food or an egglaying site, catching prey, and mating are a few examples of complex behaviour. Experimental studies of details of behaviour have provided significant information about the properties of the sense organs. Patterns of behaviour range from comparatively simple reflex responses (*e.g.,* the avoidance of adverse stimuli, the grasping of a rough surface on contact with the claws) to the elaborate behavioural sequences involved in hunting, capturing, and eating prey. The highest developments of behaviour, found in social insects such as the ants, bees, and termites are based on the instinct principle.

Environmental stimuli

An interesting example of a behavioural pattern is that found in the leaf-cutter bee *Megachile.* The female bee first locates a site for her nest in rotten wood and shapes the nest into a long tunnel; then she seeks out preferred shrub leaves from which to build a cell and cuts first a disc for a cell cap, then a series of oval pieces for the walls. After preparing the nest, she stores a mixture of pollen and honey, lays an egg, and finally closes the cell with more cut leaves. The leaf-cutter bee repeats this sequence until the nest is filled. Each act can be performed only in this set sequence. The insect does not stop to repair any damage to the nest but proceeds undeterred to the next step in her behavioural pattern.

The honeybee society is more flexible than that of the leaf-cutter bee. Behavioural sequences of individuals are predictable, but the choice of acts or duties within the hive can be influenced by the needs of the colony. A capacity for learning does exist, and must exist, in any insect that has to find its nest; but learning capacity plays a relatively small part in the overall pattern of honeybee behaviour.

Insect societies. Both in complexity of behaviour and learning capacity, solitary bees and wasps are the equals of social wasps or honeybees. Social insects, however, have developed a division of labour in which the members must do the work required at the proper time. If the society is to succeed, its needs must be communicated to the individual, and the individual must act. These needs may be met by a temporary change in behaviour during which appropriate instinctive acts are performed or by changes in development that lead to the appearance of appropriate castes. Commonly, both behavioural and developmental changes are initiated by pheromones, which act as chemical messengers that convey information from one member of a colony to another.

The queen substance

Insect societies are gigantic families, the offspring of a single female. In the honeybee the single queen in the hive secretes the pheromone known as the queen substance (oxodecenoic acid); it is taken up by the workers and passed throughout the colony by food sharing. So long as the queen substance circulates, all members are informed that the queen is present. If the workers are deprived of queen substance, they proceed at once to build queen cells and feed the young larvae with a special salivary secretion known as royal jelly to produce more queens.

Pheromones liberated by termite soldiers or reproductive adults control the development of soldiers and reproductive forms. Alarm substances and other pheromones control much of the behaviour in ants. A remarkable form of communication is the dance language in the honeybee, in which the direction and approximate distance of a foraging site can be conveyed by one worker to another. For additional information on the social behaviour of bees, ants, and termites, see below in sections on the Hymenoptera and the Isoptera. See also BEHAVIOUR, ANIMAL.

Ecology. *Terrestrial insects.* Insects feed on every sort of organic matter, and their methods of feeding and digestion have become modified accordingly. The major climatic hazards faced by terrestrial insects are temperature extremes and desiccation. Different species function best at various optimal temperatures. If conditions are too hot, an insect seeks out a cool, moist, and shady spot. If exposed to the sun an insect positions itself so as to present the smallest amount of body surface to the heat. If conditions are too cool, insects remain in the sun to warm themselves. Many butterflies must spread their wings to collect heat before they can fly. A moth raises its temperature by vibrating its wings or "shivering" before taking flight. The heat generated in this way is conserved by hairs or scales that maintain an insulating layer of air around the body. The optimum muscle temperature for flight is from 38° to 40° C (100° to 104° F).

In extremely cold weather the danger for insects is freezing, and insects that survive winters in cold latitudes are called cold hardy. A few insects (*e.g.,* some caterpillars and aquatic midge larvae) tolerate ice formation in body fluids, although it is probable that the cell contents do not freeze. In most insects, however, cold hardiness means resistance to freezing. This resistance results partly from accumulation of large quantities of glycerol as an antifreeze and partly from physical changes in the blood that permit supercooling, without freezing, to temperatures far below the freezing point. Resistance to desiccation includes development of hard waterproofing waxes and exaggeration of water-conserving mechanisms.

Aquatic insects. Major adaptational changes—apart from remarkable modifications of the legs for swimming—concern respiration of aquatic insects. Some occur in insects that rise to the water surface to take atmospheric air into their tracheal systems. Mosquito larvae use only the last pair of abdominal spiracles, which open at the tip of a respiratory siphon. Water beetles (*e.g., Dytiscus*) have converted the space between the protective sheaths on the hindwings (elytra) and the abdomen into an air storage

Respiratory adaptations

chamber. Air-breathing insects can prolong the period of submergence by trapping air among their surface hairs. This air film acts as a physical gill and makes possible oxygen uptake from water.

Other adaptations to an aquatic environment have occurred in larvae that obtain all their oxygen from the water. In midge larvae, abundant tracheae (breathing tubes) supply the entire thin cuticle. Caddisfly larvae (Trichoptera) and mayfly larvae (Ephemeroptera) have tracheal gills. In large dragonfly larvae, the gills are inside the rectum, and the water is pumped in and out through the anus.

Protection from enemies. Insects may derive some protection from a horny or leathery integument; but they also have various chemical defenses. Some caterpillars carry among their body-surface hairs special irritating hairs, which break up into barbed fragments containing a poisonous substance that causes intense itching and serves as a protection against most birds.

Dermal glands of many insects discharge repellent or poisonous secretions over the cuticle; other insects are protected by poisons that are present continuously in the blood and tissues. Such poisons often are derived from the plants on which the insects feed. In many hymenopterans (ants, bees, wasps) accessory glands of the female, which usually pour out a secretion over the egg, have become modified to produce toxic proteins. These poisons, injected into the nervous system of the prey of solitary wasps, paralyze it; in this state the prey serves as food for the wasp larva. Similar stings are used by hymenopterans, including ants, wasps, and bees, for self-defense.

Concealment

Concealment is an important protective device for insects. Vast numbers hide beneath stones or the bark of trees. Others rely on protective coloration. Although insect colours depend partly on pigmentation in the outer body covering (cuticle), the most important pigments occur in epidermal cells below the cuticle. Butterfly and moth pigments are deposited inside flattened hairs, or scales, which cover the wings. Some of the most brilliant insect colours are not the result of pigmentation; they are physical interference colours produced by fine laminae in the surface of the scales. Protective coloration may take the form of camouflage (cryptic coloration) in which the insect is confused with its background. The coloration of many insects copies a specific background with extraordinary detail. Stick insects (*Carausius*) can accommodate their colour to that of a changing background by moving pigment granules in their epidermal cells. Some caterpillars have patterns that develop in response to a background; however, these are irreversible. Insects such as caterpillars, which rely on cryptic coloration, combine it with a rigid deathlike position.

Alternatively, insects that are well provided with chemical defenses generally show conspicuous warning, or aposematic, coloration. Experiments have proved that predators such as birds quickly learn to associate such coloration "labels" with nauseous or dangerous prey. Finally, insects without nauseous qualities may gain protection by mimicry, that is, by developing the conspicuous coloration found in distasteful species (see also COLORATION; MIMICRY).

Population regulation. The factors that limit the numbers of insect species are complex. Experimental studies of a population of grain beetles in a jar containing wheat show that the complexities increase if a second species is added. With insects in natural habitats, competing not only with members of their own species but with numerous other species as well, the obstacles to survival become increasingly great. Competition among species is reduced to some extent by adaptation of species to niches, or habitats, for which other insects do not compete.

Competition

Formerly, controversy arose over whether numbers were always density dependent (*i.e.,* limited by the density of the species itself) or whether catastrophic actions, notably the vagaries of weather, were often of prime importance. It has since become recognized that the ultimate factor in the control of numbers is competition within the species for food and other needs; but in many circumstances, before competition for food becomes significant, numbers are reduced by external factors. Competition within a species often is reduced by wholesale migration to new localities. Migration may occur by active flight, as in aphids and locusts, largely directed by the wind. Another important factor in the regulation of populations is balanced polymorphism of species, in which the prevalence of individuals with given characteristics changes according to the action of natural selection as the state of the environment changes. (See also BEHAVIOUR, ANIMAL; and BIOSPHERE.)

FORM AND FUNCTION

External features. *Cuticle.* The insect is covered by the cuticle, a layer of inert material laid down by a single sheet of epidermal cells. It consists mainly of chitin, a carbohydrate also known as polyacetylglucosamine, and sclerotin, a hard substance composed of protein tanned by quinones. The cuticle, which has a superficial layer of waterproofing wax to prevent loss of water by evaporation, also serves as the skeleton to which the muscles are attached. In insects (*e.g.,* caterpillars), in which the cuticle is soft and flexible, the skeleton is of the hydrostatic type; that is, body fluid pressure, maintained by muscle tension beneath the body wall, provides the firmness necessary for the function of muscles involved in movement. In insects with hard bodies, the cuticle is made up of hardened areas called sclerites; the flexible joints between sclerites provide a permanently rigid external skeleton. At the back of the head and in the thorax, hardened ingrowths of the cuticle, known as apodemes, furnish a kind of internal skeleton for muscular attachment.

Head. In present-day insects the primitive segments are grouped into three regions known as head, thorax, and abdomen. The first six segments have fused to form the head; the appendages of these segments have become modified into antennae that bear numerous sense organs and mouthparts that convey food to the mouth. Eyes also are prominent on the head. In most insects the mouthparts, adapted for chewing, consist of several parts; behind the upper lip or labrum is a pair of hard, toothed mandibles. These are followed by a pair of structures called first maxillae, each consisting of a bladelike lacinia, a hoodlike galea, and a segmented palp bearing sense organ. The paired second maxillae are partly fused in the midline to form the lower lip, or labium. Sometimes a median tonguelike structure, called the hypopharynx, arises from the floor of the mouth.

Mouthparts

Insect mouthparts have been modified strikingly and reflect particular methods of feeding. The dipterans (true

From H. Weber, *Grundriss der Insektenkunde* (1966); Gustav Fischer Verlag

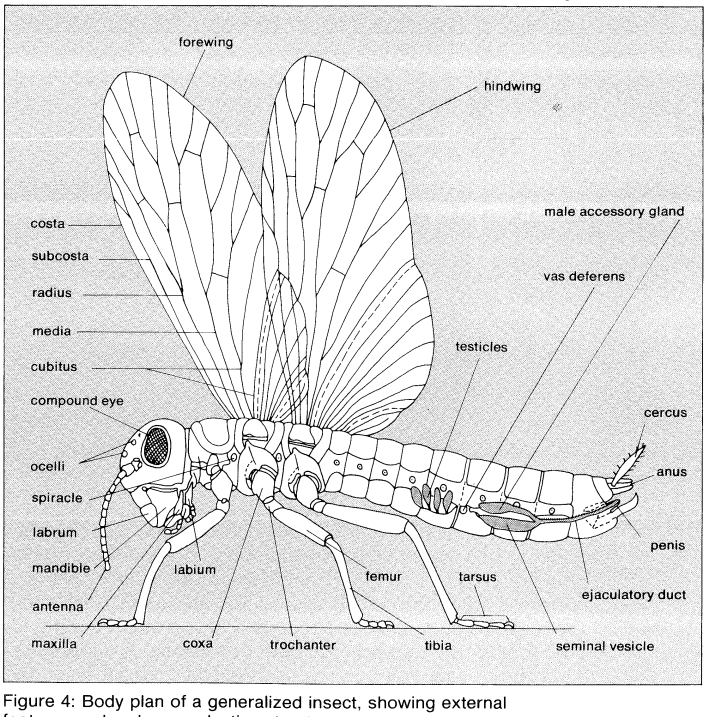


Figure 4: Body plan of a generalized insect, showing external features and male reproductive structures.

flies) provide instructive examples. In the primitive bloodsucking flies (*e.g.,* the horsefly *Tabanus*) the mandibles and maxillae form serrated blades that cut through the skin and blood vessels of the host animal. The epipharynx and hypopharynx are elongated and grooved so that, when apposed, they form a tube for sucking blood. The tonguelike labium is used for imbibing exposed fluids. Dipteran mouthparts have evolved in two directions. In the mosquitoes (Culicidae) the mandibles, maxillae, epipharynx, and hypopharynx have become exceedingly slender stylets that form a fine bundle and are used for piercing skin and entering blood vessels. The labium, elongated and deeply grooved, serves only as a sheath for the stylet bundle. In the housefly *Musca,* however, mandibles and maxillae have been lost; the tonguelike labium alone remains and serves for feeding on exposed surfaces. Certain flies related to *Musca* have reacquired a capacity to suck blood; however, since they have lost both mandibles and maxillae, a new bloodsucking mechanism has developed. Labial teeth have evolved for cutting through the skin, and the labium itself is plunged into the tissues. The stable fly *Stomoxys* has an arrangement of this kind. In the tsetse fly *Glossina,* the labium has become a fine, needlelike structure normally protected by a sheath formed from the palps of the lost maxillae.

Other mouthpart modifications of the mouthpart components provide the cutting and sucking mouthparts of fleas (Siphonaptera), plant-sucking insects (Homoptera), bloodsucking bugs (Heteroptera), honeybees (Hymenoptera), and nectar-feeding butterflies (Lepidoptera).

Thorax. The insect thorax consists of three segments (called the prothorax, mesothorax, and metathorax), which may be fused but are usually recognizable. Each segment has four groups of hard plates (sclerites); the groups are the notum (upper), the pleura (sides), and the sternum (underside). Thoracic sclerites are located on a given segment by using an appropriate prefix (pro-, meso-, meta-); for example, the notum (upper sclerite) of the prothorax is the pronotum.

Each segment bears a pair of legs, and, in the mature insect, the mesothorax and metathorax typically carry a pair of wings. Each leg always consists of five parts: a coxa articulated to the thorax, a small trochanter, a femur, a tibia, and a tarsus with one to five segments. The tarsal segments often carry claws with adhesive pads between them (arolia or pulvilli); these enable the insect to hold onto smooth surfaces. The legs may be modified for leaping, burrowing, grasping prey, or swimming in various ways.

From H. Weber, *Grundriss der Insektenkunde* (1966); Gustav Fischer Verlag

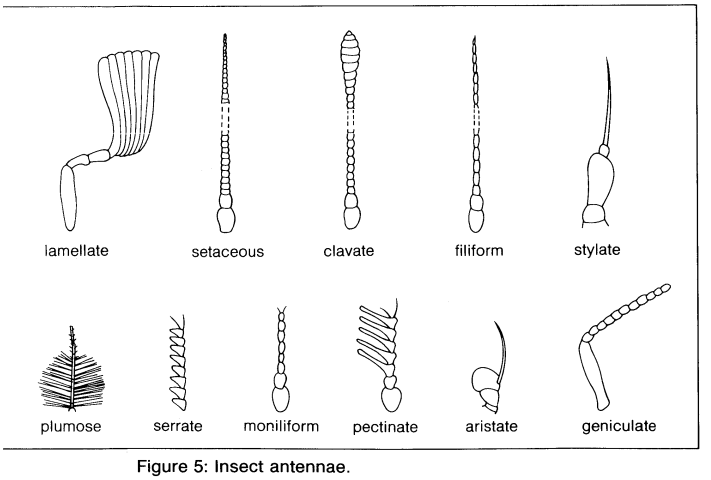


Figure 5: Insect antennae.

The wings at rest may be extended permanently on each side, as in some dragonflies (Odonata), or held erect above the body, as in mayflies (Ephemeroptera); in most insects, however, they are folded against the abdomen. The wing consists of cuticular sacs that bud out from the wall of the thorax; the sacs become flattened during development, and the two membranes, pressed together, are stiffened by thickenings of the cuticle that form cylindrical veins carrying tracheae, nerves, and circulating blood to all parts of the wing. Wings utilized for flight commonly are made of thin membranous cuticle. In some insects, notably beetles (Coleoptera), the wings of the middle segment of the thorax have become thick and horny and serve as protective sheaths (elytra) of the membranous hindwings.

Locomotion and flight

The locomotion of insects is effected by muscles acting on the external skeleton. In leaping insects (*e.g.,* grasshoppers, fleas) the force of muscle contraction is used to compress a pad of an elastic protein, resilin; when the catch mechanism is released, the stored energy in the protein molecule is used to project the insect into the air. Insect flight is achieved by flapping the wings; during these movements the wing blade, twisted as it passes from elevation to depression, produces the same effect as the rotating propeller of an aircraft. Muscles capable of changing this inclination control the direction of flight. The chief flight muscles control flight in one of two ways: in dragonflies, directly on a lever at the base of each wing; but, in most insects, indirectly by deforming the shape of the thorax. The longitudinal muscles of the thorax depress the wings that are articulated with it; the vertical muscles elevate them.

In butterflies, the number of wing beats per second may be as low as 8 to 12, while the rate in mosquitoes may exceed 600. These rates can exceed the frequency of contraction and relaxation of muscles responding to nerves because the muscles, after they have begun contracting and relaxing, respond to the alternating elastic tension in the thoracic wall, where the frequency is determined by the natural periodic oscillation of the thorax. The flight of insects, despite their small size, conforms to the aerodynamic laws that regulate the flight of aircraft.

Abdomen. The abdomen consists of a maximum of 11 segments, although this number commonly is reduced by fusion. Appendages are usually absent except in caterpillars, which use up to five pairs of abdominal prolegs in walking, and in adult insects where the appendages at the hind end have become transformed into external genitalia. In the male these genitalia are paired claspers used to hold the female; in the female, three pairs of valvulae are used to manipulate eggs during oviposition. In some insects, notably crickets and cockroaches, two feelers, or cerci, at the hind end of the abdomen bear sense organs.

Nutritive requirements

Internal features. *Digestive system.* The nutritive requirements of insects are much the same as those of mammals—water, inorganic ions, and essential amino acids (*i.e.,* those that cannot be synthesized by the animal). The requirements for preformed fat and carbohydrate vary with the species. Although vitamins of the B group are needed by insects, neither vitamins A nor D are required, and many insects can synthesize ascorbic acid (vitamin C). On the other hand, insects cannot synthesize adequate quantities of cholesterol; thus, in effect, cholesterol can be defined as a vitamin for insects.

Insects that feed solely on some restricted diet (*e.g.,* sterile blood, plant juices, refined flour) have special cells termed mycetocytes that harbour symbiotic microorganisms; these organisms, transmitted through the egg to the next generation, benefit their host by furnishing it with an internal source of vitamins and perhaps other essential nutrients. If the symbiotic micro-organisms are removed experimentally, an insect fails to grow if not provided with a diet rich in vitamins.

The digestive system consists of a foregut formed from the mouth region (stomodaeum), a hindgut formed similarly from the anal region (proctodaeum), and a midgut (mesenteron). The foregut and hindgut are lined by cuticle continuous with that on the body surface. The mouth is followed by the muscular pharynx, which functions in sucking and swallowing, and the esophagus, which may be enlarged to form a crop. The crop discharges into the midgut, sometimes, as in cockroaches, by way of a muscular gizzard or proventriculus. The termination of the midgut is marked by the attachment of the malpighian tubules, the chief organs of excretion. The hindgut commonly consists of a narrow ileum followed by a larger and often thick-walled rectum, which discharges at the anus.

Digestive enzymes, secreted not only by the salivary glands but also by the cells of the midgut and its divertic-

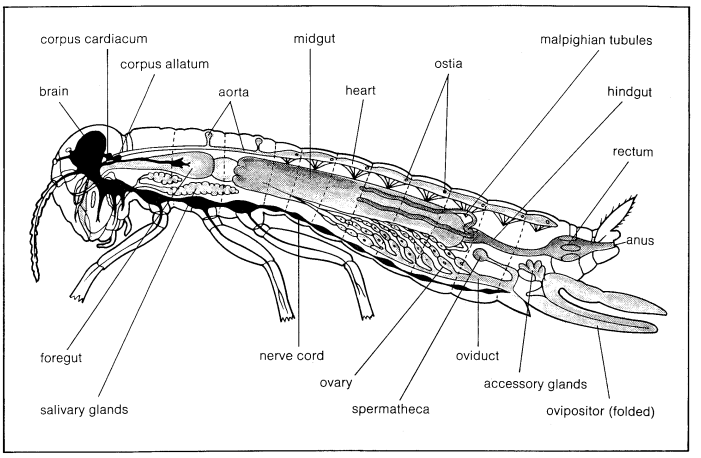


Figure 6: Internal features of a generalized female insect.
From H. Weber, *Grundriss der Insektenkunde* (1966); Gustav Fischer Verlag

ula, vary with the diet of the insect. The most important enzyme secreted by the salivary glands is amylase; the midgut secretes several enzymes including protease, lipase, amylase, and invertase. The products of digestion are absorbed chiefly in the midgut.

The hindgut receives food residues from the midgut as well as waste products from the malpighian tubules. The end products of nitrogen metabolism are uric acid, small amounts of amino acids, and urea; in aquatic insects, ammonium salts may be a major form for nitrogen excretion. In the rectum, the epithelial cells lining the gut wall often are enlarged, particularly in restricted areas where they form rectal glands. The epithelial cells of these glands are supplied richly with tracheae and function in the reabsorption of water and ions. The rectal contents of insects that inhabit dry environments commonly are reduced to dry fecal pellets prior to discharge. In many insects, particularly those which feed on relatively dry foods (*e.g.*, beetles infesting stored grain), the upper segments of the malpighian tubules are bound by a sheath to the rectal surface and form a cryptonephridial system that serves to increase the capacity of the rectum for reabsorbing water and salts. The products of digestion, discharged into the hemocoele, or general body cavity, are transported by the circulatory fluid, or hemolymph, to the organs.

Circulatory system. The circulatory system is an open one, with most of the body fluid, or hemolymph, occupying cavities of the body and its appendages. The one closed organ, called the dorsal vessel, extends from the hind end through the thorax to the head; it is a continuous tube with two regions, the heart or pumping organ, which is restricted to the abdomen, and the aorta, or conducting vessel, which extends forward through the thorax to the head. Hemolymph, pumped forward from the hind end and the sides of the body along the dorsal vessel, passes through a series of valved chambers, each containing a pair of lateral openings called ostia, to the aorta and is discharged in the front of the head. Accessory pumps carry the hemolymph through the wings and along the antennae and legs before it flows backward again to the abdomen.

From H. Weber, *Grundriss der Insectenkunde* (1966); Gustav Fischer Verlag

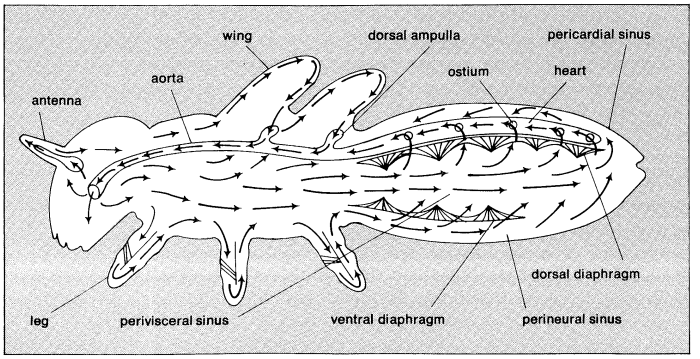


Figure 7: Circulatory system of a generalized insect.

The circulating hemolymph, or blood, is not important in respiration but functions in transporting nutrients to all parts of the body and metabolic waste products from the organs to the malpighian tubules for excretion. It contains free cells called hemocytes, most of which are phagocytes that help to protect the insect by devouring micro-organisms. An important tissue bathed by the hemolymph is the fat body, the main organ of intermediary metabolism. It serves for the storage of fat, glycogen, and protein, particularly during metamorphosis. These materials are set free as required by the tissues for energy production or for growth and reproduction.

Respiratory system. The respiratory system consists of air-filled tubes or tracheae, which open at the surface of the thorax and abdomen through paired spiracles. The muscular valves of the spiracles, closed most of the time, open only to allow the uptake of oxygen and the escape of carbon dioxide. The tracheal tubes are continuous with the cuticle of the body surface. The tracheae are stiffened by spiral thickenings or threadlike ridges called taenidia, which branch repeatedly, becoming reduced in cross section and ending in fine thin-walled tracheoles less than one micron in diameter. The tracheoles insinuate themselves between cells, sometimes appearing to penetrate into them, and push deeply into the plasma membrane. Tracheae

From G. Grandi, *Istituzioni di Entomologia Generale* (1966); Calderini

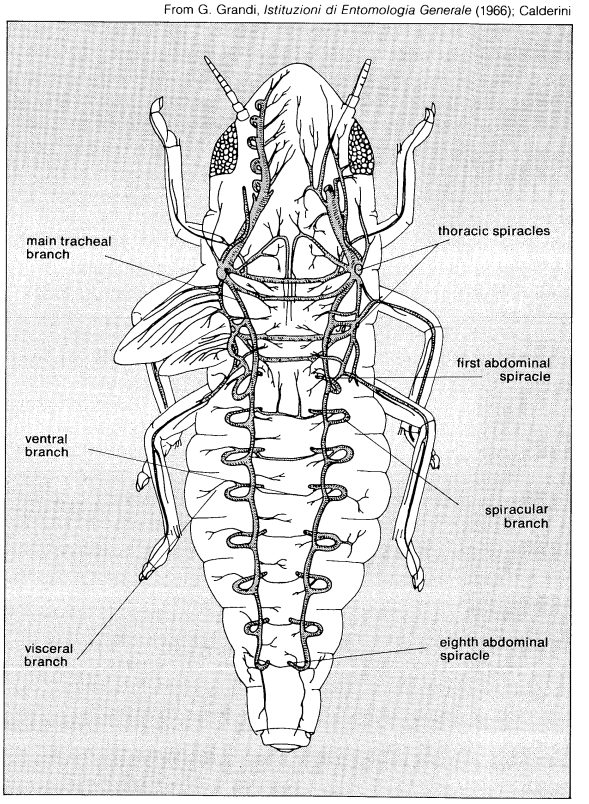


Figure 8: Respiratory system of a generalized insect.

Although movements of oxygen and carbon dioxide occur solely by gaseous diffusion in sedentary insects, the system is ventilated mechanically in active species. Pumping movements of the abdomen provide the force necessary to drive out streams of air at some spiracles and suck them in at others. The taenidia keep the tracheae distended, thus allowing free passage of air. In addition, the most active insects have large thin-walled dilatations of the tracheae called air sacs, which serve to increase the volume of air displaced during respiratory movements. Both lack of oxygen and accumulation of carbon dioxide provide stimuli to nerve centres that induce increased respiration during muscular activity.

Reproductive system. The reproductive system consists of the sex glands, or gonads (male testes and female ovaries), the ducts through which the sexual products are carried to the exterior, and the accessory glands. The two testes are made up of a variable number of follicles

in which the spermatocytes mature and form packets of elongated spermatozoa. Spermatozoa, liberated in bundles with heads held in a cap of gelatinous material, accumulate in the vesicula seminalis, a dilated section of the male sexual duct (vas deferens).

Each of the two ovaries consists of a number of ovarioles. The ovarioles converge upon the two oviducts, and the oviducts unite to form a common oviduct down which the ripe eggs are discharged. Each ovariole consists of a germarium and a series of ovarial follicles. The germarium is a mass of undifferentiated cells that form oocytes, nurse cells, and follicular cells. The nurse cells provide nourishment for the oocytes during the early stages of their growth; follicular cells, which invest the enlarging oocyte as a continuous epithelium, provide the materials for yolk formation and, in the final stages, lay down the eggshell or chorion. The ovarial follicles increase progressively in size as the oocytes grow to form ripe eggs.

During copulation, bundles of spermatozoa are sometimes introduced directly into the female vagina by means of the male copulatory organ, or aedeagus. Secretions from the accessory glands of the female activate the sperm, the sperm bundles disperse, and the free spermatozoa make their way up to the receptaculum seminis, or spermatheca, where they are stored, ready to fertilize the eggs. In most insects, the male accessory glands secrete materials that form a tough capsule, or spermatophore; spermatozoa are encased in this spermatophore, which is inserted into the entrance of the vagina. The spermatophore walls commonly contain a gelatinous substance that swells upon exposure to secretions of the female and forces out the spermatozoa. The vagina serves both for receiving sperm and for laying eggs.

Spermatheca

The terminal segments of the abdomen of females sometimes are modified to form an ovipositor used for depositing eggs. In butterflies and moths (Lepidoptera) a second copulatory canal independent of the vagina has been evolved, so that the sperm enter by one route, and the eggs are deposited by another.

The eggshell, or chorion, commonly provided with an air-filled meshwork, provides for respiration of the developing embryo. The chorion is also pierced by micropyles, fine canals that permit entry of one or more spermatozoa for fertilization. As the egg passes down the oviduct before egg laying, the micropyles come to lie opposite the duct of the spermatheca; at this stage fertilization occurs. Eggs must be waterproof to prevent desiccation; each egg has a layer of waterproofing wax, sometimes over the entire shell surface, more often lining the inside.

Nervous system. The central nervous system consists of a series of ganglia that supply nerves to successive segments of the body. The three main ganglia in the head (protocerebrum, deutocerebrum, and tritocerebrum) commonly are fused to form the brain, or supraesophageal ganglion. The rest of the ganglionic chain lies below the alimentary canal against the ventral body surface. The brain is joined by paired connectives to the subesophageal ganglion, which is linked in turn by paired connectives to the three thoracic and eight abdominal ganglia (numbered according to segment). In most insects the number of separate ganglia has been reduced by fusion. The last abdominal ganglion always serves several segments. In homopterans and heteropterans all the abdominal ganglia usually fuse with mesothoracic and metathoracic ganglia; and in the larvae of higher flies (Cyclorrhapha), the ganglia of the brain, thorax, and abdomen form one mass.

Each ganglion is made up of nerve-cell bodies that lie on the periphery and a mass of nerve fibres, the neuropile, that occupies the centre. There are two types of nerve cells, motor neurons and association neurons. Motor neurons have main processes, or axons, that extend from the ganglia to contractile muscles, and minor processes, or dendrites, that connect with the neuropile. Association neurons, usually smaller than motor neurons, are linked with other parts of the nervous system by way of the neuropile.

Cell bodies of the sense organs, called sensory neurons, lie at the periphery of the body just below the cuticle. Sensory neurons occur as single cells or small clusters of cells;

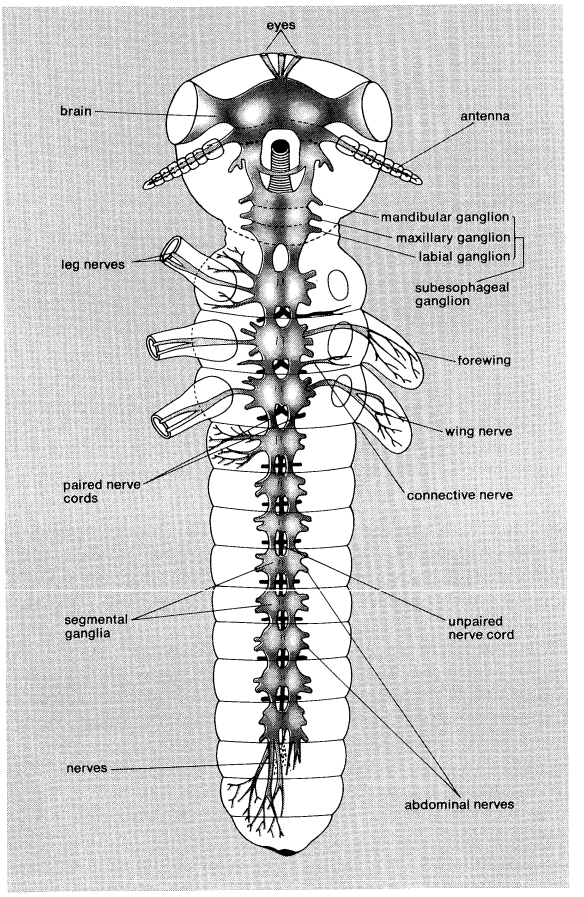


Figure 9: Nervous system of a generalized insect.
From G. Grandi, *Introduzione allo Studio della Entomologia* (1951); Edizioni Agricole

the distal process, or dendrite, of each cell extends to a cuticular sense organ (sensillum). The sensilla are usually small hairs modified for perception of specific stimuli (*e.g.*, touch, smell, taste, heat, cold); each sensillum consists of one sense cell and one nerve fibre. Although these small sense organs occur all over the body, they are particularly abundant in antennae, palps, and cerci. The sense cell of each sensillum gives off a proximal process, or sensory axon, which runs inward to the central nervous system, where it enters the neuropile and makes contact with the endings of association neurons. Bundles of both sensory axons and motor axons, which are enclosed in protective membranous sheaths, constitute the nerves.

Sensilla

Tactile hairs may be sensitive enough to perceive air vibrations and thus serve as organs for sound reception. Tympanal organs (eardrums) are present in certain butterflies and grasshoppers. Mechanical sensilla (chordotonal organs) below the surface of the cuticle serve for perception of internal strains and body movements.

Eyes. The eyes are of two kinds, simple eyes, or ocelli, and compound eyes. In the adults of higher insects both types are present. The visual sense cells are derived from the epidermis, as are those of other sense organs, and are connected to the optic ganglia (a part of the brain) by sensory axons. Each visual sense cell has a zone at its surface, which, on exposure to light, gives rise to chemical products that stimulate the sense cell, called the retinula cell, and initiate the nerve impulse in the sensory axon. The light-receptive zone, or rhabdom, of the retinula cell commonly has a rodlike form; because it lies perpendicular to the surface, light passes lengthwise along it. In the simple eyes (ocelli) a lens-shaped area of cuticle lies over the group of retinula cells that form the retina. Since the optical structure is primitive, the visual image received is crude; ocelli can perceive only light, darkness, and movement.

The compound eye, made up of a number of facets, resembles a honeycomb; each facet overlies a group of six or seven retinal cells that surround the rhabdom. Each

The compound eye

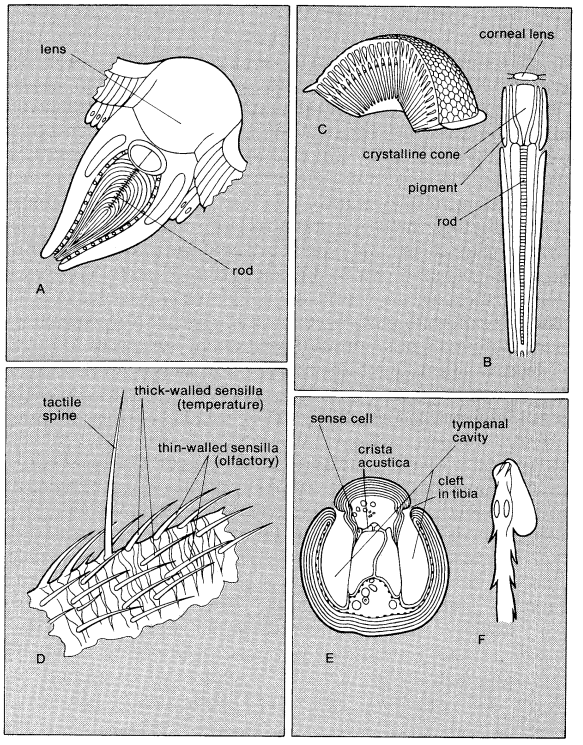


Figure 10: *Sensory receptors of insects.*
(A) Detailed structure of a single ocellus (simple eye); there is only one retinal rod below the lens. (B) Single ommatidium of a compound eye. (C) Compound eye showing ommatidia, each with its own lens. (D) Surface view of part of an antennal segment of *Rhodnius*. (E) Anterior tibia of *Decticus* at the level of tympanal organ. (F) Foreleg of *Tettigonia* showing slitlike openings of tympanal sacs.

From V.B. Wigglesworth, *The Life of Insects*, Copyright ©1964 by V.B. Wigglesworth; World Publishing Co.

of the retinal units below a single facet is termed an ommatidium. The number of facets varies. For example, there are only a few dozen facets in the eye of the primitive apterygote *Collembola,* while the eye of the housefly *Musca* has some 4,000, and the highly developed eye of the dragonfly may contain up to 28,000.

During light reception, rays from a small area of the field of view fall on a single facet and are concentrated upon the rhabdom of the retinula cells below. Since each point of light differs in brightness, all the ommatidia that form the retina receive a crude mosaic of the field of view. Unlike the image in a camera or in human eyes, the mosaic image in the compound eye is not inverted but erect. The fineness of the mosaic and, therefore, the degree of resolution improves with increasing numbers of facets. It is estimated that the eye of the honeybee has visual acuity equal to 1 percent that in man.

Each ommatidium commonly is shielded by a curtain of pigmented cells that prevent the spread of light to neighbouring ommatidia. This is termed an apposition eye. In the eyes of insects that fly at night or in twilight, however, the pigment can be withdrawn so that light received from neighbouring facets overlaps to some extent. This is termed a superposition eye. The image formed is brighter but not as sharp as that formed by the apposition eye. In addition to perceiving brightness, the eyes of insects can perceive colour as well as some other properties of light.

EVOLUTION AND PALEONTOLOGY

Origin of insects. The most primitive insects known are found as fossils in rocks of the Middle Devonian Period and lived about 350,000,000 years ago. The bodies of those insects were divided then, as now, into a head bearing one pair of antennae, a thorax with three pairs of legs, and a segmented abdomen. Those insects originated with the terrestrial branch of the phylum Arthropoda. The Arthropoda, whose origin is thus far unknown, probably arose in Precambrian times, perhaps as much as 1,000,000,000 years ago. Some arthropods colonized the open sea and have become the present-day class Crustacea (crabs, shrimps) and the now-extinct Trilobita. Other arthropods colonized the land. This terrestrial line persists chiefly as the classes Onychophora, Arachnida (spiders, scorpions, ticks), the myriapods (consisting of Diplopoda [millipedes], Pauropoda, Symphyla, and Chilopoda, or centipedes), and finally the class Insecta.

Apterygotes

The most primitive insects today are found among the wingless (apterous) hexapods; sometimes known collectively as apterygotes, they include proturans, thysanurans, diplurans, and collembolans. It is agreed generally that insects are related most closely to the myriapod group, among which the Symphyla exhibit most of the essential features required for the ancestral insect form (*i.e.,* a Y-shaped epicranial suture, two pairs of maxillae, a single pair of antennae, styli and sacs on the abdominal segments, cerci, and malpighian tubules). There is, therefore, general agreement that the insects probably arose from an early symphylan-like form.

Insect fossil record. The insect fossil record has many gaps. Among the primitive apterygotes, only the collembolans (springtails) have been found as fossils in the Devonian Period. Ten insect orders are known as fossils, mostly of Late Carboniferous and Permian times. No fossils have yet been found from the Late Devonian or Early Carboniferous periods, when the key characters of present-day insects are believed to have evolved; thus, early evolution must be inferred from the morphology of extant insects.

It has become evident that insect evolution, like that of other animals, was far more active at some periods than at others. There have been geological epochs of "explosive" evolution during which many new forms have appeared. Those epochs may have followed some modification or innovation in body function, or new developments favoured by climatic changes or evolutionary advances of other animals and plants. During those periods of evolutionary change, new methods of feeding and living led to diversity of insect mouthparts and limbs, the origin of metamorphosis, and other changes.

Insect phylogeny. Figure 11 is a simplified family tree of the presumed evolutionary history of winged insects (Pterygota) throughout the geological periods from the Devonian to the Recent. The apterygotes, which are regarded as survivors of primitive insect stock, are omitted from the family tree. Dark lines indicate the periods during which the various orders have been found as fossils. Some lines stop at the names of orders now extinct and known only as fossils. Light lines indicate the hypothetical origin of various orders. Many insect types, traces of which have not yet been discovered, must have been produced during the explosive periods of evolution in Carboniferous and Permian times.

The primitive wingless insects (Figure 11) gave rise to a paleopterous stock. Descendants of this stock included ancient fossil types that flourished in Permian times, such as the giant dragonflies or Protodonata (some of which had a wing span of more than half a metre) and dragonflies and damselflies (Odonata) and mayflies (Ephemeroptera), both of which have persisted with little change to the present. The primitive insect stock also gave rise to a neopterous stock, believed to include the progenitors of the remaining insect orders. The Orthoptera (grasshoppers) and the Plecoptera (stoneflies) have been found as fossils even in late Carboniferous times. The Isoptera (termites), Embioptera (webspinners), and Dermaptera (earwigs), though doubtless of ancient origin, have not been found yet as fossils dated earlier than the Mesozoic Era.

The evolutionary radiation (Figure 11), believed to have given rise to the orders listed above in the Middle Carboniferous Period, is thought to have produced also a paraneopterous stock, which formed the base for a new evolutionary radiation during the Permian Period. Present-day derivatives of this stock evolved into the Psocoptera (psocids), Mallophaga (chewing lice), Anoplura or Siphunculata (sucking lice), Thysanoptera (thrips), Heteroptera (true bugs), and Homoptera (*e.g.,* aphids).

Several phylogenetic lines (Figure 11) are exopterygote (*i.e.,* insects with simple metamorphosis) some of which, such as Mallophaga and Anoplura, are secondarily wing-

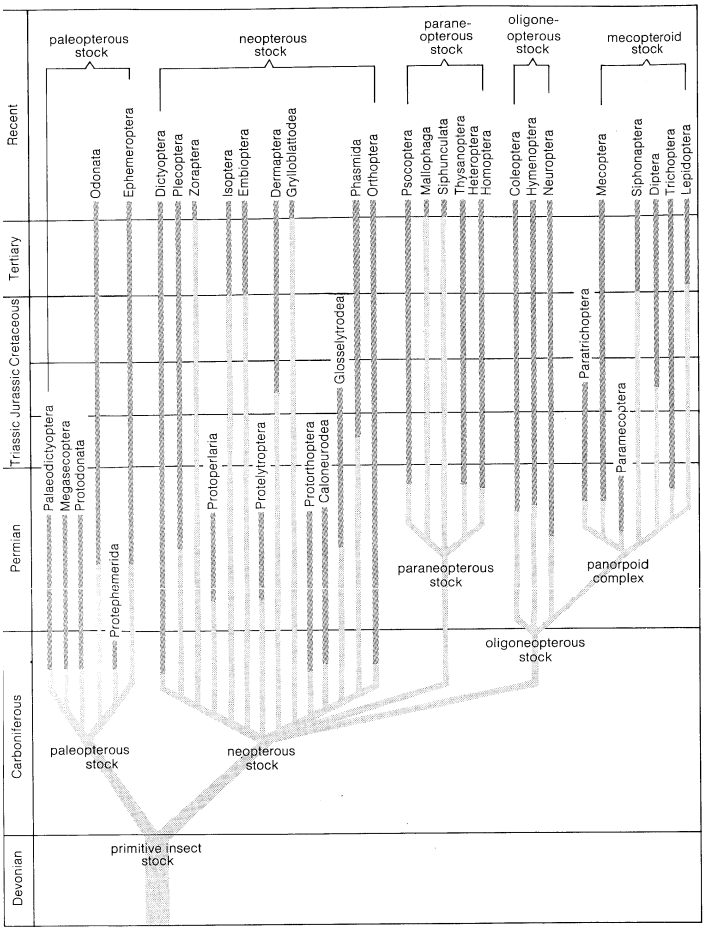


Figure 11: Phylogenetic relationships of insects.
From J. Smart, "Explosive Evolution and the Phylogeny of Insects," *Proceedings of the Linnean Society of London* (1963)

less. The remaining orders are endopterygote (insects with complete metamorphosis). They are shown in Figure 11 as derivatives of an oligoneopterous stock, which gave rise to Neuroptera (lacewings), Hymenoptera (ants, wasps, and bees), and Coleoptera (beetles) in the Early Permian Period; the early ancestry of these orders is obscure, however, and the earliest fossils closely resemble present-day forms. One line from the evolutionary radiation (Figure 11) at the beginning of the Permian gave rise to a mecopteroid stock, and there is good evidence that a sub-radiation of these mecopteroid orders (sometimes called the panorpoid complex) provided the origin for the present Mecoptera (scorpionflies), Diptera (true flies), Siphonaptera (fleas), Trichoptera (caddisflies), and Lepidoptera (butterflies and moths).

Evolution. *Wings and flight.* Insect wings develop as paired outgrowths from the thorax, stiffened by ribs, or veins, in which run tracheae. These tracheae follow a consistent pattern throughout the Pterygota, and their specific modifications (known as venation) are important in classification and in estimations of the degree of relationship between groups. The basic consistency of venation suggests that wings have been evolved only once among the insects, that is, all the Pterygota (as shown in Figure 11) arose from a single stem. By the time (toward the end of the Carboniferous) fossil insects are found, wings are developed fully. In the Paleoptera (Figure 11) the wings are held aloft above the back, as in mayflies, or held extended permanently on each side of the body, as in dragonflies. Throughout the Neoptera there is a wing-flexing mechanism (secondarily lost in butterflies) that enables the wings to be folded back to rest on the surface of the abdomen.

First appearance of winged insects

Winged insects must have made their appearance very early in the Carboniferous, more than 300,000,000 years ago; but there is no fossil evidence to show the way they evolved. One hypothesis is that wings arose as fixed planes extending sideways from the thorax and that these planes were used, perhaps in some large leaping insect, for gliding. Later muscles developed, first to control inclination and then to move the wings in flapping flight. Another hypothesis is that wings may have originated from large thoracic tracheal gills, similar to the movable tracheal gills along the abdomen of some mayfly larvae. Such outgrowths could have been useful to insects exposed by the drying up of a temporary aquatic habitat and might have carried them in rain-bearing winds to a new watery home. It is likely that the most primitive symphylan-like insects were terrestrial; throughout insect evolution, however, independent adaptations to aquatic habitats have occurred. Usually the pattern is one in which the adults leave the water and disperse. Many pterygote insects have become secondarily wingless, sometimes as single species or groups of species within large orders, sometimes as entire orders (the parasitic lice, Mallophaga and Anoplura, and the fleas, Siphonaptera).

Metamorphosis. It generally is agreed that insect metamorphosis evolved as adult insects gradually adopted different modes of life from those of larvae. The characters of larva and adult became genetically independent; in response to natural selection, therefore, each was able to evolve independently of the other. Mouthparts, limbs, and other morphological features were modified in different directions and in higher groups. Where these differences were extreme, an intermediate pupal stage evolved to bridge the morphological gap between larva and adult. It seems quite probable that the development of metamorphosis occurred more than once during the evolution of insects.

Feeding methods. Insects did not evolve in a constant environment. Throughout geological time there were prodigious changes in climate; in addition, evolution was continuous among all other animals and plants. Geologically the selection pressures among insects were changing continuously. At the end of the Mesozoic Era the first flowering plants appeared. Insect evolution has paralleled that of the flowering plants; they have evolved together. As Lepidoptera (butterflies and moths), Hymenoptera (ants, bees, and wasps), Diptera (true flies), and Coleoptera (beetles) began to feed upon flowers, nectar, or pollen, flowering plants came to rely more and more upon insects—rather than upon the wind—for transferring their pollen. Flowers evolved nectaries, scents, and conspicuous colours as attractants for those insects that could effect cross-pollination. Insects likewise evolved appropriate mouthpart modifications for extracting nectar from flowers.

Parallel evolution with flowering plants

During the Mesozoic warm-blooded animals (mammals and birds) first appeared; by the dawn of the Tertiary Period, they had become predominant among the earth's large animals. The warm fermenting excrement and the decaying dead bodies of mammals furnished excellent nutrient media for many insect larvae, notably among the Diptera and Coleoptera. The adults in both groups found their nourishment in flowers. Some heteropterans (true bugs) and dipterans pierce the skin of birds and mammals and feed on their blood. The Anoplura (sucking lice) and the Siphonaptera (fleas) have become so specialized for this type of parasitic existence that their relationships to other insects are not yet known with certainty.

Continuing evolution. Evolution is occurring among present-day insects. They exhibit a balanced genetic polymorphism; in other words, in response to small environmental changes, one genetic form, more successful than another, will become more plentiful. Sometimes there is no visible difference between these forms, the advantage presumably lying in some physiological change. It is advantageous for a species to have a gene pool from which favourable characters can be selected so that the species can respond to environmental changes. Changes within a species may occur progressively over a large geographical area. Such a progressive genetic change is termed a cline; in some cases insects at the extremes of the cline are so unlike that they are taken as separate species and may be infertile when crossed.

One well-known example of evolution in action among insects is industrial melanism (accumulation of the black pigment melanin); many butterflies inhabiting industrial

areas have become almost black during the past century; black forms are more tolerant of pollution and less conspicuous to predators. Another example of this cline type of evolution is the development of insect strains resistant to an insecticide that has been applied heavily in an area for several years. In many parts of the world houseflies have become highly resistant to DDT.

CLASSIFICATION

Distinguishing taxonomic features. The class Insecta is divided into orders on the basis of the structure of the head, including eyes, mouthparts, and antennae; the thorax, including legs and wings; and the abdomen, including segmentation, spiracles, and appendages (cerci, styli, furcula). The genitalia and their accessory structures, usually located in the ninth abdominal segment, are important in classification. Other taxonomic criteria include bristles (their form and arrangement are known as chaetotaxy), sensory receptors (spines, hairs, sensilla, tympanal organs), pattern of wing venation, and position of mouthparts. In addition, type of metamorphosis and form of larva and pupa are used to distinguish insects.

Annotated classification.

CLASS INSECTA

Body divided into head, thorax, and abdomen; head with 1 pair of antennae, mouthparts consisting of a pair of mandibles and 2 pairs of maxillae, the 2nd pair fused medially; thorax with 3 pairs of legs and usually 1 or 2 pairs of wings; segmented abdomen lacks walking appendages; genital opening near anus; metamorphosis usually occurs; about 1,000,000 species named; worldwide distribution.

Subclass Apterygota

Primitively wingless; metamorphosis slight or absent (ametabola); adult with 1 or more pairs of pregenital appendages; adult mandibles articulate with head capsule at a single point.

Order Protura. Minute in size; mouthparts entognathous (withdrawn in head capsule) and piercing; no antennae or compound eyes; abdomen with 11 segments, well-developed telson (last abdominal segment); tracheal system present or absent; malpighian tubules reduced to papillae; metamorphosis slight.

Order Thysanura (bristletails). Mouthparts ectognathous (exposed) adapted for biting; antennae many-segmented, only basal segment with muscles; compound eyes present or absent; tarsi of legs with 2 to 4 segments; 11-segmented abdomen ends in segmented median filament plus a variable number of lateral, styliform, pregenital appendages and a pair of many-segmented cerci; tracheal system and malpighian tubules present; metamorphosis slight or absent.

Order Diplura (or *Entotrophi*). Mouthparts entognathous; antennae with many segments, all with muscles; compound eyes and ocelli absent; tarsi of legs with one segment; abdomen without terminal median filament and with lateral styliform appendages on most pregenital segments, ends in paired cerci; tracheal system present; malpighian tubules absent or vestigial.

Order Collembola (springtails). Mouthparts entognathous, adapted for biting; antennae with 4 segments, the first 3 with muscles; compound eyes absent; abdomen 6-segmented, usually with a ventral tube on segment I, a retinaculum on segment III, and a forked springing organ (furcula) on segment IV; tracheal system usually absent; no malpighian tubules; metamorphosis absent.

Subclass Pterygota

Winged or secondarily wingless; metamorphosis; adults without pregenital abdominal appendages; adult mandibles (unless greatly modified) articulating with head capsule at 2 points.

Division Exopterygota (hemimetabola)

Metamorphosis simple, sometimes slight; pupal instar rarely present; wings develop externally; immature stages commonly resemble adults in structure and habits.

Order Ephemeroptera (mayflies). Soft-bodies with short setaceous (bristle-like) antennae and vestigial mouthparts; wings held vertically at rest, hind pair much reduced; intercalary veins and many crossveins present; abdomen with long cerci, and with or without a medial caudal filament; larvae (nymphs) aquatic, campodeiform (elongated and flattened) with tracheal gills of varied form; true adult preceded by a subimago (winged instar).

Order Odonata (dragonflies). Predaceous insects with biting mouthparts; 2 pairs of elongate membranous wings, each with a complex network of small crossveins and a conspicuous stigma; compound eyes very large and prominent; antennae short and filiform (threadlike); abdomen elongated and slender with male accessory armature on 2nd and 3rd ventral segments; larvae (nymphs) aquatic, with labium modified to form a prehensile organ for catching prey; breathing by rectal or caudal gills.

Order Plecoptera (stoneflies). Soft-bodied insects, some large with long, bristle-like antennae; mouthparts of biting type, but weak; wings membranous, folded back over the abdomen in repose; tarsi of legs with 3 segments; abdomen usually bears a pair of long, jointed cerci; young (nymphs) aquatic, campodeiform, usually with long antennae, cerci, and tracheal gills of varied type.

Order Dictyoptera (cockroaches, mantids). 2 pairs of wings; thickened forewings called tegmina; hindwings folded longitudinally fanwise; hindlegs similar to middle ones, adapted for running; tarsi of legs 5-segmented; mandibulate mouthparts, adapted for chewing.

Order Grylloblattodea (grylloblattids). Wingless; eyes small or absent; female ovipositor well developed; all legs similar, adapted for running; tarsi of legs 5-segmented; mandibulate mouthparts, adapted for chewing.

Order Phasmida (stick and leaf insects). Often wingless; when winged, tegmina often shorter than wings; all legs similar, adapted for walking; mandibulate mouthparts; no tympanum; female ovipositor short, often concealed.

Order Orthoptera (grasshoppers, crickets). 2 pairs of wings (forewings called tegmina); femur of hindleg enlarged for jumping; tarsi of legs usually with 3 or 4 segments; special auditory and stridulatory (sound-producing) organs often present; mandibulate mouthparts, adapted for chewing.

Order Dermaptera (earwigs). Elongated insects with chewing mouthparts; forewings reduced to very short leathery tegmina devoid of veins; hindwings semicircular with radially disposed veins, folded fanwise and then transversely; many species wingless; cerci unjointed and modified to form heavily sclerotized forceps.

Order Embioptera (webspinners). Gregarious insects inhabiting silken funnels; chewing mouthparts; both pairs of wings similar, with the radial vein (R) greatly thickened, other veins often reduced; short 2-segmented cerci; female wingless, larvalike.

Order Isoptera (termites or white ants). Social insects polymorphic in form (*i.e.,* live in large communities consisting of reproductive forms, wingless sterile soldiers, and young stages, or workers); biting mouthparts; wings alike, elongated, membranous, capable of being shed by basal fractures; anterior wing veins strongly sclerotized; fine network between other veins; cerci short; genitalia rudimentary in both sexes; show affinities with cockroaches.

Order Psocoptera (booklice or psocids). Small or minute insects with long filiform antennae, delicate membranous wings (though many are wingless), head with Y-shaped epicranial suture, enlarged post-clypeus (sclerite on the face); maxilla with a rodlike lacinia (inner lobe) partly sunk into head capsule; labial palps much reduced; cerci absent.

Order Zoraptera. Minute winged or wingless insects with 9-segmented moniliform (beadlike) antennae; Y-shaped epicranial suture; normal maxillae and labial palps; wings with simplified venation, capable of being shed at basal fracture lines; cerci short, with single segment.

Order Mallophaga (chewing lice). Wingless ectoparasites, chiefly on birds; eyes reduced, no ocelli, antennae with 3 to 5 segments; mouthparts modified for biting; prothorax distinct and free; mesothorax and metathorax often partly fused; thoracic spiracles ventral; cerci absent.

Order Anoplura or Siphunculata (sucking lice). Wingless ectoparasites of mammals; eyes reduced or absent; antennae with 3 to 5 segments; mouthparts, highly modified for piercing and sucking, retracted into head when not in use; thoracic segments fused; thoracic spiracles dorsal; cerci absent.

Order Thysanoptera (thrips). Minute slender-bodied; short antennae, piercing mouthparts; wings, when present, very narrow with much reduced venation and long marginal setae; cerci absent, metamorphosis includes 1 or 2 inactive pupal instars.

Order Homoptera (cicadas, hoppers, whiteflies, aphids, scale insects). Plant feeders; mouthparts adapted for sucking; beak arises from back of head; wings, when present, usually number four; front wings with uniform structure, either membranous or slightly thickened; wings at rest usually held rooflike over body; male scale insects with only 1 pair of wings; ocelli present or absent; compound eyes usually well developed.

Order Heteroptera or Hemiptera (true bugs). Mouthparts adapted for piercing-sucking, with slender segmented beak arising from front of head; basal portion of front wing (called a hemelytron) thickened and leathery, with tip membranous;

hindwings entirely membranous; wings at rest held flat over abdomen with membranous tips of front wings overlapping; ocelli present or absent; compound eyes usually well developed.

Division Endopterygota (holometabola)

Metamorphosis complex, accompanied by a pupal instar; immature stages differ from adult in structure and habits; wings develop internally during larval stages.

Order Megaloptera (alderflies, dobsonflies). Head prognathous (mouthparts located anteriorly on a horizontal head); biting mouthparts; filiform antennae; 2 pairs of large similar wings, at rest held rooflike or nearly flat over abdomen; larvae elongated, with biting mandibles.

Order Raphidiodea (snakeflies). Head prognathous, elongated; biting mouthparts; filiform antennae; prothorax elongated, cylindrical; 2 pairs of similar elongated wings, at rest held rooflike over abdomen; larvae elongated, flattened, with biting mandibles.

Order Neuroptera (lacewings). Head hypognathous (mouthparts located ventrally on a vertical head); antennae varied; 2 pairs of similar wings, at rest held rooflike over abdomen; larvae elongated or broad, with piercing, sucking jaws.

Order Mecoptera (scorpionflies). Slender carnivorous insects with elongated filiform antennae, head usually a downward directed rostrum, with biting mouthparts; legs long, slender; wings similar, membranous, carried longitudinally and horizontally in repose; abdomen elongated with short cerci; male genitalia prominent (recalling a scorpion's stinger); larva eruciform (caterpillar-like) with biting mouthparts, sometimes with abdominal feet as well as 3 pairs of thoracic legs; pupae exarate (appendages free).

Order Trichoptera (caddisflies). Small to moderate mothlike insects with bristle-like antennae; mandibles vestigial or absent; wings membranous, hairy, at rest held rooflike over the back; larvae aquatic, more or less eruciform, usually living in cases held by means of hooked caudal appendages; pupae exarate with strong mandibles.

Order Lepidoptera (butterflies, moths). 2 pairs of membranous wings, few crossveins; wings, body, and appendages covered with broad scales; mandibles normally vestigial or absent; maxillae forming coiled sucking proboscis; larvae eruciform, with spiracles on most segments, often with 5 pairs of prolegs on the abdomen in addition to the 3 pairs of thoracic true legs; pupae usually obtect (appendages glued to body).

Order Coleoptera (beetles). Minute to large in size; modified forewings are horny or leathery (elytra), usually meet to form a straight middorsal suture; hindwings membranous, folded below elytra, often vestigial or absent; mouthparts adapted for chewing; prothorax large and mobile, mesothorax much reduced; larvae varied (campodeiform and eruciform), pupae exarate, without articulated mandibles.

Order Hymenoptera (ants, bees, wasps, sawflies). 2 pairs of membranous wings, venation often much reduced; hindwings smaller and connected to forewings by a row of hooklets; mouthparts primarily of biting type but often adapted for sucking fluids; abdomen usually constricted at the base (except sawflies) with first segment fused to metathorax; ovipositor always present and modified for sawing, piercing or stinging; larvae usually legless with distinct head, in sawflies eruciform with locomotory appendages; pupae exarate, usually in a cocoon.

Order Diptera (true flies). 1 pair of membranous wings; hindwings modified to form halteres; sucking mouthparts, usually forming a proboscis, sometimes adapted for piercing; mandibles rarely present; labium usually expanded distally to give a pair of fleshy lobes; prothorax and metathorax small and fused, large mesothorax; larvae eruciform and legless, often with head reduced and retracted; tracheal system varied; pupae either free and exarate or enclosed in the hardened larval cuticle (puparium).

Order Siphonaptera (fleas). Small, wingless, laterally compressed; adults ectoparasitic on warm-blooded animals; compound eyes absent; usually a pair of ocelli; short antennae in grooves on head; mouthparts modified for piercing and sucking; both maxillary and labial palpi present; thoracic segments free; coxae of legs very large; larvae slender, eruciform and legless; pupae exarate, enclosed in cocoon.

Critical appraisal. The classification into orders presented above is acceptable to most entomologists. Although the apterygotes are classified here as four unrelated orders of the subclass Apterygota, three of these groups (proturans, collembolans, diplurans) are considered by some as separate classes equivalent to the class Insecta. In such a classification the subclass Apterygota includes only the thysanurans and relatives. For this alternative classification, see below in the section *Apterygote.*

The order Orthoptera infrequently includes, in addition to the grasshoppers and crickets, three other closely related groups (classified above as distinct orders): dictyopterans (cockroaches and mantids), grylloblattids, and phasmids (stick and leaf insects). The term orthopteran often is used as a common name for these four groups. For an expanded classification of orthopterans as four distinct orders, see below in the section *Orthopteran.*

The Mallophaga (chewing lice) and Anoplura (sucking lice), classified here as orders, sometimes are grouped as suborders of an order Phthiraptera, a group of obligate permanent ectoparasites of birds and mammals. For this alternative classification, see below in the section *Phthiraptera.*

The homopterans and heteropterans, here classified as separate orders, sometimes are considered as suborders of an order Hemiptera. Both groups have piercing-sucking mouthparts; for this reason they are believed to be related closely to each other. Some entomologists, however, consider distinguishing features other than the mouthparts sufficiently important to accord full ordinal status to each group. For further discussions of the taxonomic problems involved in classifying Homoptera and Heteroptera, see below in the introduction of these sections.

The term neuropteran frequently is used to describe three closely related groups, classified here as three distinct orders: Neuroptera (lacewings), Raphidiodea (snakeflies), and Megaloptera (dobsonflies and alderflies). Although the tendency has been to classify these groups as distinct orders, they sometimes are placed in the order Neuroptera. For a more detailed classification of the neuropterans as three orders, see below in the section *Neuropteran.*

Among the lepidopterans, members of the family Micropterigidae are more primitive than existing trichopterans (caddisflies). Although some entomologists treat them as a distinct order (Zeugloptera), see the *Annotated classification* following *Lepidoptera* for a scheme that places them in the order Lepidoptera.

The aberrant parasitic *Stylops* and its allies have been treated as the order Strepsiptera. The tendency now, however, is to include them in the order Coleoptera.

(V.B.W.)

MAJOR INSECT GROUPS

The remainder of this article consists of a review of the major groups of insects, identified at the ordinal (*i.e.,* order) level of biological classification and arranged in accordance with the *Annotated classification* above.

There are some exceptions to this scheme, however. Three groups—the apterygotes, the orthopterans, and the neuropterans—are not treated as orders, because of divided opinion regarding their taxonomic status (see above *Critical appraisal*). Extinct groups (indicated in the *Annotated classifications* by a dagger [†]) are treated throughout this article under the headings *Evolution and paleontology.* Some additional information on insect evolution is provided in GEOCHRONOLOGY: *Fossil record.*

Apterygote (springtails, bristletails, and allies)

Apterygotes have long been defined as primitive wingless insects of the subclass Apterygota (class Insecta), distinct from the subclass Pterygota or winged insects. Used in this sense, the term apterygotes commonly includes four groups of primitive insects: proturans, collembolans, diplurans, and thysanurans. The taxonomic status of these groups, however, remains unsettled.

A typical apterygote is wingless and has six legs. Although the presence of six legs was once an important feature in the identification of true insects, three groups commonly known as apterygotes (*i.e.,* proturans, collembolans, diplu-

rans) are now considered by some entomologists to be offshoots from the main insectan stem of evolution and have been given independent taxonomic status as classes equivalent to the class Insecta. The term apterygotes, therefore, has been redefined to include only those groups thought to be ancestors of pterygotes *i.e.,* the thysanurans (*e.g.,* silverfish, firebrats, and bristletails), together with the Microcoryphia (a group closely related to the Thysanura) and the extinct Monura. For completeness of discussion, and because of the similarities of these primitive hexapods, the proturans, collembolans, and diplurans, as well as the thysanurans, microcoryphians, and monurans are included in this article.

From (monuran) A.G. Sharov, *Basic Arthropodan Stock with Special References to Insects* (1966), Pergamon Press; (microcoryphian) R.J. Tillyard, "Primitive Wingless Insects," *New Zealand Journal of Science and Technology* 7(4) (1924); (others) *Invertebrate Identification Manual* by R.A. Pimental © 1967 by Litton Educational Publishing, Inc., reprinted by permission of Van Nostrand Reinhold Company

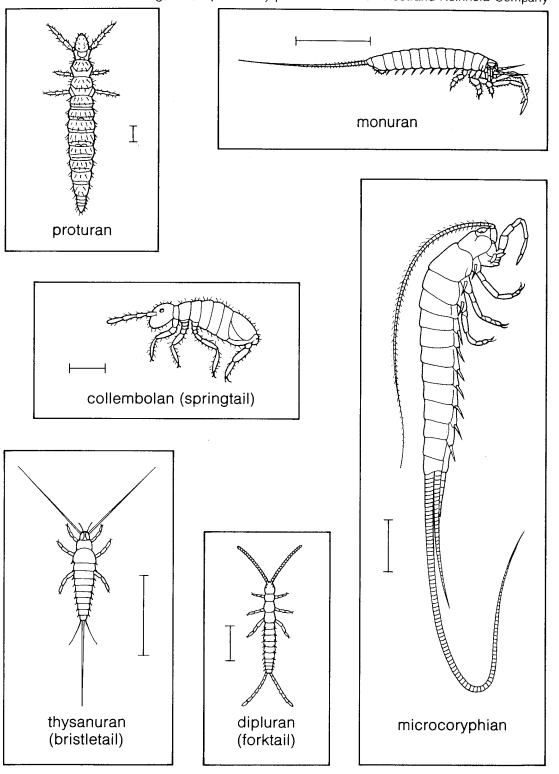


Figure 12: Types of hexapods. The line scales indicate an average length (in millimetres).

GENERAL FEATURES

Protura are minute (to two millimetres in length), elongated, white, and lack antennae. Distributed throughout the world in soil and leaf litter, they number about 152 species. Collembola are diverse in form, coloration, and habitat. Most species are less than three millimetres in length, but some range to ten millimetres. They are elongated or globular insects with antennae and may have a furcula (ventral abdominal springing organ). Collembolans occur in soil and leaf litter throughout the world including Antarctica. There are over 3,500 known species. Diplura are white or yellowish, blind, elongated with long antennae, and less than 10 millimetres (0.4 inch) in length although one group attains 50 millimetres. Their two tail filaments (or threadlike structures) can be long and thin, short and thick, or in the form of pincers. Diplurans are widely distributed in soil, leaf litter, and rotting logs. Over 400 species have been described. Thysanura and Microcoryphia are mainly long with three elongated tail filaments. Mostly 5 to 20 millimetres in length when fully grown, these insects are widely distributed in leaf litter; some live in ant and termite nests. Microcoryphians have large compound eyes and styli (*i.e.,* bristlelike processes) on legs and abdomen; thysanurans have small compound eyes and styli on abdomen only. More than 350 species of Thysanura and Microcoryphia are known. The extinct Monura were about 11 millimetres in length. Only two species are known from fossils.

NATURAL HISTORY

Life cycle. After hatching, the immature primitive wingless hexapods (called nymphs) change to adults with only slight metamorphosis. The young are similar to adults, changing little from molt to molt until sexual maturity is attained. Molting may continue throughout adult life. The greatest changes occur in the Protura, the only anamorphic hexapod group (*i.e.,* increase in number of body segments occurs at time of molting); the complete number of segments is present only after the third molt. There are at least six stages between molts (instars), and the last is the adult. Little is known about the postembryonic development of Diplura. They feed on living and dead vegetable matter and fungi; one group preys on other small invertebrates.

The life cycles of the Collembola are diverse although generally there are three stages (juvenile, postmaturity growth, senile molts) with no significant growth. Females may lay up to 800 eggs that hatch in 2 to 40 days. Three to 12 juvenile molts occur in intervals ranging from 11 days to a year; up to 50 molts occur in a lifetime (from 4–18 months). Collembola feed on living or decaying plant material, including fungi, algae, and spores; a few feed on carrion or are predatory.

In Thysanura there may be more than 40 instars; the adult one occurs after about 12 molts. Some silverfish live for seven years; they reach sexual maturity in two or three years and molt three to five times each year. In Microcoryphia there are six instars including adults. Both groups feed on decaying or dried vegetable material. Domestic silverfish eat plant and animal remains, paper, and artificial silk.

Reproduction

Reproduction in nonpterygote groups is mainly sexual, but parthenogenesis (reproduction without fertilization) can occur. Males deposit sperm packets, often haphazardly, that are taken up by females and stored until the time for fertilization. In proturans small external genitalia of unusual form surround the gonopore in both sexes; collembolans lack external genitalia, but the gonopore and surrounding area differ; in diplurans external genitalia are absent or vestigial. Thysanurans and microcoryphians have external genitalia similar to those of the pterygotes; however, the aedeagus in males is used to deposit sperm drops, and not as a copulative organ. The deposition and pickup of sperm drops in thysanurans and microcoryphians must take place during each adult stage if young are to be produced since the contents of the female sperm receptacle are lost with each molt.

Ecology. Most primitive hexapods live in soil and leaf litter found near the ground or associated with plants and rotting logs in moist regions. Collembolans are also found in aquatic environments, caves, permanent ice fields and snow, and insect and animal nests. Moisture is an important ecological factor in habitat selection. Most primitive wingless hexapods have a thin external covering and must guard against dehydration.

The condition and nature of soil are important factors in the distribution and nature of collembolan populations. Fungi in soil are especially important as collembolan food. Physical factors known to affect collembolans often do so secondarily; the primary effect is on the soil microflora. Changes in macroflora also affect collembolan populations. Direct cultivation (*e.g.,* plowing) has an immediate harmful effect on collembolan populations but does little long-range damage. Fertilizers increase the numbers of collembolans in soil, and herbicides have no effect. Some insecticides are lethal, while others are not. Increases in soil collembolans following insecticide applications are probably due to lethal effects on predators. Estimates of soil collembolans average about 100,000 per cubic metre. Soil collembolans usually form aggregations in soil. Certain species are more prevalent in a given layer (*e.g.,* leaf litter, the fermentation layer, the humus-forming layer, or deep in the soil). Typically small species and young insects occur in deeper layers, although daily and seasonal vertical movements may occur. Occasional swarming of large numbers of collembolans occurs after unusually wet conditions.

The most important collembolan predators are mites.

Others are pseudoscorpions, staphylinid and carabid beetles, an empid fly, and dacytine ants. Occasional collembolan feeders include spiders, fish, frogs, miscellaneous ants, and pulmonate snails. Collembolans are parasitized by gregarines, nematodes, virus, bacteria, fungi, and gordian worms.

FORM AND FUNCTION

The bodies of proturans, collembolans, diplurans, thysanurans, and microcoryphians are divided into three segments: head, thorax, and abdomen. The three-segmented, leg-bearing thorax has one pair of legs on each segment. The extinct monurans, considered an offshoot of the Thysanura–Microcoryphia stem, have no separate abdomen or thorax; three pairs of legs occur on the first three of 14 segments. The monuran body, although similar to present day Thysanura, has one tail filament or telson (the median tail filament in present-day Thysanura–Microcoryphia and winged mayflies). Many proturans and collembolans have entirely cutaneous respiration (breathing through the skin). In some proturans, collembolans, diplurans, and microcoryphians, a breathing system of tracheae and spiracles (external openings) occurs in the trunk, while in thysanurans the tracheal system is like that in pterygote insects.

Head. Proturans have anterior mouthparts (they are prognathous) and a pair of sense organs. Collembola are either prognathous or hypognathous (mouthparts are ventral) and may or may not have a pair of postantennal sense organs and lateral ocelli (simple eyes). The diplurans lack eyes and sense organs, while the thysanurans have simple lateral eyes and are hypognathous. Microcoryphia have both compound eyes and ocelli. Monurans have large compound eyes. Mouthparts are important taxonomically in primitive wingless hexapods. Movable mouthparts are generally recognized as walking limbs that were attached to ancestral body segments that became fused to form the posterior part of the head. The Protura, Collembola, and Diplura are entognathous (*i.e.,* the mouthparts are withdrawn inside the head capsule); Monura, Thysanura, and Microcoryphia are ectognathous (*i.e.,* the mouthparts are external, as in pterygote insects). Mandibles are present in the biting mouthparts of Diplura, Thysanura, Microcoryphia, Monura, and most Collembola. Although proturan mouthparts are modified for sucking, mandibles are retained. However, in collembolans with sucking mouthparts, mandibles may or may not be retained.

Mouthparts

Body. In general the three segments of the thorax are distinct, and each bears one pair of legs. Each leg has a terminal claw. Proturan legs are simple except that the anterior legs are longer, serving as tactile organs; the anterior tarsus has sensory organs. The three pairs of legs are similar to each other in the nonpterygote groups.

The abdomen of Protura undergoes anamorphosis: the first and second instars have 9 segments, the third 10, and the rest 12. Collembola have a maximum of six abdominal segments; fusion occurs between segments in some groups. Diplura, Thysanura, Microcoryphia, and the extinct Monura have 11 abdominal segments. The final abdominal tailpiece is the telson (see ARTHROPODS).

Modified abdominal structures. Primitive hexapods have abdominal structures that represent modified remnants of ancestral walking limbs. Many hexapods have cerci (sensory appendages) on the eleventh abdominal segment; these aid in identification of the telson. The Protura, Collembola, and Monura lack cerci. In Diplura a pair of cerci arise from the small terminal segment. Cerci can be long with numerous segments, short with a central duct and terminal pore, or modified into a pair of pincers for holding prey. They are important taxonomic criteria in separating hexapod groups. The extinct monurans have pairs of abdominal structures, called ventral styli, that are modified limbs. Similar styli are present in Thysanura and Microcoryphia; these hexapods also have one or two pairs of vesicles on some abdominal segments, as do proturans. Collembolans have a single tube containing a pair of vesicles, a single median tenaculum (catchlike process that holds furcula in place), and a single median furcula (springing organ); each organ is a modified pair of ancestral limbs.

Sensory appendages

Locomotion. The primitive wingless hexapods are walking terrestrial invertebrates. Proturans, however, use only the middle and posterior legs for locomotion; the anterior pair of legs are held above the head and used as tactile organs. Collembolans have a furcula. This ventral abdominal springing organ enables them to spring rapidly into the air. Apparently thysanurans and microcoryphians have developed the median tail filament (telson) to assist in supporting the abdomen during movement. Microcoryphians can also spring with considerable agility by using the body and tail filaments.

EVOLUTION AND PALEONTOLOGY

Evolution of the hexapods has occurred in conjunction with that of the myriapods. Both apparently had a common origin from early crustaceans. Among the myriapods, the Symphyla bear the closest relationship to hexapods, and both may have arisen from an ancestral symphylan stem. The evolution of Protura, Collembola, Diplura, and Thysanura plus Pterygota has usually been considered linear, with offshoots at successive intervals. The hexapodous state may have arisen several times; if this were the case, Protura, Collembola, and Diplura could have arisen separately. The extinct Monura and the Microcoryphia are offshoots from a thysanuran stem; earlier specialization may have led to the winged insects (pterygotes).

Fossil species

There are few fossil species of the primitive wingless hexapods. One extinct collembolan family (Protentomobryidae) contains a species (*Protentombrya walkeri*) of the Cretaceous (approximately 100,000,000 years) Period of Canada. The oldest fossil collembolan species, *Rhyniella praecursor* (family Neanuridae), is found in the Middle Devonian (approximately 370,000,000 years) sandstone of Scotland. Other species known from the Baltic amber include one campodeid dipluran. The extinct order Monura includes two species, *Dasyleptus lucasi* of the Upper Carboniferous (approximately 280,000,000 years) of France and *D. brongniarti* of the Siberian Permian (approximately 252,000,000 years) deposits. The extinct family Triassomachilidae (order Microcoryphia) includes *Triassomachilis uralensis* of the Triassic (approximately 210,000,000 years) deposits of Russia. Other extinct species occur in the genera *Machilis, Praemachilis,* and *Parastylus.* One thysanuran family (Lepidotrichidae), thought to be extinct, has now been found in California.

CLASSIFICATION

Distinguishing taxonomic features. Apterygotes differ from pterygotes in lacking wings and undergoing simple metamorphosis. They differ also in the structure of the thorax and in the development of abdominal appendages.

Mouthparts are an important criterion in separating these hexapods into two groups, those with entognathous mouthparts (Protura, Collembola, and Diplura) and those with ectognathous mouthparts (Thysanura, Microcoryphia, and the extinct Monura). These groups are further differentiated on the basis of antennae (number of segments) if present, placement of eyes (simple and compound) if present, segmentation and specialization of legs, number and modifications of abdominal appendages, and number of abdominal segments.

Annotated classification. Three previous orders (Protura, Collembola, and Diplura) are treated here as separate classes. The Thysanura, Microcoryphia, and Monura are placed in the subclass Apterygota (class Insecta). Groups indicated by a dagger (†) are extinct and known only from fossils.

CLASS PROTURA

Entognathous mouthparts.

Order Protura

Antennae and eyes absent; anterior pair of legs elongated and sensory; 8th abdominal segment with paired openings to large abdominal glands; 3 pairs of ventral abdominal appendages; sparse setae (hairs or bristles) important for classification; scales absent.

Suborder Eosentomoidea

Tracheal system present; claw of middle and posterior legs claw-shaped; 8th abdominal segment with striate band; lids to gland openings small, unornamented.

Family Eosentomidae. Pairs of abdominal appendages alike, 2-segmented with 5 setae.

Suborder Acerentomoidea

Tracheal system absent; claw of middle and hindlegs broadly boat-shaped; lids to gland openings large and with or without teeth on posterior border.

Family Protentomidae. Abdominal appendages 2-segmented with up to 4 setae, 2 pairs with terminal vesicles.

Family Acerentomidae. Abdominal appendages 2-segmented with up to 4 setae, anterior pair with terminal vesicle.

CLASS COLLEMBOLA

Entognathous mouthparts.

Order Collembola

1 pair of 4- to 6-segmented antennae; a pair of postantennal organs; 0 to 8 simple eyes (ocelli) on each side of head; abdomen of 6 or fewer segments including telson, a ventral tube on 1st abdominal segment, usually a ventral furcula on 4th abdominal segment, tenaculum on 3rd abdominal segment; hairs, setae, bristles, or scales present.

Suborder Arthropleona

Head prognathous (except Poduridae, Actaletidae); postantennal organ present or absent; trunk elongated, most segments distinct.

Section Poduromorpha. Trunk segments essentially similar; 1st thoracic segment distinct with dorsal setae; antennae short; scales absent.

Family Onychiuridae. Mouthparts mandibular; complicated sensory organs on antennae; ocelli absent.

Family Hypogastruridae. Mouthparts mandibular; sensory organs of antennae simple; ocelli usually present.

Family Neanuridae. Mouthparts modified for sucking, mandibles present or absent; sensory organs of antennae simple; ocelli usually present.

Family Poduridae. 1 species, *Podura aquatica;* hypognathous head; mouthparts mandibular; ocelli present.

Section Entomobryomorpha. Trunk segments usually dissimilar; 1st thoracic segment reduced dorsally, without dorsal setae; antennae usually long; scales present or absent.

Family Actaletidae. Head hypognathous; mouthparts mandibular.

Family Isotomidae. Postantennal organ usually present; scales absent.

Family Oncopoduridae. Postantennal organ present or absent; scales present.

Family Tomoceridae. Postantennal organ absent; scales present, distinctly ribbed or fluted.

† *Family Protentomobryidae.* Extinct.

Family Entomobryidae. Postantennal organ generally absent; antennae and furcula long; scales present or absent.

Suborder Symphypleona

Head hypognathous; postantennal organ absent; thorax and abdomen fused, segmentation lost or indistinct.

Family Neelidae. Antennae small, shorter than head, situated midway or on anterior half; ocelli absent.

Family Sminthuridae. Antennae longer than head, situated on posterior half; ocelli usually present.

CLASS DIPLURA

Entognathous mouthparts.

Order Diplura

Antennae multijointed; compound eyes and ocelli absent; abdomen of 11 segments, telson included or absent; paired ventral styli; paired cerci of variable form; setae, rarely scales.

Family Campodeidae. Antennae longer than head plus thorax; 1st abdominal segment with lobelike appendages; cerci as long as abdomen, multijointed.

Family Projapygidae. Antennae shorter than head plus thorax; 1st abdominal segment with styli; cerci less than half abdomen length, fewer than 10 joints; scales absent.

Family Japygidae. Antennae shorter than head plus thorax; 1st abdominal segment with styli; cerci paired terminal forceps; scales absent.

CLASS INSECTA

Ectognathous mouthparts.

Subclass Apterygota

Primitive wingless insects.

† Order Monura

Extinct; compound eyes present; trunk not obviously separated from thorax; telson a long median tail filament.

† *Family Dasyleptidae.* Fine hairs; abdominal styli 2-jointed.

Order Microcoryphia

Compound eyes present; 3 ocelli; telson a long median tail filament; cerci as lateral tail filaments; styli on thoracic legs and abdomen; setae and scales.

† *Family Triassomachilidae.* Extinct.

Family Machilidae. Abdominal segments with 1 or 2 pairs of vesicles, median ventral sclerites visible.

Family Meinertellidae. Abdominal segments with 1 pair of vesicles, median ventral sclerites almost invisible.

Order Thysanura

Small compound eyes present or absent; trunk segments flattened; telson a long median tail filament; telson and cerci long or short; abdominal styli unjointed.

Family Lepismatidae. Eyes present; scales usually present.

Family Nicoletiidae. Eyes absent; scales present or absent.

Critical appraisal. Detailed studies of morphology, physiology, embryology, and life histories are required to further elucidate relationships of these hexapods.

The standard system of classification has been class Insecta; subclass Apterygota; orders Protura, Collembola, Diplura (or Aptera), and Thysanura; subclass Pterygota. This system is retained in many textbooks, but it is no longer accepted by most specialists, who have shifted the entognathous orders to separate subclasses or classes, and the culmination is the classification used above. Another classification is superclass Hexapoda, class Protura, class Collembola, class Diplura, class Thysanura, class Insecta. Under this system apterygotes are primitive wingless classes of Hexapoda. (K.A.J.W.)

Ephemeroptera (mayflies)

The order Ephemeroptera comprises the group of insects known as mayflies. Other common names for the winged stages are shadfly, sandfly, dayfly, fishfly, and drake. The aquatic immature stage, called a nymph or naiad, is widely distributed in freshwater; a few species can tolerate the brackish water of marine estuaries.

The winged stages attract attention through mass emergences when they may make roads slippery, clog gutters, and taint the air with an odour of decay. Mayfly nymphs are important in the energy transfer cycle that occurs in freshwaters (winged stages cannot feed). Some species are carnivorous, but the majority of nymphs feed on diatoms, algae, higher plants, and organic detritus; nymphs are devoured in turn by many carnivorous animals, especially fishes.

GENERAL FEATURES

Appearance. Winged mayflies have large compound eyes; short, bristlelike antennae; and functionless mouthparts and digestive tracts. Their membranous wings include a front pair and a hind pair, which is much smaller than the front pair or absent; in repose, the wings are held together upright over the body like those of a butterfly. The adult mayfly has two or three threadlike tails, usually as long as, or longer than, the body.

Nymphal characters

Nymphal characters include a single claw terminating each of the six legs. The surface of the thoracic region of the body is strongly rounded outward and bears the developing wings in external pads on the upper surface; the abdominal region usually is long and slender. Gills are attached to the outer edge of the upper surface of some of the ten segments into which the body is divided. The body of the nymph terminates in three, less often two, slender tails. Adult mayflies of North American species range in body length, exclusive of tails, from 2.5 millimetres, (0.1 inch) for *Caenis* to 32 millimetres (more than an inch) for *Hexagenia.*

Distribution and abundance. Worldwide, about 2,000 species of mayflies have been described, nearly 600 of them from North America north of Mexico. The order is represented in all continents except Antarctica. In areas of high biological productivity (*e.g.,* gravel-bottomed, hard-water, temperate-zone streams), as many as 1,400 nymphs have been found in one square foot of surface, and one gravel riffle has yielded 33 species.

NATURAL HISTORY

Life cycle. The life cycle of mayflies consists of four stages: egg, nymph, subimago, and imago. Eggs, which

vary widely in size and surface detail, may be oblong, oval, or rounded. Depending on the species, a female may produce fewer than 50 or more than 10,000 eggs. Eggs are laid in water and either settle to the bottom or adhere to some submerged object; they often hatch in about two weeks but may, under certain circumstances, undergo a period of varying duration in which no growth occurs. This cessation of growth, known as diapause, is a highly effective adaptation that enables the insects to avoid environmental conditions hostile to developing nymphs or to emerging winged stages.

Nymphal life may be as short as two weeks or as long as two years; an annual cycle is most common. As many as 50 molts (periodic shedding of skin) may occur, depending on the species and the environment. When growth is complete, the nymphal skin splits down the back and a winged form, called the subimago, or dun, emerges. The subimago flies from the surface of the water to some sheltered resting place nearby; after an interval lasting a few minutes to several days, but usually overnight, the skin is shed for the last time, and the imago, spinner, or adult stage emerges. Mayflies are the only insects that molt after developing functional wings. The subimago resembles the imago in overall appearance, although it is softer and duller in colour than the adult. The wings of the subimago, generally rather opaque, are tinted with gray, blue, yellow, or olive; heavy pigmentation along the veins may give the wings of the subimago a mottled appearance that rarely persists in the imago. Legs and tails of the subimago are shorter than are those of the imago. It is often incorrectly assumed that the two stages are different species.

Subimago and imago stages

Mating and egg deposition. Mating takes place soon after the final molt. In most species death ensues shortly after mating and oviposition (egg deposition). Winged existence may last only a few hours; *Hexagenia* males, however, may live long enough to engage in mating flights on two successive days, and female imagos that retain their eggs may live long enough to mate on either of two successive days. Groups of male imagoes perform a mating flight, or dance, over water as dusk approaches, flying into any breeze or air current. Individuals may fly up and forward, then float downward and repeat the performance. Females soon join the swarm, rising and falling as the dance continues. The male approaches the female from below and behind and grasps her thorax with his elongated front legs; mating is completed on the wing. After her release by the male, the female deposits her eggs and dies. A few species are ovoviviparous—*i.e.*, eggs hatch within the body of the female.

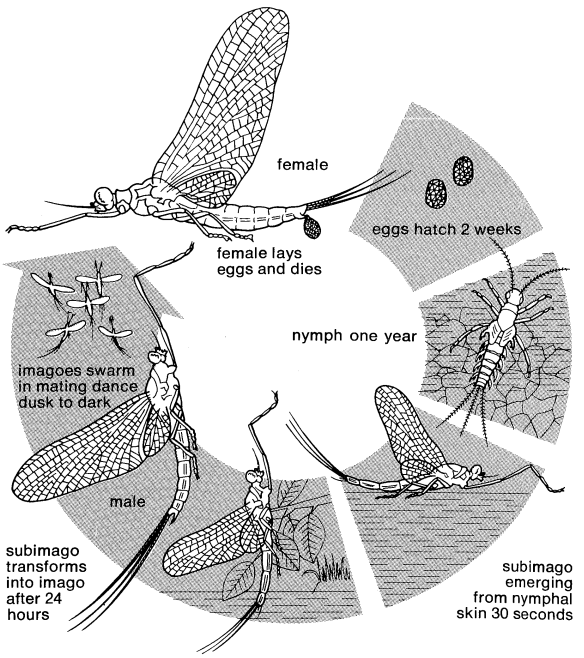


Figure 13: Life cycle of the mayfly.

Methods of oviposition vary. Some species drop the rounded egg mass from a height of several feet in a manoeuvre suggestive of dive-bombing; in others, the female flies low over the water's surface, striking it at intervals with the tip of her abdomen and washing off a few eggs each time she strikes the water. Still other females extrude the eggs from two oviducts as two long packets, which usually adhere to each other; they may be dropped from a foot or more above the water, but more often, the female falls to the surface with wings extended and squeezes out the eggs as she dies. In a fourth type of oviposition, the female alights on some object protruding from the water and crawls under the surface, depositing the eggs while submerged. Females, unless they drop the eggs from a height of several feet, are vulnerable to feeding fishes. Mayflies sometimes mistake blacktopped roads for streams, forming swarms over them, and drop eggs on road surfaces.

Ecology. Mayfly nymphs are preyed upon by carnivorous invertebrates and fishes. Winged stages are devoured in flight by birds, bats, and predatory insects, including dragonflies, robber flies, and hornets. When at rest, mayflies may be preyed upon by spiders, beetles, birds, and certain mammals, especially flying squirrels in North America. During their transformation to the adult stage and especially during oviposition by females, mayflies are vulnerable to predation by fishes; artificial lures used by fishermen are patterned after them.

FORM AND FUNCTION

Adaptations of form and function presumably determine distribution. The legs and jaws of some nymphs are modified for burrowing in silt or sand; other species are flattened, so that they can enter narrow crevices or cling to bottom materials in swift currents; long, slender legs and body adapt others for clambering on submerged vegetation. Strong swimmers are long and slender and occupy a variety of habitats. Gills may be platelike, feathery, or filamentous; often they are modified for specialized functions.

PALEONTOLOGY AND CLASSIFICATION

Recognizable mayflies occur in the fossil record of the Upper Carboniferous Period (about 325,000,000 to 280,000,000 years ago), and they appear to have been abundant during the Permian. Represented largely by wing impressions, the fossil record is so incomplete that most systems of classification and interpretations of relationships are based on characteristics of recent forms, chiefly their morphology.

Distinguishing taxonomic features. Characteristics of the male genitalia are the most reliable means for identification of adult species. Many other features, including patterns of veins in the wings, affect generic and other higher categories of classification.

Annotated classification. The classification below is modified from that of George Edmunds, Jr. (1962).

ORDER EPHEMEROPTERA (mayflies)

Soft-bodied insects; life cycle consisting of 4 stages—egg, nymph, subimago, imago; wings membranous, at rest held vertically upward; hind wings reduced; mouthparts and digestive system of adults nonfunctional; only insect to molt after developing functional wings; antennae bristlelike; 5 superfamilies—Heptagenioidea (families Siphlonuridae, Baetidae, Metretopodidae, Oligoneuriidae, Heptageniidae, and Ametropodidae); Leptophlebioidea (families Leptophlebiidae, Ephemerellidae, Tricorythidae); Ephemeroidea (families Behningiidae, Potamanthidae, Euthyplociidae, Polymitarcidae, Ephemeridae, Palingeniidae); Caenoidea (families Neoephemeridae, Caenidae); Prosopistomatoidea (families Baetiscidae, Prosopistomatidae); about 2,000 species on all continents except Antarctica.

Critical appraisal. In addition to the scheme above, another classification increases the number of families to 25. Before the taxonomy is established with certainty, more attention must be given to nymphal characters, not only their morphology but also their physiology and behaviour. Many species remain undescribed. (J.W.L.)

Odonata (dragonflies)

Large, active by day, and often strikingly coloured, dragonflies (order Odonata) are among the few insects that have secured a major place in folklore and art. Though they cannot inflict injury to man, longstanding vernacular names such as "horse stinger" and "devil's darning needle" testify to their formidable appearance. In Japan, where a journal (*Tombo*) is devoted solely to reports of their biology, dragonflies traditionally have been held in high regard.

Dragonfly is the common name for large, winged, sun-seeking insects comprising the order Odonata. Adult dragonflies have two pairs of narrow transparent wings, a long slender abdomen, and prominent eyes; they can be recognized by the skewed thorax, forwardly directed legs, and small inconspicuous antennae. Typically they fly near ponds and rivers in which their larvae develop. Dragonflies are predatory insects; the adults catch small flying insects on the wing. Although they consume mosquitoes and other insects troublesome to man, they are catholic feeders and not considered of appreciable value in controlling specific pests.

Distribution and size range

The order Odonata is small, well known, and widely distributed. The world fauna probably does not greatly exceed the 5,000 species now known. Dragonflies occur from the tropics, where they are most numerous and varied, to the latitudinal limit of trees. At a medium-sized pond, populations seldom exceed a few hundred adults of each species. Though the structure of adults is uniform, there is a wide range of sizes, and the wingspans of the smallest and largest species are about 1.8 and 19.3 centimetres (0.7 and 7.5 inches), respectively.

NATURAL HISTORY

Since it is fully aquatic, the larva (or nymph) differs markedly in structure and behaviour from the flight-oriented adult. During development the larva molts approximately 8 to 15 times and passes through a corresponding number of stages or instars; the number of stages varies within and between species. There is no pupal stage.

A few minutes after hatching from an egg, the sheathlike cuticle of the first instar larva (prolarva) splits and releases the spider-like second instar. The early instar larvae feed actively on various small water animals including minute Crustacea and Protozoa; during later instars, the larva feeds on midge larvae, aquatic beetles and snails, and even small fishes. Wing pads appear at an early stage and grow larger with subsequent molts. Toward the end of the last instar, the organs become adult inside the larval skin. A few days later the larva climbs out of the water and molts to disclose the adult, a process known as emergence. The largest dragonflies usually leave the water after sunset and take to flight just before sunrise; therefore, their emergence is seldom witnessed. Smaller species may emerge during the daytime.

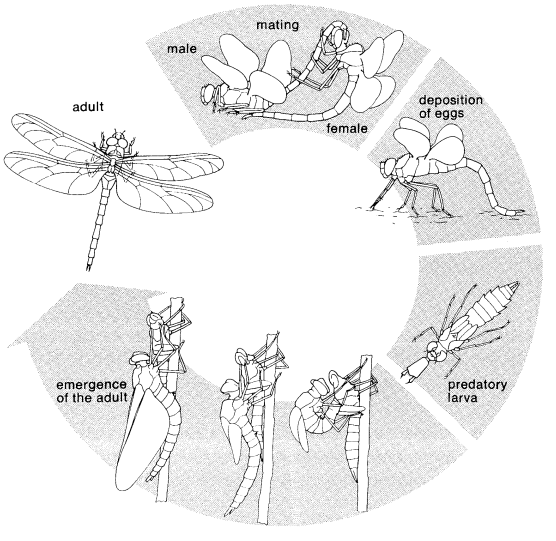


Figure 14: Life cycle of the dragonfly.

The newly emerged adult dragonfly is soft, reproductively immature, and lacks full coloration. One of its first actions is to fly away from water. Adult life consists of two stages, a prereproductive (or maturation) period that lasts two days to two weeks (depending on the species and weather) and a reproductive period that lasts one to six weeks.

Mating behaviour

During the first stage the adult feeds actively, away from water. The second stage begins when, as a sexually mature adult, the dragonfly flies to the mating rendezvous, usually a pond or river where the eggs will be laid. Males assemble there slightly earlier than females and space themselves along the shore or over the water. Each male, once established, defends an area of characteristic extent, much as birds defend territories. When a male adult approaches or enters a "territory" occupied by another individual of the same species, the occupant acts aggressively and a contest often ensues; thus, territories are held by the most vigorous males. If a female adult approaches or enters a territory the resident male tries to mate with her. In some species mating is preceded by a courtship display during which, by her responses, the female accepts or rejects the male. Usually, however, there is no evident prelude, and copulation occurs immediately.

The mating posture is unique among insects. With the claspers at the end of his abdomen, the male grasps the top of the female's head or prothorax; then, by his movements he induces her to bring the tip of her abdomen forward so that it meets the intromittent organ at the base of his abdomen; in this way the female receives a sperm capsule. This "copulation wheel" can be formed in flight, though mated adults usually alight promptly. Copulation lasts between a few seconds and several hours, depending on the species. After mating, the female may lay eggs immediately or after a delay of hours or days. If she lays immediately, she may do so alone, with her partner still attached in tandem, or with him hovering nearby darting at other males that approach. In this way females lay eggs with little interference.

Eggs are laid in one of two ways. Species with a well-formed ovipositor place them within or on plant tissue, above or in the water. Some climb beneath the water surface to lay and remain submerged for an hour or more. Species without an ovipositor wash the eggs off the end of the abdomen or drop them onto the water surface. Eggs of some species are laid in running water and usually possess adhesive or tangling devices that prevent their being swept downstream.

Requirements for reproduction

In order to reproduce, the adult dragonfly requires warm weather, food, and water in which to lay its eggs. In equatorial waters reproduction occurs throughout the year; elsewhere, however, the life cycle is modified so that the reproductive stage coincides with the appropriate season. Where rainfall is a seasonal constraint, as in the tropics, the adult survives dry periods in one of two ways, either by retiring to humid woodlands (*Lestes, Gynacantha*) or by becoming nomadic and following the moving rainbelt, laying eggs in temporary pools that appear in its path (*Pantala, Tramea*). In both cases the larval stage is brief (one to three months), and the adult stage relatively long. Outside the tropics, where the seasonal constraint is imposed by low temperatures rather than drought, winter is passed in a resistant stage. In a few species the resistant stage is the immature adult and sexual activity is postponed until spring (*Aciagrion, Indolestes, Sympecma*); in other species it is the egg (*Lestes, Aeshna*); but most often the resistant stage is the larva, which, depending on the species and location, may require one to five years to complete development. Thus the larval stage lasts considerably longer than the adult one, which may last only a few weeks. In temperate regions, each species flies at a characteristic time of year, as determined by larval responses to climatic variables such as temperature and day length.

Dragonfly larvae are eaten by fishes, birds, and each other; predators at the time of emergence include birds and small crocodiles. Once they have left the emergence site, the adult insects have few regular enemies. In flight they are able to evade almost all predators by their agility.

Larvae and adults are parasitized by flukes (Trematoda) and water mites (Hydracarina), and eggs by minute wasps.

FORM AND FUNCTION

Adaptations to flight

Adult structure. The principal structural features of adult dragonflies reflect adaptations to flight. The large wings are strengthened with a network of veins, and each has a thickened patch (the pterostigma) on the anterior edge near the tip. In certain migratory Anisoptera (*Libellula, Pantala, Tramea*), the expanded hindwing permits gliding flight. The sloping thorax accommodates large wing muscles and sets the legs forward where they can grasp prey. Where the head joins the thorax, a delicate orientation organ maintains equilibrium during flight. The large compound eyes, acutely responsive to movement and form, play an important role when adults capture food or interact with other dragonflies. In most Anisoptera the compound eyes meet dorsally, and in certain Aeshnidae that fly only in subdued light (*e.g., Gynacantha*) the eyes occupy almost the entire surface of the head.

The unique method of copulation exhibited by Odonata can be regarded as a special way of transferring a sperm capsule. Instead of placing the capsule on the ground or a leaf for the female to pick up (as certain primitive insects do), the male puts it on his own body, where accessory genitalia at the base of the abdomen are especially constructed to receive it and transmit it to the female during flight. Interlocking structures on the male claspers and female head or prothorax make the copulation wheel secure.

Adaptations to aquatic environment

Larval structure. Larval structure, though broadly similar to that of the adult, reflects demands of the aquatic environment, particularly with respect to respiration, locomotion, feeding, and concealment. Oxygen is replenished by diffusion through gills; in most Zygoptera these are three leaflike plates at the tip of the abdomen, while in most Anisoptera they are outgrowths of the hindgut wall, ventilated as water is pumped in and out of the anus. Both respiratory systems are used in emergencies as methods for locomotion; the jet propulsion employed by Anisoptera is particularly effective. Unlike the adult, the larva exhibits wide variation in form. Species living within fine sediment are cylindrical, with a terminal siphon that maintains respiratory contact with the water above (*Aphylla*); those residing on the surface of mud or sand often are flattened, with lateral spines (*Ictinogomphus*); those living among plants near the surface are streamlined and very active (*Anax, Lestes*).

The larva captures prey in an unusual way. Typically it remains motionless until it detects a victim by sight or touch. When the prey comes within range, the larva shoots out a prehensile organ (the labium) that grasps the prey and draws it back to the mandibles. In the resting position the labium lies beneath the head and thorax, sometimes obscuring the front of the head like a mask. It can be extended almost instantaneously by a localized increase in blood pressure that is controlled by a muscular diaphragm in the abdomen.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

The extinct Protodonata possessed a complete series of alternate convex and concave wing veins; the two suborders, Palaeodonata and Archodonata, are known from the Upper Permian and Upper Carboniferous, respectively. The Odonata, in which the series of wing veins lacks the posterior media and anterior cubital, contain four extinct and three living suborders. The extinct ones are Meganisoptera (Upper Carboniferous and Upper Permian), which included the gigantic *Meganeura monyi* with a wing-span of 70 centimetres; Protanisoptera (Lower and Upper Permian); Protozygoptera (Lower Permian); and Archizygoptera (Triassic and Jurassic). Living suborders are listed below.

Distinguishing taxonomic features. Features used by taxonomists in classifying members of the order Odonata are, for adults, the shape and venation of the wings, distance between the compound eyes, form and development of caudal appendages, and presence of an ovipositor. Larvae are classified according to the type and form of respiratory organ, shape and armature of labium, number and arrangement of body spines, and shape of abdomen.

Annotated classification. The classification given here is that of F.C. Fraser (1957); it takes account of the fossil record of this group, which is unusually long and rich. In this scheme the superorder Odonatoidea (Odonatoptera of Lameere) comprises two orders: the Protodonata and Odonata.

ORDER ODONATA

Carnivorous insects with aquatic larvae (except *Megalagrion*), terrestrial and winged adults; adults with two pairs of similar richly veined wings and a pterostigma, small movable prothorax, large obliquely set thorax with forwardly placed legs, intricate accessory genitalia on second and third sternites in the male; larvae with labium modified to seize prey, gills in hindgut or as caudal or lateral abdominal structures; about 5,000 known species, of which all but two are Zygoptera or Anisoptera.

Suborder Zygoptera (damselflies)

Upper Permian to present; both pairs of wings similar, petiolate at the base; eyes more than one diameter apart; male adult with four caudal appendages, two superior and two inferior, female with ovipositor well developed; larva with caudal or lateral gills, gizzard with 8–16 radially symmetrical fields; one extinct, 17 living families; small, delicate, weakly flying insects; long slender abdomen attains greatest development in *Mecistogaster; Megalagrion oahuense,* small forest dweller in Hawaii, only known member of the Odonata with a terrestrial larva; size: body length about 1.9–15.0 cm, wingspan about 1.9–19.3 cm.

Suborder Anisoptera

Jurassic to present; well represented in Tertiary; hindwings broader at base than forewings; wings not petiolate at base; eyes less than one diameter apart or touching; male adult with three caudal appendages, two superior and one inferior, female with ovipositor well developed (Aeshnidae, Petaluridae) or reduced; larva characterized by rectal gills, anus closed by three robust spines, gizzard with four to eight dental folds; eight living families; robust, strongly flying, agile insects; some accomplished migrants; size: body length about 1.8–11.9 cm, wingspan about 2.4–15.7 cm.

Suborder Anisozygoptera

Mainly Mesozoic; adults intermediate in structure and appearance between the other two living suborders; larvae resembling Anisoptera; 11 families, all extinct except one (Epiophlebiidae), known from two similar species inhabiting highland streams in Japan and Nepal; size: body length about 5.1 cm, wingspan about 6.5 cm.

Critical appraisal. The members of the order Odonata, as the only living descendants of the Odonatoidea, occupy a uniquely isolated position in the phylogeny of insects, having evidently separated from all other winged groups at a very early stage of evolution. There is general agreement among specialists regarding the status and affinities of living families and genera of the Odonata and, with few exceptions, classifications based on the adult and larva correspond. The evolution of the two major suborders has, however, been a matter for continuing debate among entomologists. Some of these authorities consider that the suborders Zygoptera and Anisoptera arose independently from the Protodonata. According to Fraser (1957), the Anisoptera, in their descent from the Protodonata, passed through a zygopterous stage. This view of the genealogy of the order Odonata has gained wide acceptance. (P.S.C.)

Orthopteran (cockroaches, mantids, stick and leaf insects, grasshoppers, crickets)

Orthopteran has come to be regarded as a common name for several related groups of insects that exhibit considerable morphological, physiological, and paleontological diversity. Although sometimes these insects are combined into the order Orthoptera, generally, several orders are implied in the term orthopteran. Among the orthopterans, cockroaches and mantids are placed in the order Dictyoptera; the grylloblattids (order Grylloblattodea) and walking sticks (order Phasmida) are given ordinal rank also. On the other hand, members of the suborders Ensifera (katydids, crickets, and camel crickets) and Caelifera (pygmy sand crickets, grasshoppers, and locusts) are considered to comprise the order Orthoptera. For completeness of discussion, all of these groups, handled here as four separate orders, are included in this article.

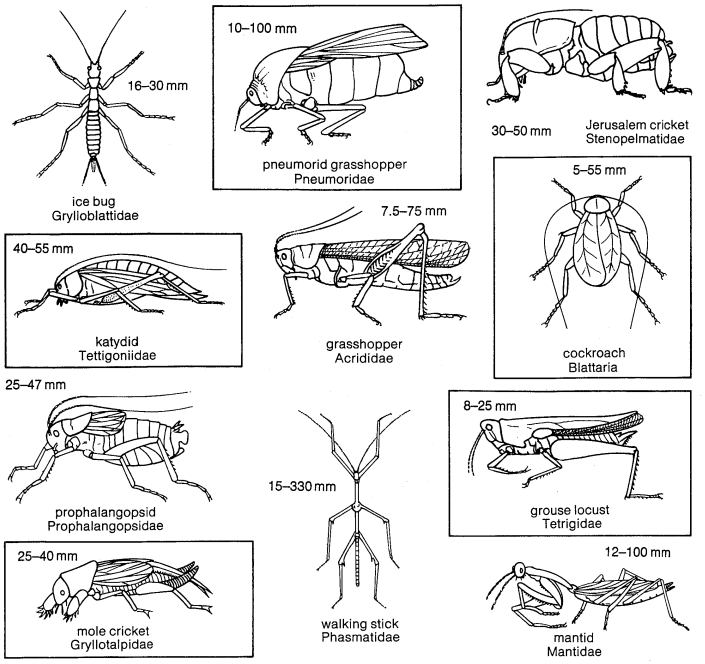


Figure 15: Diversity among orthopterans.

Orthopterans, abundant in tropical regions throughout the world in both numbers of species and individuals, are common in the summer months in temperate regions, when their relatively large size and chirping sounds attract considerable attention. Zoologists have long been interested in cockroaches, one of the oldest insect groups known. Most of the 24,000 species of orthopterans are plant feeders, with mouthparts adapted for chewing. Locusts, known as pests since biblical times, are very destructive to agricultural products.

GENERAL FEATURES

Orthopterans may be bizarre in appearance, unusually large in size, or show peculiar behaviour. They range in size from a few millimetres to more than 30 centimetres. Some tropical walking sticks resembling tree twigs are more than 30 centimetres (11.8 inches) long, and others, much smaller, resemble leaves of plants. The size range of present-day cockroaches is typical of the diversity of body size among orthopterans: tiny flightless cockroaches (*Attaphila*), living as commensals in the nests of ants, are only two millimetres long when mature, whereas a species of *Megaloblatta* found in South America reaches 10 centimetres in length with a wing span of almost 19 centimetres.

Distribution and abundance

Approximately 24,000 species of orthopterans have been identified. Throughout the U.S. there are about 1,300 species; not all of them inhabit any one place. There are more species (*e.g.*, 282 in Arizona) in southern and southwestern sections of the U.S. than in the North (*e.g.*, 103 in all the New England states). Of the 600 species found in Europe, Great Britain has only 35, including four established, introduced species (adventives).

The largest families of orthopterans are worldwide in range, although all have decreased numbers of species in cold temperate zones. Few mantids or walking sticks, for example, occur outside tropical or subtropical areas. There are about 20 mantid species and 27 walking stick species in the southern regions of the U.S., compared with the 400 mantid species and 600 walking stick species that have been identified in Central and South America. A few northern groups include the grylloblattids and several genera of grasshoppers.

IMPORTANCE

Among the orthopterans are many species that are either harmful to agricultural products or are pests. Grasshoppers are capable of causing widespread devastation of the agricultural crops grown in many countries throughout the world. In cattle-growing regions there often is competition between grasshoppers and livestock for available forage. Mormon crickets (a common name for species of the genus *Anabrus* that originated during the early years of the Mormon settlement in Utah) are major pests of both crops and open rangeland in the western part of the U.S. during seasons that are favourable for their development. Cockroaches, known throughout the world as domestic pests, are a frequent nuisance, especially in warm-temperate to tropical areas. Although cockroaches occasionally carry organisms such as bacteria or parasites that produce intestinal diseases, they are more generally considered to be mechanical carriers of contaminating filth.

Mantids, predators on other insects, have become adapted to resemble the flowers, tree trunks, or grass stems on which they await their prey. Crickets, katydids, and grasshoppers are known for the songs they produce using stridulatory mechanisms, and research concerned with song production is an active field. The biology of migratory grasshoppers or locusts involves hormones that promote transformation of nonmigratory, solitary, short-horned grasshoppers into gregarious hordes of locusts capable of causing great destruction. This transformation has been studied in attempts to control these pests.

NATURAL HISTORY

Simple metamorphosis

Life cycle. *General features.* Since orthopterans undergo simple metamorphosis and have externally developing wings, they are known as hemimetabolous insects. The grylloblattids are wingless, and all large orthopteran groups contain a few wingless species, even though the basic structure of the orthopteran thorax proves their relationship to winged insects. A typical orthopteran life cycle has three stages: egg, nymph, and adult. Usually eggs are

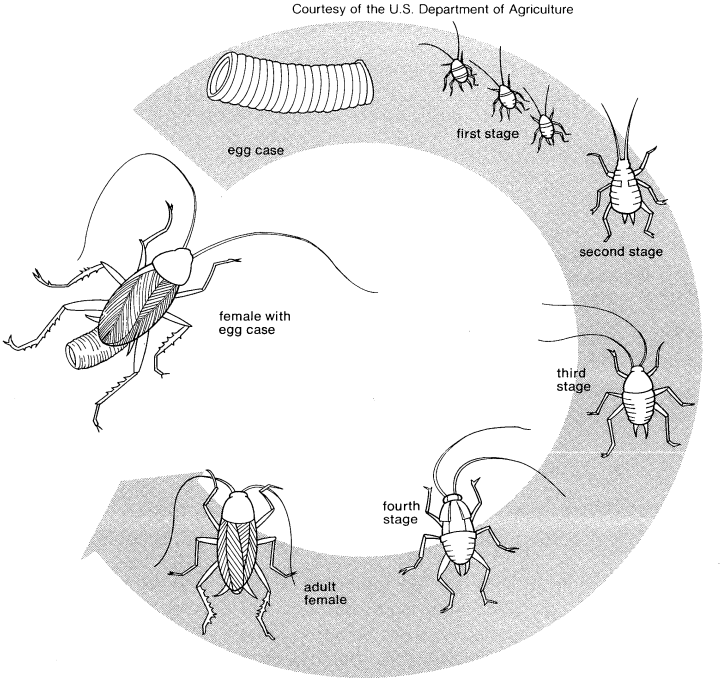


Figure 16: Orthopteran life cycle.

deposited outside the body on the ground or on vegetation; however, in some cases (*e.g.*, viviparous cockroaches), eggs are incubated in a brood chamber within the body of the female, and nymphs are born alive. Nymphs resemble adults except for their smaller size and lack of development of reproductive organs and wings; there is no pupa, or resting stage. In most orthopteran groups, the hatching insect that wriggles from the egg is not a fully formed nymph with freely moving legs; actually it is little more than an active embryo and is still enclosed in a thin membrane. This stage is called a vermiform larva; shedding of the enclosing membrane occurs at the intermediate molt. The shapeless skins shed by young grasshoppers crawling

from egg pods or by mantids leaving an egg case are examples of such exuviae (cast skins).

The number of nymphal stages between the intermediate molt and adulthood varies from about 4 to 13. Generalizations are approximately as follows: cockroaches, 5–13; mantids, 4–9; grylloblattids, about 8; crickets and katydids, 5–9; walking sticks, 4–6; and grasshoppers, 4–9, most often 5–7.

Egg cases

Egg laying. Egg-laying habits are distinctive in many orthopteran species. Among cockroaches, only one family (Blaberidae) is viviparous; the other four families contain species that have well formed egg cases (oothecae). Among these families, some species carry the oothecae protruding from the body until hatching time is near; others, however, deposit their egg cases within several days of formation. Usually oothecae contain from a few to more than 30 eggs arranged in two rows. Along one edge of the ootheca is a seam that bears a keel, or ridge; the shape of the ridge varies in species that carry the ootheca externally. Minute openings from the base of the keel to the interior of the ootheca are known to be a ventilating device in some species. The ootheca is first carried in the body with the keel uppermost; in certain groups, however, it is rotated prior to deposition so that the keel is on one side. Details concerning positioning of the ootheca and other aspects of egg laying were not correlated with behaviour patterns or group affinities until a basis for understanding their significance was established.

Mantids lay eggs in clusters of less than 10 to more than 300. Usually they are laid in regular layers surrounded by a viscous quick-drying liquid that provides a light but tough protective covering. The egg masses of most mantids have a distinctive shape and size. Although most mantid species attach oothecae to vegetation, these egg cases may be attached to rocks in some environments or placed in grooves in the sand and covered over in the desert.

Grylloblattid eggs are laid in damp moss, decaying logs, or in pockets between broken rocks and wet soil. The eggs, about three millimetres long and usually black in colour, are laid loosely or inserted into the hatching site with the ovipositor.

Of the walking stick species studied, most have eggs that look like small seeds and are dropped loosely on the ground. At least one species, however, attaches its eggs to foliage, and a large, heavily bodied species of the southeastern U.S. (*Anisomorpha buprestoides*) scratches a depression in sandy soil with its front and middle legs, deposits eggs in it, and covers them with sand.

Crickets and katydids utilize their ovipositors to insert eggs into soil or plant material. The eggs of tree crickets (*Oecanthus*), for example, are inserted in rows in the canes of blackberries and various other stems; the eggs of field crickets (*Gryllus*) are laid in soil; the flattened eggs of certain katydids (*Scudderia*) are forced between the upper and lower epidermis layers at the margin of tree leaves; and the eggs of other katydids (*Microcentrum*) are laid in overlapping rows lengthwise on twigs of trees.

Most grasshoppers lay their eggs in soil; a few drill holes in dead wood or place their eggs at the bases of grass clumps or on the surfaces of leaves. Before laying the eggs, the female manipulates the valves of her ovipositor to make a hole in suitable soil (the type varies with the species). During the digging process, the abdomen is greatly extended, and the female manipulates the ovipositor valves to open and close and rotate on the long axis. Then she deposits several dozen eggs in the hole. The eggs are surrounded by a mucilaginous mass (the egg pod) that dries in a cylindrical shape. The number of egg pods laid by one female and the number of eggs in each pod vary according to the species and local conditions. The egg pods are laid over an interval of several weeks.

Parthenogenesis

A few orthopteran species have females only; therefore, reproduction occurs without fertilization (parthenogenesis). Only rarely are species that normally reproduce bisexually parthenogenetic; and when parthenogenesis does occur in bisexual species, it is usually only partially successful because the few nymphs that do hatch often are deformed and fail to reach maturity. In the laboratory, however, there have been a few cases in which several generations have been produced parthenogenetically, proving that there is an inherent capability in these bisexual groups for reproduction without males.

Growth and life span. Increases in the number of antennal segments occur during the development of some orthopterans. The German cockroach, *Blattella germanica,* has been studied in detail in this regard, and it has been shown that a newly hatched nymph has about 24 segments in each antenna. Each of the succeeding nymphal stages (there are usually six in this species) shows an increase in the number of segments until, by the time the adult stage is reached, the average number is 94. The two basal segments do not divide; the third segment, as well as certain other ones, is a growth centre that divides during molting. It is customary for grasshoppers to have 20–30 antennal segments when mature; this is about twice the number present in the first-stage nymph.

The life span of orthopterans depends in part upon whether or not there are long periods during winter or at other times (*e.g.,* a dry season) when the insects are quiescent. Some species habitually spend several months resting during unfavourable periods; such species have one generation per year, and the life span of an individual is approximately one year. The portion of the life span spent as an adult varies, but is likely to be about one or two months. In some species egg maturation is delayed several months after the final molt, and the females do not lay the eggs until they have matured. In other species eggs are not hatched until one or more years after they are laid; therefore, more than one winter or dry season is passed in the egg stage, and a single life cycle can occupy two or more full years. For instance, a walking stick commonly found in the U.S., *Diapheromera femorata,* often has some eggs that hatch the year following deposition and others that hatch after two winters have been passed amid dead leaves on the ground. There are some orthopterans that develop very slowly, and their life cycles require several years for completion; an example is a North American cockroach (*Cryptocercus punctulatus*) that inhabits rotting logs, feeds on decaying wood, and attains maturity after six or seven years. Grylloblattids live for five to seven years. The time required for domestic cockroaches to reach maturity varies with species and environmental conditions. The German cockroach completes nymphal growth in about 95 days, but the American cockroach (*Periplaneta americana*) needs about 225 days. Similarly, adults live from a month or two to several years, depending on conditions and species.

Reproduction. *General features.* Typically each female has paired ovaries consisting of tubes in which eggs develop, moving posteriorly into a single oviduct as they "ripen." The oviduct leads to a vagina and then to the exterior where there is either a simple or specialized ovipositor consisting of paired appendages called ovipositor valves. Attached to the oviduct or vagina is a sac (called the spermatheca) for storage of male sperm; as eggs move down the oviduct, they are fertilized by the sperm. The typical male contains paired testes that produce vast numbers of slender active sperm; these are stored in enlargements of the tubes leading posteriorly from the testes.

Spermatophores

Accessory gland secretions provide not only the medium for carrying the sperm but also a material that solidifies to form a thin-walled sac, or reservoir, containing some sperm and fluid. This reservoir, called the spermatophore, is almost universally found among orthopterans.

Grasshopper spermatophores consist of a bladderlike reservoir and a spermatophore tube. The spermatophore is formed during the first two minutes or so of copulation, after which the tube extends from the male genital organs to the spermathecal duct of the female. The spermathecal duct opens at the base of the ovipositor valves. Sperm pass to the female during copulation; after sperm transfer is complete, parts of the spermatophore may remain attached to both male and female. In some orthopterans, particularly crickets and katydids, the entire spermatophore is attached to the female.

Courtship behaviour. All orthopteran groups have species that show definite courtship behavior prior to actual mating. Male cockroaches are attracted particularly

by females that are virgin and in a receptive condition. Such females frequently secrete pheromones. Pheromones, chemical substances secreted by certain insects, influence the behaviour of other individuals of the same species. Antennae of males of a domestic roach, *Periplaneta americana,* have specialized sense organs that detect the odour of female *P. americana* pheromones; upon detection of the odour, the male initiates searching movements, first with the antennae, then with the palpi. Finally, the male, with folded wings raised and fluttering, actively searches out the female. If the female is still receptive when he finds her, the male protrudes his posterior abdominal segments, pushes under the end of the female, and grasps the terminal ventral segments of the female with his genital hooks; then he expels the spermatophore, which becomes attached to the spermathecal opening of the female. The entire process lasts up to an hour.

Mating calls

Among the grasshoppers, species with coloured hind wings and the habit of making sounds during flight use hovering and other special flight patterns to attract the attention of females. Crickets and katydids have the most dramatic courtship displays because "songs" enter into the precopulatory behaviour. Females of some species are receptive only to the specific song of a male of the same species; in others, however, mating calls are not necessary, and a female will mate with a male who is unable to sing because his wings have been removed. Here, as in grasshoppers, a variety of mating positions are assumed.

A striking sequel to mating occurs frequently in mantids when the female eats the male. There is a popular opinion that mantid males always are eaten, but many escape under natural conditions. But in the close confines of a small cage cannibalism of the male is more common.

Ecology. In a broad sense, ecology represents the sum total of interrelations between organisms and their environment. In the case of orthopterans, the basic requirements of food and moisture; shelter, including protection from weather and from enemies; favourable habitats, involving special niches such as caves or deserts; as well as preferred seasons and conditions conducive to successful reproduction, are involved.

Predators, plant feeders, scavengers

Food. Mantids are the only orthopteran group that feeds almost entirely on insects, but some members of primarily plant-feeding groups also capture and devour insects. For instance, tree crickets (*Oecanthus*) regularly eat flowers, leaves, and other plant parts, in addition to many aphids and other weak insects. Some katydids are active predators of insects. Most cockroaches are typical scavengers, but some are specialized feeders. *Cryptocercus,* for example, digests cellulose in decaying logs with the aid of symbiotic protozoans in its intestines. All walking sticks and a majority of grasshoppers are plant feeders. Although many grasshoppers feed on a wide variety of plant species, some are restricted to a single plant species or one group of plants. Some orthopterans consume only certain parts of plants; for example, coneheaded katydids of the genus *Neoconocephalus* feed mainly on seeds of grasses. Plant preferences among some leaf-eating orthopterans change with the seasons of the year; in other leaf-eating groups feeding habits are dependent on the stage of the life cycle—*i.e.,* nymphs do not eat the same plants as adults of the same species. Moisture requirements of orthopterans vary, as evidenced by the habitats they occupy. Some can absorb water from a drop on the cuticle (skin); others obtain it from water vapour in the air if the relative humidity is sufficiently high.

Habitat. Shelter utilized by orthopterans ranges from general hiding places amid living plant foliage or dead leaves on the ground to special structures such as subterranean galleries in soil or the recesses in caves occupied by various crickets. Some Gryllacrididae are leaf rollers and produce silk to maintain the rolled shape of their hiding places. The loose bark of trees and logs and the water-filled leaf bases of bromeliads often shelter certain genera of cockroaches, some of which are semi-aquatic in their habits. In Africa a few cockroaches (*Cyrtotria*), of elongated and cylindrical body shape, are adapted to enter round holes in hollow plant stems where they sometimes live. With the exception of cockroaches, most conspicuous orthopterans are active by day (diurnal), although there are nocturnal species in every group. Although grylloblattids are essentially nocturnal, they are sometimes active on cloudy days or in winter. The majority of crickets and katydids are nocturnal, as are many walking sticks and some mantids; however, many mantids prey on insects that visit flowers by day.

The degree of moisture, types of vegetation, and altitude above sea level influence the location of orthopteran communities. Grasshoppers breed in the Himalayan Mountains at altitudes as high as 6,000 metres (about 18,000 feet), and each mountain altitudinal zone has distinctive species; fully winged, actively flying species are usually not restricted to a single zone but are found in adjacent ones. On the other hand, at high altitudes there are proportionately greater numbers of grasshoppers with short, nonfunctional wings or none at all.

Caves as habitats

Caves are a special habitat occupied by orthopterans on all continents. The long-horned grasshoppers and the crickets are the principal orthopteran representatives; nearly 200 species of these two groups have been found in caves. In addition, more than 30 cockroach species inhabit caves; and a third group, the grylloblattids, has at least one cavernicolous species in Japan and three in the U.S. Some of these orthopterans inhabit lava tubes and fissures resulting from past volcanic action. Air currents and high humidity in these tubes and fissures produce an "ice cave" condition. In the U.S., the grylloblattids and a few dozen cave crickets (Gryllacrididae) are the principal cavernicoles. It is noteworthy that a bone from a bison skeleton, found in a French cave in the 1920s, bore a prehistoric carving that depicted *Troglophilus,* a European cave cricket.

Usually the orthopteran species found on a given continent are distinct from those of other continents, especially if the land masses are well separated. For example, there are about 2,000 species and 500 genera of grasshoppers in Africa; although several of the genera are found in North America, none of the species is. Some species in North Africa, however, also inhabit southern Europe and western Asia. Explanations for distinct continental species are found not only in the far, overwater distances involved but also in the long periods of isolation that have occurred and in the different conditions that have prevailed in past geologic periods.

Oceanic islands have been populated, in part, by species that were transported by hurricanes, floating debris, birds, and in recent centuries by human activities. The 85 species of Hawaiian orthopterans include four distinct genera of crickets and katydids comprising about 45 of the 85 species. Evidently some of these evolved following the establishment of a few parent species. The remainder are believed to have been established as a result of the activities of man. Since cockroaches are scavengers, they are often found where man is found; *i.e.,* in his buildings and campsites. Early man probably spread cockroaches as he moved about seeking food. Modern commerce has been even more helpful to these unwelcome travellers. An analysis of the distribution of 11 domiciliary species found in the U.S. and related species found elsewhere suggested that five reached America in shipping from West Africa; another African species might not have reached America directly; two probably came from Europe; two might have come from the Orient; and one was native to the West Indies. Thus man has played an important role in spreading cockroaches throughout the world.

Taxonomic significance of ecology

The familiar, large, black field crickets of the U.S. are good examples of ecological differences among similar species. There are six native species in the eastern U.S.; several others occur in the western states, and at least two introduced species have become established in the Gulf States. Using the general appearance of dead specimens, the native eastern species are very similar and difficult to distinguish. For many years, the taxonomy of these species was unsettled. Entomologists using behaviour rather than morphology as a major taxonomic criterion have found that five of the six species have distinct songs; that four of them overwinter as nymphs, the other two chiefly as eggs; and that, to a considerable extent, the habitat preferences

are different—*i.e.,* one species lives in abandoned fields, another in leaf litter of open woodland. Laboratory attempts to crossbreed males and females of different species have been unsuccessful; the pair either failed to copulate or, if they did, produced unhealthy hybrids. Ecological differences also are important for other groups of closely related orthopterans.

FORM AND FUNCTION

Adaptive features. *Movement.* Orthopterans exhibit various adaptations for movement; some are present in an entire family or suborder, others are peculiar to certain genera. The head of mantids is borne by the prothorax in such a way that it is easily turned to face in different directions. Since the mantid diet consists almost entirely of insects, vision is critically important and is unusually well developed. The best known orthopterans with specialized front legs are mantids; the principal leg segments are hinged and spined for seizing and holding prey. Some Orthoptera, especially certain groups of Tettigonioidea, also have front legs with long spines that enable them to hold other insects, although the hinging is not comparable to that in mantids.

Burrowers

Although some cockroaches burrow in soil, sand, or decomposing wood, the principal burrowers are found among the Orthoptera. In both groups the legs, especially the front tibiae, are short and strong, with heavy spurs. Mole crickets, false mole crickets, and sand crickets are accomplished burrowers. Small tunnels serve as shelter and as egg-laying locations, and roots or tubers encountered while burrowing are sometimes used as food. Several genera of camel crickets (Gryllacrididae) in the southwestern U.S. have conspicuous, sometimes basket-shaped, clusters of spurs on the hind legs. They often live on sand dunes and burrow chiefly for shelter. A few desert-living grasshoppers, some in the southwestern U.S., others in Africa, exhibit what has been called "self-burial." Instead of making an elongated cylindrical burrow, the grasshopper rests on the surface of the sand or moves forward and backward, manoeuvring its legs until it has submerged itself and covered its body with sand. The apparent purpose is protection.

Hind legs of Orthoptera, though useful in walking, are used primarily for leaping. Particularly important are the large muscle in the femur, the hinged attachment of tibia to femur, and the tendon extending within the leg from the femur to the end of the tarsus. In a few semi-aquatic Orthoptera, the hind tibia is broadened as a paddle or equipped with fringed spurs to permit effective swimming strokes in water. The ability to run swiftly is common among cockroaches and some mantids. Cockroaches escape enemies by running; mantids utilize their running ability both to escape predators and to catch prey. Some mantids that live on the ground, in deserts, or on tree trunks in the tropics are active runners; however, the majority of mantids stalk their prey slowly or wait quietly until an unsuspecting insect moves nearby.

Body shape is important to many orthopterans, either allowing them to live in places where adequate shelter from weather and enemies is provided or affording them concealment through camouflage. Most cockroaches have flat bodies that enable them to hide beneath stones, under other objects on the ground, or under the loose bark of logs. Examples of orthopterans whose camouflages resemble parts of plants are members of the Phasmida; some of them resemble leaves, others look like twigs or rough pieces of small limbs from trees. Several katydids and grasshoppers resemble leaves; some are green or brown, others have spots that resemble leaves affected by plant diseases. There are some slender grasshoppers that live among grasses, where they conceal themselves by clinging lengthwise to stems and remaining motionless or by quickly sidling around behind stems.

Colour

Camouflage. There are two basic types of insect colours. Structural colours occur when irregular cuticle or scale surfaces break up and reflect certain wave lengths of light. Metallic lustres of some orthopterans (*e.g.,* silvery patches on some grasshoppers) are examples. Most orthopteran colours are due to pigments; often they are located in the cuticle, but sometimes they occur in some deeper body layer. The pigments may be naturally occurring ones or, like melanin, dependent on an oxidation process or a hormonal balance that influences metabolism; these latter pigments are present in varying amounts in different individuals of the same species.

Among some orthopterans, especially grasshoppers, body colours tend to simulate the colour of the habitat background. This is particularly true of species inhabiting rocky or sandy environments. In some cases, colour changes occur rapidly; this was demonstrated by certain light gray African grasshoppers that became black after being caged a few days on dark burnt-over ground. In other cases more time is required. Colour changes usually involve the effect of bright light on integumentary pigments. Among some orthopterans, however, light must enter the eyes, and a rhythm related to some nervous-endocrine mechanism is apparently involved.

An unusual and rapid colour change occurs in an Australian alpine grasshopper (*Kosciuscola tristis*), which lives at above 5,000 feet elevation. The adult male, bright greenish blue on the upper part of its body at temperatures above 25° C (77° F), is dull and blackish below 15° C (59° F). At intermediate temperatures, correspondingly intermediate shades of colour occur. Detailed experiments by Australian entomologists prove that temperature, not light intensity, relative humidity, or degree of crowding is the controlling factor. The epidermal cells of the integument contain brown and blue granules; at warmer temperatures on sunny days the blue granules, in a discrete layer uppermost in the epidermal cells, are near the surface of the integument. At night or on cloudy days, the brown granules migrate from the bottom of the epidermal cells and change places with the blue granules. Thirty minutes is sufficient time for a colour change to take place.

Glandular secretions

Defense. There are no known stinging orthopterans but many have chemical mechanisms in the form of glands that produce irritating fluids or repugnant odours. The disagreeable smell of some cockroaches, especially when disturbed, is well known. Examples are several species of *Eurycotis* in Florida and tropical America; both sexes have a large gland in the hind part of the abdomen between the sixth and seventh segments. An acidic, milky fluid consisting of several chemical constituents is emitted either as an oozing liquid or as a three-foot spray. Another cockroach (*Diploptera*) has a defense gland that ejects a mixture of quinones from the second abdominal spiracles. Ants, beetles, and other predators become confused and avoid these cockroaches when they release their secretions; however, certain mantid predators are not affected.

Man may handle most walking sticks safely, but a large, heavily bodied species in the southeastern U.S. (*Anisomorpha buprestoides*) sometimes forcibly ejects a milky fluid that is extremely irritating if introduced into the human eye. This species has a pair of circular pores on the thorax leading to reservoirs of the fluid; each reservoir has circular muscles that permit ejection of fluid without the general body contraction characteristic of some grasshoppers. When handled, most grasshoppers and some other orthopterans regurgitate from the mouth a brown fluid that superficially resembles molasses. Release of the fluid from the forward part of the alimentary canal is triggered by a response of the nervous system to pressure on certain parts of the body, especially the sides of the thorax or the femurs. Some grasshoppers have other defense mechanisms (*e.g.,* some exude fluid through spiracles or from special glands opening on the body or even leg joints). Sometimes hissing sounds and blowing of bubbles from spiracles accompany secretion.

Physiology and biochemistry. *Body composition.* Several grasshopper species have been analyzed chemically. They consist of (by dry weight) roughly 50 percent to 75 percent crude protein, 4 percent to 18 percent fats, 4 percent to 16 percent carbohydrates, and 3 percent to 19 percent percent ash.

The tough and usually hard outer body wall (exoskeleton) of orthopterans is called the integument or cuticle; its most important component is chitin, a stable polysaccharide chemically similar to plant cellulose. Chitin makes the

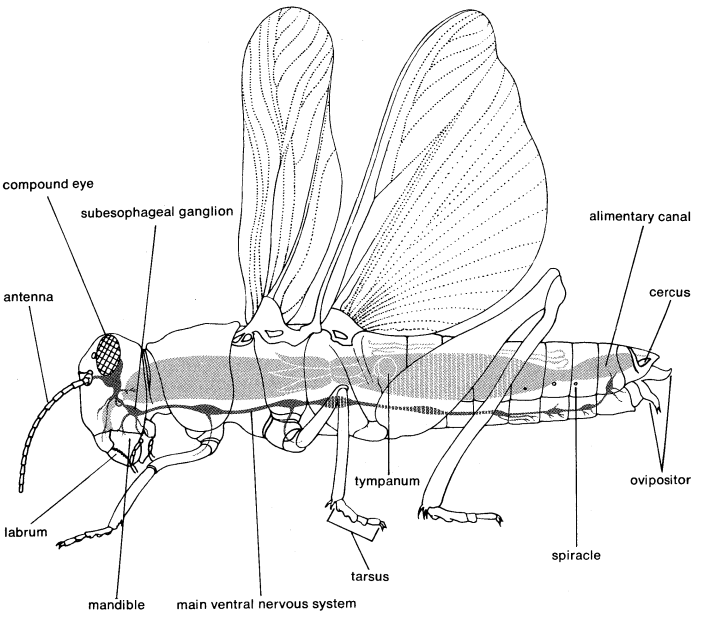


Figure 17: Internal and external body plan of adult orthopteran.
From H. Weber, *Grundriss der Insektenkunde* (1966); Gustav Fischer Verlag, Stuttgart

cuticle strong and flexible but does not provide rigidity. Sclerotin, the horny substance of the cuticle formed by a tanning-like process involving protein produced in the exoskeleton, is found in hard body plates (sclerites), leg spurs, and sharp tubercles; sclerotin is responsible for the rigidity of these structures. A heavily "sclerotized" cuticle is one that is hard and usually dark coloured.

Hormones. The importance of hormones in the biology of orthopterans has been revealed by research. Together with the related pheromones, which tend to coordinate individuals within the population of a species instead of regulating function within an individual, hormones are important in many activities of orthopterans related to mating and reproduction. Other activities involving hormones in grasshoppers include control of fat accumulation in metabolism, control of peristalsis in the malpighian tubules (excretory organs attached to the posterior part of the alimentary canal), secretion of an enzyme at hatching time for dissolving the cuticle that encloses the embryo, and control of the number of molts in nymphal growth.

Activities involving hormones

Detailed studies on the reproduction of cockroaches have disclosed an interrelated series of neurological and glandular functions that combine to control mating and egg production. Frequently, dorsal abdominal glands of the male aid in attracting the female to a mating position. In several cases, once a female has mated and an ootheca is being carried, mechanical pressure of the ootheca causes a stimulation to be transmitted to glandular bodies closely associated with the cerebral ganglia and called corpora allata; this in turn inhibits development of additional eggs in the ovarioles until laying and subsequent removal of pressure occur. In other cases, virgin females are receptive to mating just when yolk deposition is occurring in the first oocytes of developing eggs. Following mating, the mechanical stimulation of the inserted spermatophore inhibits further attraction of the female to the male abdominal glands until after the first group of eggs is deposited.

Phase polymorphism in locusts

Locust is a common name for several species of short-horned grasshoppers that often increase suddenly in numbers and undertake mass migration. Thus, a locust has both solitary and gregarious phases. Gregarious locusts outnumber solitary ones, migrate both as nymphs and adults, and travel in swarms. Swarming adults are tremendously destructive to crops. Typically, gregarious locusts have darker bodies and longer wings compared with solitary forms. Colour changes in adults are correlated with maturation of reproductive organs.

Hormones and pheromones are involved in many stages of locust development. Solitary locusts can transform into gregarious ones as a result of hormonal changes induced by crowding. The presence of mature male locusts under conditions of crowding stimulates a maturation hormone that causes females to mature rapidly. Head glands in the female are stimulated to release another hormone that speeds egg maturation. A favourable season followed by an unfavourable one may cause gregarious locusts to develop. In a favourable season with enough food, the population of solitary locusts increases. If the next season is a poor one, the solitary locusts are forced to crowd together where food is available. Crowding exposes the females to male secretions, females and their eggs respond by maturing rapidly, a population explosion occurs, and a locust horde results. In *Schistocerca gregaria,* the attainment of reproductive activity is sometimes synchronized with environmental contact with certain aromatic shrubs that produce terpenoids in season.

Sound production and hearing. Some orthopterans make conspicuous sounds, while others produce sounds that are outside the range of human hearing. In both cases sound production is important to behaviour necessary for success of the species concerned. Except for Grylloblattodea, in which sound production is unknown, all major groups of orthopterans produce some sort of sound, though sound production is widespread only in crickets, katydids, and grasshoppers.

Stridulation

The stridulatory mechanism of grasshoppers involves moving the hindleg across the folded front wing (tegmen). Serrations, or pegs, which vary in shape, number, and location among different species, are located on the inner surface of the femur and rub across special raised veins of the tegmen, creating a characteristic lisp; sometimes the serrations are on the tegminal veins. In the hindwings of other grasshoppers are stiff veins that make a crackling sound (crepitation) in flight.

Among male crickets and katydids, a front wing with an enlarged transverse vein near its base bears teeth that rasp when shuffled across a scraper on the other front wing. The row of teeth is called the file, and the membrane to which it is attached vibrates when the teeth move over the scraper. During stridulation the tegmina are lifted at an angle of 15° to 40° to the surface of the abdomen, then rapidly opened and closed (shuffled); sound is produced during the closure.

The best known auditory organs of orthopterans, the tympanic organs on each side of the abdomen, are found in both sexes of grasshoppers and on the front tibiae of most crickets and katydids. There are auditory nerves running from special cells beneath a tympanic membrane (a thin area of cuticle, backed by an air sac and free to vibrate) to a ganglion of the central nervous system. In addition to these evident tympanic structures, other less evident auditory organs occur in the orthopterans. Many orthopterans, however, have no conspicuous tympana and are entirely dependent for sound reception on sensory hairs located on cerci, the head, other parts of the body, and an auditory organ called Johnston's organ, which is widespread in the second segment of the antenna.

EVOLUTION AND PALEONTOLOGY

Cockroach fossils

Cockroaches are the most abundant and the earliest fossilized orthopterans found; fossils have been discovered at various localities in North America, France, and the U.S.S.R. Several hundred Carboniferous Period (about 280,000,000 to 345,000,000 years old) and at least a hundred Triassic cockroach species (about 190,000,000 to 225,000,000 years old) have been described; they differ from present Blattaria chiefly in details of wing venation. In addition, some early species had long ovipositors, unknown in recent species. Only a few fossil mantids are known, the oldest in Baltic amber of the Oligocene (about 26,000,000 to 38,000,000 years ago). Some phasmatid-like Jurassic species (about 136,000,000 to 190,000,000 years old) are believed to be primitive walking sticks.

Although fewer in number than the Blattaria, fossil Orthoptera have contributed to orthopteran classification. Ensifera occur from the Triassic to the present; apparently Ensifera and Caelifera separated as distinct evolutionary lines as early as the Carboniferous. The earliest fossils in

the acrididoid line had long antennae. Shorter antennae, reduction of tarsal segments to three, and reduction in length of the ovipositor occurred by the Eocene Period (about 38,000,000 to 54,000,000 years ago).

By the late 19th century, all principal groups of orthopterans except Grylloblattodea were represented in collections; however, the order Orthoptera, broader in scope than it is at the present time, included earwigs and other groups. Gryllacridids were not placed in a separate family, and Phasmida were considered a family closely related to Mantidae because both are walking rather than jumping in habit. By the 20th century, however, basic morphological studies, as well as extensive reports on fossils, contributed new insights into the fundamental relationships of the major groups. The Grylloblattids were first reported in 1914, and numerous publications since then have analyzed their phylogenetic significance. In the late 1930s extensive studies of fossils were correlated with important work on current species, especially concerning the Orthoptera (restricted sense). Meanwhile, comparative studies of wing venation, the proventriculus, reproductive organs, and behaviour have steadily advanced the knowledge of group relationships. Additional details and supporting data, for example, were given in an extensive phylogenetic study in 1968. The rank of a suborder (Acridomorpha) for grasshoppers alone (Eumastacoidea through Acridoidea of this article) has not been evaluated sufficiently. Many of the earliest fossil orthopterans were different enough from any present ones to justify the recognition of separate though extinct families.

CLASSIFICATION

Distinguishing taxonomic features. Among the distinctive features of orthopterans are their wings, which, when present, usually number four. The two forewings, generally long and narrow, are many-veined and somewhat thickened. Among the Orthoptera, Dictyoptera, and Phasmida the forewings, hardened and of a leathery consistency, are known as tegmina. The hindwings, broad with many veins, usually are folded fanlike beneath the forewings when at rest. The females have ovipositors. Some are concealed by ventral abdominal segments; others are as long as the body. Orthopterans have mandibulate mouthparts adapted for chewing and undergo simple metamorphosis.

The classification of orthopterans into orders and families is based chiefly on comparative morphology, on indications of ancestry derived from fossils, and on relationships suggested by patterns of behaviour and the physiology of body systems. Similar anatomical structures that have been significant in deducing orthopteran relationships include: tarsal segments; hindlegs; wings; stridulatory and auditory organs; head capsule; thoracic sclerites; ovipositor; male genitalia; and proventriculus. In recent years correlations between behaviour and mating and reproduction, especially in cockroaches and crickets, have been used widely to support classification of families and subfamilies as well as distinctions between species.

In the grylloblattids, cockroaches, and mantids, the cerci are segmented; segments are lacking in katydids, crickets, grasshoppers, and walking sticks. The external male genitalia are sometimes concealed; in Grylloblattodea (grylloblattids), Dictyoptera (cockroaches and mantids), and Phasmida (walking sticks), the genitalia are asymmetric; in the Orthoptera (katydids, crickets, grasshoppers, and locusts) they are symmetrical. The Orthoptera have the femur of the hindleg enlarged for jumping; other groups have hindlegs similar in size to the middle legs. In mantids the front legs are modified for seizing prey, in mole crickets for digging. The tarsus consists of five segments in grylloblattids, mantids, and cockroaches; there are usually five segments in walking sticks, usually only three or four in Orthoptera. The antennal segments vary from fewer than seven to many long and setaceous ones.

Annotated classification.

ORTHOPTERAN

Common name for several orders of related insects; wings, when present, number 4; chewing mouthparts; mostly plant feeders; size range from 2 mm to 30 cm; more than 24,000 species; worldwide distribution.

Order Dictyoptera

Hindlegs similar to middle ones, adapted for running; tarsi 5-segmented; wing pads not reversed during nymphal stages (front wings remain above hindwings); antennae usually filiform, with more than 30 segments; cerci many-segmented; without auditory tympanum.

Suborder Blattaria (cockroaches)

Head usually concealed from above by shieldlike pronotum; two ocelli represented by pale areas (fenestrae); front legs adapted for running; proventriculus (gizzard) heavily armed on inner lining, with longitudinal folds between the teeth; members from Carboniferous to present; worldwide distribution; sizes 2 to 100 mm, average about 15 mm; more than 3,000 species.

Family Cryptocercidae. Wingless; blackish; cerci (sensory appendages) concealed by apical segments of abdomen; eyes small; size usually at least 15 mm; found in China, Manchuria, U.S.; 3 species.

Family Blattidae. Middle and hindfemurs with numerous strong spines similarly arranged on both ventral margins; ovipositor a plowlike (vavular) modification of apical ventral segment of abdomen; male subgenital plate symmetrical with widely spaced styli; size range about 10 to 35 mm; worldwide distribution; about 550 species.

Family Polyphagidae. Middle and hindfemurs mostly with few strong spines ventrally; spines on front and back ventral margins arranged differently; anal area of folded hindwing not plaited fanlike; postclypeus usually enlarged; range in size about 3 to 35 mm; worldwide distribution; about 200 species.

Family Blattellidae. Variable, mostly small species; mostly with numerous similar spines on ventral margins of middle and hindfemurs; ranging in size from 3 to 100 mm, with an average of about 12 mm; worldwide distribution; about 1,700 species.

Family Blaberidae. Mostly large species; spines of middle and hindfemurs variable; male subgenital plate asymmetrical; viviparous or ovoviparous; from 10 to 60 mm in size; worldwide in distribution; about 650 species.

Suborder Mantodea (mantids)

Head usually conspicuous anterior to a narrow pronotum, seldom concealed by a broad pronotum; 3 distinct ocelli; front legs adapted for seizing prey; proventriculus not heavily armed, inner lining with fine anastomosing ridges between teeth of moderate size.

Family Mantidae. Oligocene to present (fossils rare); size from 20 to 155 mm, average about 60 mm, worldwide distribution; about 2,000 species.

Order Grylloblattodea (Notoptera)

Legs similar, adapted for running; 5-segmented tarsi; wingless; eyes small or absent; no ocelli; antennae fairly long, filiform, about 25–40 segments; female with well-developed ovipositor, resembling Tettigoniidae (see below under *Orthoptera*); cerci long, with 8 to 9 segments.

Family Grylloblattidae (ice bugs). Recent (no fossils); size about 20–25 mm; found in northwestern North America, eastern Siberia, Japan; about 15 species.

Order Phasmida (Cheleutoptera or Phasmatoptera)

Legs similar, adapted for walking; tarsi nearly always 5-segmented; often wingless; when winged, tegmina often shorter than wings; wing pads not reversed as nymphs; ovipositor short, often concealed; male genitalia asymmetrical; cerci short, unsegmented; no tympanum or conspicuous stridulatory organs; usually very slender, elongated, sometimes broadened, even leaflike.

Family Phasmatidae (walking sticks, stick insects). Variable, rarely less than an inch long; a broadly conceived family; Triassic to present; 15 to 325 mm in size with average about 70 mm; worldwide distribution centred in warm countries; about 2,000 species.

Family Phylliidae (leaf insects). Depressed, leaflike, winged; tegmina short in males, covering most of abdomen in females; antennae long, pubescent in males, very short in females; all members recent; about 60 mm in size; Indo-Malayan distribution from Ceylon to Fiji; about 25 species.

Order Orthoptera

Hindlegs almost always enlarged and adapted for jumping; tarsi usually 3- or 4-segmented, occasionally fewer than 3 or as many as 5 segments; front wings more or less thickened; wing pads reversed (hindwing partly covering front wing) during late nymphal stages; cerci nearly always unsegmented; specialized auditory and stridulatory organs often present.

Suborder Ensifera

Antennae usually long, with more than 30 segments; auditory organs, if present, consist of tympanum at base of front tibia; many species with stridulatory mechanism at base of tegmen; ovipositor usually present as rounded or flattened bladelike, or elongated cylindrical spearlike, structure.

Superfamily Prophalangopsoidea. Tarsi compressed or cylindrical, 3- or 4-segmented; tympanum present; simple stridulatory vein on each male tegmen.

Family Prophalangopsidae. Triassic to present. 25–47 mm in size; recent species found in India and northwestern North America; 3 species.

Superfamily Tettigonioidea. Tarsi usually depressed, 4-segmented, at least on middle and hindlegs; tympanum present; stridulatory organ specialized, with left tegmen uppermost and scraper (at edge of lower tegmen) often better developed on right tegmen.

Family Tettigoniidae (katydids, bush crickets). Jurassic to present; size range from 15 to 120 mm, average about 35 mm; distribution worldwide; about 3,000 species.

Family Phasmodidae (phasmodids). Very slender, resembling Phasmatidae (see above); head prognathous (*i.e.,* mouth directed forward instead of downward); hindfemurs not enlarged; all members recent; found in Australia; size range from 25 to 85 mm; 2 species.

Superfamily Gryllacridoidea. Tarsi depressed, usually 4-segmented; front wings without stridulatory mechanism; usually no tympanum; ovipositor usually flattened, bladelike; 12 to 50 mm in size with average 25 mm.

Family Stenopelmatidae (stenopelmatids, Jerusalem crickets). Heavy body, often wingless; front legs usually strongly armed for digging; ovipositor inconspicuous; members from Miocene to present; found in southeastern Asia, South Africa, North America, and Central America; about 35 species.

Family Gryllacrididae (gryllacridids). Usually winged with folded wings surrounding abdomen at rest; found from Jurassic to present; chiefly tropical, most nocturnal, in trees and shrubs; one species in U.S.; about 550 species.

Family Schizodactylidae (schizodactylids). Winged or wingless; when winged, wing tips very long, rolled in a coil at rest; tarsal segments 2 and 3 with conspicuous lateral lobes; all members recent, found in southern Asia, South Africa; about 8 species.

Family Rhaphidophoridae (camel crickets and allies). Wingless; usually of humpback appearance; tarsi compressed; legs often very long and slender; found from Oligocene to present; worldwide distribution; about 300 species.

Family Lezinidae. Wingless; cerci with apical segmentation; found in Asia and Africa; about 10 species.

Family Henicidae. Usually wingless; front tibia with 1 or more dorsal spurs; front coxa usually with spine; represented in Old and New World tropics; about 150 species.

Superfamily Grylloidea. Tarsi 3-segmented; tegmina of males usually highly developed for stridulation; tympanum usually at base of front tibia; ovipositor most often round in cross section, spearlike, other times somewhat flattened or reduced; 2 to 50 mm in size with average 18 mm.

Family Gryllidae (crickets). Front legs usually not for digging; ovipositor usually elongated; species from Triassic to present; worldwide distribution; about 2,400 species.

Family Gryllotalpidae (mole crickets). Front legs highly modified for digging, with tibiae expanded and with fingerlike dactyls; ovipositor vestigial; represented from Oligocene to present; distribution worldwide; about 65 species.

Suborder Caelifera

Antennae short, with less than 30 segments; auditory organs, if present, at base of abdomen in form of tympanum; ovipositor usually consists of paired valvular appendages adapted for digging.

Superfamily Tridactyloidea. Tarsi 1- or 2-segmented; antennae short, with 12 or fewer segments.

Family Tridactylidae (pygmy sand crickets). Front tarsi 2-segmented; hindtibiae with unsegmented tarsus, no "claws," 2 long "paddles"; usually winged; 3 to 18 mm in size; worldwide distribution; about 60 species.

Family Cylindrachaetidae (false mole crickets). Body elongated, cylindrical, wingless; antennae 7- to 8-segmented; eyes reduced; ocelli absent; front tibia with 4 or 5 dactyls; cerci very short. All members recent; inhabit Australia and southern South America; size 40–60 mm; about 8 species.

Superfamily Tetrigoidea. Pronotum strongly elongated, extending backward over most or all of abdomen; tegmina short, scalelike; hindwings usually fully developed; no tympanum; front and middle tarsi 2-segmented; hindtarsus 3-segmented.

Family Tetrigidae (pygmy grasshoppers or grouse locusts). Miocene to present; 8–25 mm in size with average about 14 mm; worldwide distribution; about 700 species.

Superfamily Eumastacoidea. Abdominal spiracles situated in lateral membrane between terga and sternums; basal abdominal terga without auditory tympanum; basal segment of hindtarsus with serrated margins or at least with an external tubercle (all U.S. species wingless).

Family Eumastiacidae (eumastacids or monkey grasshoppers). Sides of abdomen without stridulatory organ; antennae relatively short; frontal costa forked below median ocellus; represented from Miocene to present; worldwide in warm countries; about 200 species.

Family Tanaoceridae (tanaocerids). Specialized stridulatory organ on side of 3rd abdominal tergum, rubbed by hindfemur; antennae very long; frontal costa not forked; found in southwestern U.S.; 3 species.

Superfamily Pneumoroidea. Body robust; male fully winged, female short winged; hindlegs short; tympanum absent; male with stridulatory ridges on side of 3rd abdominal tergum.

Family Pneumoridae (pneumorids). All members recent; restricted to southern Africa; 11 mm to 107 mm in size; about 18 species.

Superfamily Proscopioidea. Body sticklike, superficially like Phasmatidae (see above under *order Phasmida*); head elongate, conical, usually oblique or vertical; wings rarely present, tympanum absent; no stridulatory organ.

Family Proscopiidae (proscopiids). All representatives recent; from Costa Rica to southern South America; size 30 to 180 mm with average about 60 mm; about 100 species.

Superfamily Acridoidea. Typically robust, with both pairs of wings and tympanum present, but very variable; pronotum not or briefly prolonged over abdomen; abdominal spiracles located in terga; tarsi 3-segmented, basal segment without serrated margins; size 8 to 120 mm with average about 28 mm.

Family Acrididae (grasshoppers). Many of unusual appearance; stridulatory mechanism often present, but not with organ on abdomen; vertex without a definite longitudinal furrow; tympanum normally present (absent in some, especially wingless genera); ocelli small; hindtibiae usually not expanded apically for swimming; Eocene to present; worldwide distribution; more than 5,000 species.

Family Pamphagidae (pamphagids). Usually large; wings variable; many wingless bulky species in arid regions; vertex of head with longitudinal furrow; lower basal lobe of hindfemur longer than upper lobe; members recent; in Africa and southern parts of Europe and Asia; about 300 species.

Family Pyrgomorphidae (pyrgomorphids). Head conical; longitudinal furrow of vertex present; lower basal lobe of femur longer than upper lobe; tympanum usually present; no stridulatory mechanism; recent; worldwide in tropical and subtropical areas, unknown in U.S.; about 350 species.

Each of the following 7 families contains only a few species: Pauliniidae, Xyronotidae, Ommexechidae, Trigonopterygidae, Charilaidae, Lathiceridae, Lentulidae. The Xyronotidae includes 1 Mexican species only; the other families are not represented in North America.

Critical appraisal. The arrangement of orders, suborders, superfamilies, and families presented above is a consensus of current opinion; however, some entomologists recognize different relationships and additional families too minor or too little understood to be included here. In most large groups, there are a few peculiar species that vary slightly from the characteristics described. Although an attempt was made in this classification to indicate the earliest geological period in which major groups were found, available information is not always clear as to whether the insect found belongs to a modern family or an ancestral one with more primitive features and a different name. The known species shown for each group represent only living ones that have been named; it does not include fossils. In some cases modern catalogs and monographs can be used to obtain accurate counts; for others, estimates are given.

Several other orders of insects are orthopteroid in their general relationships. Among them, the Dermaptera (earwigs) differ from orthopterans by having short leathery front wings devoid of veins, hindwings with veins radiating from a central point midway of the anterior margin, and most cerci modified into pincer-like structures. Isoptera (termites) have similar front and hindwings; although in many ways they resemble some cockroaches, they differ in their elaborate caste system and their habits (*i.e.,* living in complex colonies consisting of reproductive individuals, sterile workers, and soldiers). Zorapterans show some morphological relationship to cockroaches but have two-segmented tarsi, peculiar wing venation, a primitive caste system, and other differences. Embiopterans (web spin-

ners) are also orthopteroid in basic morphology, but are notably distinct from orthopterans by the much enlarged silk-producing basal segment of the front tarsus. Plecoptera (stoneflies) are also orthopteroid, but their front and hindwings are of a similar texture (unlike orthopterans), and their immature stages are specialized for an aquatic life.

(A.B.G.)

Isoptera (termites)

Termites are cellulose-eating social insects that constitute the order Isoptera. Although they are referred to popularly as white ants, they are not closely related to ants, which are grouped with bees and wasps in a higher order of insects, the Hymenoptera. The social system of termites shows remarkable parallels with those of the Hymenoptera, but it has evolved independently.

GENERAL FEATURES

Distribution and abundance. Termites, which number in total about 1,900 species, are distributed widely, reaching their greatest abundance in numbers and species in tropical rain forests around the world. In North America termites are found as far north as Vancouver, British Columbia (*Zootermopsis*), on the Pacific coast, and Maine and eastern Canada (*Reticulitermes*) on the Atlantic coast. In Europe the northern limit of natural distribution is reached by *Reticulitermes lucifugus* on the Atlantic coast of France, although an introduced species, *Reticulitermes flavipes,* occurs as far north as Hamburg, Germany. The known European species of termites have a predominantly Mediterranean distribution and do not occur naturally in Great Britain, Scandinavia, Switzerland, Germany, or northern Russia. In the Far East *Reticulitermes speratus* ranges as far north as South Korea, Peking, and northern Japan. Termites occur also in the Cape Province of the Republic of South Africa, Australia, Tasmania, and New Zealand.

In addition to naturally occurring termites, many species have been transported inadvertently by man from their native habitats to new parts of the world. Termites, particularly *Cryptotermes* and *Coptotermes,* are transported in wooden articles such as shipping crates, boat timbers, lumber, and furniture. Because dry-wood termites (*e.g., Cryptotermes* species) live in small colonies in wood and tolerate long periods of dryness, they can survive in seasoned wood and furniture and can easily be transported over long distances. Members of the family Rhinotermitidae (*e.g., Coptotermes*) are transported in shipping crates that have contact with moisture. *Coptotermes formosanus,* widely distributed in Japan, Taiwan, and South China, has been introduced into Sri Lanka (Ceylon), the Pacific islands, South Africa, East Africa, Hawaii, and southern United States. A termite native to the U.S., *Reticulitermes flavipes,* was found in the hothouses of the Royal Palace in Schönbrunn, in Vienna, and the species was reported and described before it was discovered in the U.S. The termites presumably had been shipped from North America in wooden containers of decorative potted plants.

Importance. Termites are important in two ways. They are destructive when they feed upon, and often destroy, wooden structures or vegetable matter valuable to man. Introduced species, because they are not so well equipped as native species to adapt to changes in their new environments, tend to seek shelter in protected, man-made environments such as cultivated crops or buildings and thus are likely to become the most serious pests, causing serious damage to houses and wooden furnishings. Termites are beneficial, however, in that they help to convert plant cellulose into substances that can be recycled into the ecosystem to support new growth.

Termite control

Although of the nearly 2,000 known termite species only about 10 percent have been reported as pests, many of that group cause severe and extremely costly damage. For effective control, it is essential to determine whether the pest is a subterranean or a wood-dwelling species, as treatment methods differ.

Subterranean termites, dependent on contact with soil moisture, normally reach the wood in man-made structures through their foundations. The most common control used around a structure is a trench containing insecticide and covered by soil. Insecticides are useful around cracks and crevices in foundations and infested wood. Construction and design practices that can prevent the initial entry of subterranean termites into a structure include the use of pressure-treated wood, dieldrin-treated concrete foundation blocks, and reinforced concrete foundations that extend at least six inches (15.2 centimetres) above the ground and have no cracks or contact with any outside wood.

Dry-wood termites, which nest in the wood on which they feed and do not invade a structure from the soil, are difficult to control. Preventive measures include the use of chemically treated wood in building construction and the use of paint or other durable finish to seal cracks in wood surfaces. Fumigation is the most effective method for exterminating dry-wood termites. Another method is to pour insecticides into small holes drilled into galleries of infested wood.

No completely satisfactory method of termite control has been devised. Since most methods have depended heavily on chlorinated hydrocarbons, it is probable that alternate, safer methods of control will be developed.

From C.A. Kofoid *et al., Termites and Termite Control* (1934); originally published by the University of California Press; reprinted by permission of The Regents of the University of California

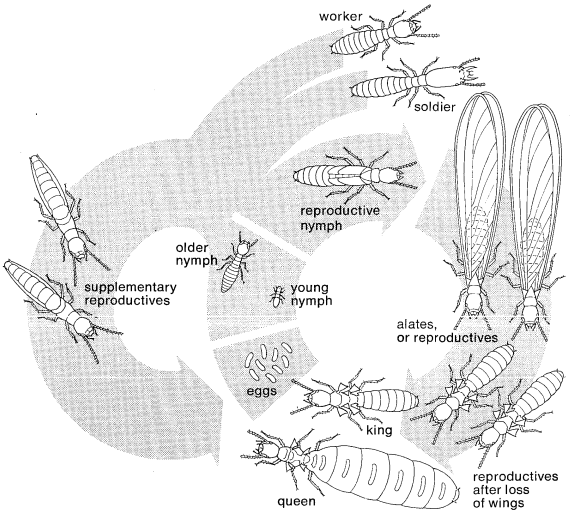


Figure 18: Life cycle of the termite.

NATURAL HISTORY

Colony organization. The termite society, or colony, is a highly organized and integrated unit, with division of labour among its members differentiated by structure, function, and behaviour into castes. The major castes in the colony are the reproductive, soldier, and worker castes; the soldiers and workers are sterile. The functional reproductives are of two types, primary and secondary (or supplementary).

Mechanisms controlling differentiation of termites into castes are not understood fully. It is known that all young nymphs are genetically identical at hatching and that all could develop into any caste (reproductive, soldier, or worker).

Regulation of castes

The castes in a colony are balanced and regulated closely; normally there are one pair of reproductives and a set ratio of soldiers to workers and nymphs. If members of any caste are lost, additional members of that caste develop from nymphs to restore the balance. Conversely, if overproduction of one caste occurs, selective cannibalism restores the balance.

Chemical substances such as pheromones and hormones play a role in differentiation, production, and regulation of castes. Both reproductive and soldier castes secrete a pheromone, a chemical substance that is transmitted through mutual licking (trophallaxis) to other members of the colony and inhibits development of reproductives or soldiers. If the caste balance of the colony is upset, some undifferentiated nymphs, which do not receive the "pher-

omone message," develop into reproductives or soldiers, thereby restoring the balance.

This inhibition theory has been confirmed by experiments with supplementary reproductive development in *Kalotermes* and *Zootermopsis.*

Pheromones may act to control caste differentiation through hormonal action, but it is not clear how this mechanism works. Activation of the corpora allata near the brain may result in release of a hormone, distinct from the juvenile hormone, that causes differentiation of a nymph into a soldier. Similarly, activity of a molt gland may be responsible for differentiation of the reproductive caste. In termites, therefore, hormones not only control molting and metamorphosis, as in other insects, but also play a role in caste differentiation.

Colony formation and development. *Swarming.* A new termite colony normally is founded by dispersion of winged adults (alates), which usually develop in a mature colony during certain seasons of the year. After molting into winged adults, alates group themselves in special chambers near the periphery of the nest for several days or weeks. Emergence and flight of alates usually is associated with high atmospheric humidity in combination with temperature, climatic, and seasonal factors that vary with the species. In some species one emergence a year may occur; in others there may be many successive flights.

Workers prepare tunnels to the surface and exit holes prior to emergence of the alates and sometimes construct launching platforms. During emergence the soldiers guard the exit holes, not only to prevent entry of enemies but also to prevent alates from re-entering the nest. At the time of emergence the alates, which normally avoid light, become attracted to it and fly out of the nest. They are weak fliers and, unless carried by the wind, descend within several hundred yards of the original colony. The flight, commonly called a nuptial or mating flight, is simply dispersal; mating occurs after the flight. Swarming from many colonies occurs simultaneously in a given area and may be synchronized closely in areas separated by hundreds of miles. An advantage of synchronization might be intercolony mating.

Dispersal, pairing, and mating

Shortly after the alates alight, they shed their wings, leaving only the base of the wing scale attached to the thorax. During a short courtship, in which the female raises her abdomen and emits a sex attractant, the pair moves off in tandem (pairing), with the male following closely behind the female. The couple then seek a nesting site; together they find a crevice or dig a hole in wood or soil that has been softened by rain and seal the hole with fecal matter. Copulation takes place only after the establishment of this nuptial chamber. During copulation, which occurs intermittently throughout the lives of the king and queen, sperm are transferred and stored in the spermatheca of the female. Since the male has no external copulatory organ, sperm are released through a median pore on the ninth sternite, or abdominal plate.

After copulation the first batch of eggs, usually few in number, is laid. In two to five years, as the colony matures, the egg-laying capacity of the queen increases, her ovaries and fat bodies develop, and her abdomen enlarges (the process is called physogastry). Physogastric queens in more advanced families (*e.g.,* family Termitidae, especially *Macrotermes* and *Odontotermes*) may become 11 centimetres long. The queen, now an "egg-laying machine," may produce as many as 36,000 eggs a day for many years. The king is one or two centimetres long. In temperate regions egg laying stops during the winter months, while in tropical species it continues throughout the year.

The first young nymphs develop into workers or pseudergates and soldiers. Only after the colonies are mature do winged adults develop. During the initial stages of colony formation, the reproductives feed the young and tend the nest; but, as the colony becomes established, the young nymphs perform these duties.

Primitive termite families have small colonies—from hundreds to thousands of individuals. More advanced families (*e.g., Rhinotermitidae, Termitidae*) have colonies that may number thousands to millions of individuals; all members are produced by the single reproductive pair. Workers and soldiers may live two to five years. The primary king and queen in higher termite families may live 60 to 70 years. The entire colony may exist for many years in species that replace the primary king and queen with secondary reproductives.

Other colonizing methods. Sometimes new colonies are formed by budding, the division or accidental separation of part of a colony from an original nest; supplementary reproductives then take over as the reproductive pair. Another method of colony formation is sociotomy, or social fragmentation; workers, soldiers, and nymphs migrate or march to a new nesting site, and the fragment develops supplementary reproductives. Sometimes an original reproductive pair joins a migrating group.

Nests. *Internal features.* Since termites have a soft cu-

(Left) P.P. Grasse and C. Noirot, *Insectes Sociaux,* vol. 7 (1961), (centre) F.J. Gay and J.H. Calaby, *Biology of Termites,* vol. 2 (1970), (right) Kumar Krishna



Termite nests.
(Left) Cross section of mound of the fungus-growing termite *Macrotermes natalensis* from the Ivory Coast of Africa. (Centre) "Compass" mound nest of *Amitermes meridionalis,* north–south aspect, from Australia. (Right) Arboreal carton nest of *Nasutitermes lacustris* from Sri Lanka (Ceylon).

ticle and are desiccated easily, they live in nests that are warm, damp, dark, and sealed from the outside environment; their nests are constructed by workers or old nymphs. In addition to providing an optimum microclimate, the nest provides shelter and protection against predators. The high relative humidity in the interior of the nest (90 to 99 percent) probably is maintained in part by water production resulting from metabolic processes of individual termites. The temperature inside the nest generally is higher than that of the outside environment.

Ventilation of nests

Since the anaerobic protozoans (necessary to digestion) in the hindguts of primitive termites cannot tolerate high concentrations of oxygen, such termites have developed a toleration for high concentrations of carbon dioxide, as high as 3 percent in some nests. Ventilation must occur in the nest, however, and may be facilitated by its architecture. For example, the subterranean nests of *Apicotermes* have an elaborate system of ventilation pores. Convection currents and diffusion through the nest wall also provide ventilation in large mounds.

Nest types. The family Kalotermitidae and the subfamily Termopsinae (family Hodotermitidae) make their nests in the wood on which they feed. These termites excavate irregular networks of galleries with no external openings (except for temporary ones during swarming); many galleries have partitions made of fecal matter and are lined or coated with plaster made of fecal matter. The Kalotermitidae live in the sound wood of stumps and branches of trees. Examples are *Neotermes tectonae,* which lives in and attacks teak trees in Java, and *Cryptotermes,* which bores into trees and furniture in various parts of the world. The Termopsinae live in damp rotten logs. Although true wood dwellers never invade soil, and their nests have no soil connections, all other termites basically are subterranean; *i.e.,* they build their nests either in soil or with soil connections and exploit food sources away from the nest.

Many species of Rhinotermitidae build nests in wood that is buried in damp soil and from which a diffused network of tunnels to food sources may radiate into the soil or above the ground in the form of covered runways. Other termites build a diffused subterranean nest with many chambers or pockets in soil and a network of galleries.

Many termites build discrete and concentrated nests. Some nests rise partly above the ground as mounds or hills; others are totally underground or arboreal. Dirt, particles of fine clay, or chewed wood glued together with saliva or excreta are used to build nests. During nest construction a termite deposits fecal matter to cement particles in place.

Arboreal nests are ovoid structures built of "carton" (a mixture of fecal matter and wood fragments), which resembles cardboard or papier-mâché. Carton may be papery and fragile, or woody and very hard. The inside of an arboreal nest consists of horizontal layers of cells, the queen occupying a special compartment near the centre. The nests always maintain connections with the ground through covered runways.

The large mounds or hills, a prominent landscape feature in the tropics, may be domelike or conical; some have chimneys and pinnacles. Longitudinal and horizontal chambers and galleries comprise the interior. Generally the outer wall is constructed of hard soil material, distinct from the internal central portion (or nursery), which is composed of softer carton material. In northern Australia *Amitermes meridionalis* builds wedge-shaped mounds, called compass or magnetic mounds, that are three to four metres (9.8 to 13.1 feet) high, 2.5 metres (8.1 feet) wide, and one metre (3.2 feet) thick at the base. The long axis always is directed north–south, and the broad side faces east–west—an orientation that probably functions to regulate temperature. Spectacular mounds are built by fungus-growing termites in Indomalaya and Africa; mounds of some African *Macrotermes* species reach a height of eight to nine metres (26.2 to 29.5 feet) and have pinnacles, chimneys, and ridges on their outer walls. Such mounds are built of fine particles of clay glued together by saliva to form an exceedingly hard substance. Inside the mounds are honeycomb-like structures on which the fungus grows.

Many termite nests harbour various other invertebrates as guests (*e.g.,* beetles, flies, bugs, caterpillars, millipedes); some (termed termitophiles), in fact, are unable to survive independent of their termite hosts. True termitophiles actually have evolved with their hosts and are species specific. Some beetles and flies have developed glands that secrete substances sought and licked by the termites. The termite nest, because it provides shelter and warmth, may be occupied also by lizards, snakes, scorpions, and some birds.

Invertebrates dependent on termites

A few termites, known as inquilinous species, live only in obligatory association with other termite species. Examples of such obligate relationships are *Ahamitermes* and *Incolitermes* species, which live only in the mound nests of certain *Coptotermes* species; the galleries of guests and hosts are completely separate. Inquilinous species feed on the inner carton material of the host nests. *Incolitermes,* however, depend on the host species not only for food but also for exit holes from the nest during swarming when alates of the inquiline and the host emerge together. Such species' tolerance is highly unusual; normally, different species of termites are hostile to one another, and host termites may attack inquilinous guests if partitions between galleries are broken.

FORM AND FUNCTION

Castes and their roles. *Reproductives.* The primary reproductives in a termite colony are usually one royal pair, a king and queen. They have developed from winged forms (alates) with hardened, pigmented bodies and large compound eyes that have flown from a parent colony and shed their wings. The primary reproductives have several important functions: reproduction, dispersal, and colony formation; in addition, during initial stages of colony

From (winged adult) A *Revision of the Nearctic Termites* by Nathan Banks and Thomas E. Snyder, U.S. National Museum Bulletin 108, Smithsonian Institution, in D.J. Borror and D.M. DeLong, *An Introduction to the Study of Insects,* Holt, Rinehart and Winston, Inc.; and (others) S.H. Skaife, *Dwellers in Darkness* (1961), Longman Group Ltd.

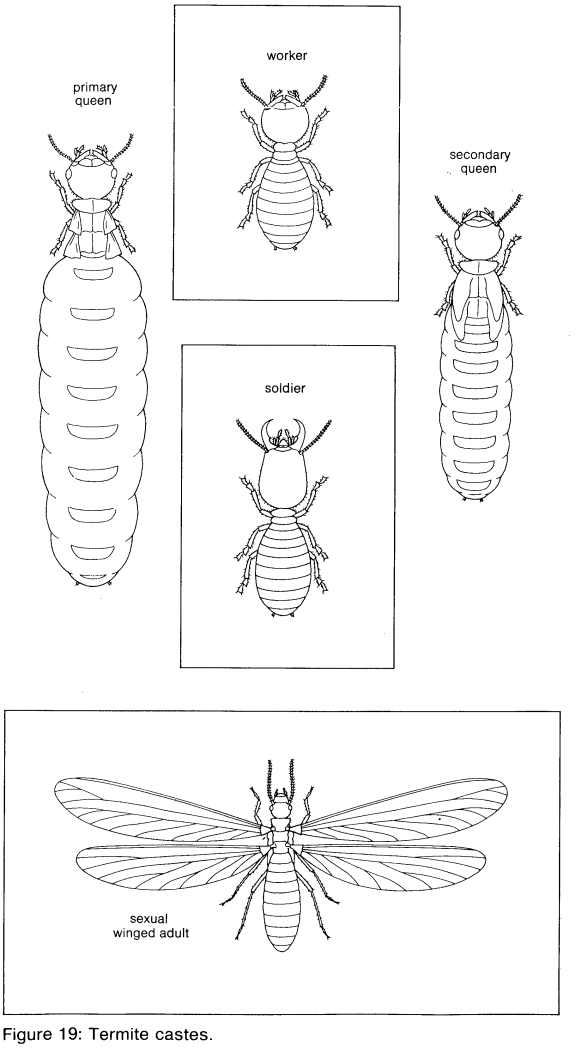


Figure 19: Termite castes.

formation, the primary reproductives perform tasks later taken over by the worker caste (*e.g.,* construction, housekeeping, care of young).

If the king and queen die, they are replaced by several supplementary reproductives that are slightly pigmented, have either short wing pads (brachypterous) or none (apterous) and reduced compound eyes. These secondary reproductives, which develop from nymphs and may be called neotenics, normally are not present in a colony as long as the primary reproductives function. If a primary reproductive is lost, a neotenic achieves sexual maturity without, however, attaining a fully winged adult stage or leaving the nest.

Workers and soldiers. The sterile castes are the workers and soldiers. Both are wingless, usually lack eyes and, though of both sexes, usually lack fully developed reproductive organs. In some species the workers and soldiers are dimorphic (of two sizes); the larger is termed a major soldier or worker, the smaller a minor soldier or worker. A few species contain trimorphic soldiers. Most termite species have both soldier and worker castes.

The worker caste usually is the most numerous in a colony. Workers are pale in colour, soft-bodied, and with mandibles and mouthparts adapted for chewing. They feed all the other members of the colony (reproductives, soldiers, and young), collect food, groom other colony members, and construct and repair the nest. The worker caste is responsible for the widespread destruction termites can cause. In some primitive termite families a true worker caste is absent, and its functions are carried out by immature individuals called pseudo-workers or pseudergates, which may molt without much change in size.

Mechanisms for defense

The primary function of the soldier caste is defense. Since most termite soldiers are blind, they probably locate enemies through tactile and chemical means. The termite soldier has a large, dark, hard-covered head; its long powerful jaws (mandibles) may be hooked and contain teeth. The head and the mandibles are used to defend the colony against predators, usually ants. The attacking mandibulate soldier makes rapid lunging movements, opening and closing its mandibles in a scissorlike action that can behead, dismember, lacerate, or grip a foe. In some mandibulate soldiers (*e.g., Capritermes*) the mandibles are an asymmetrical, snapping type, with the left mandible twisted and arched and the right bladelike. In defense, the mandibles lock together and release with a loud click, like the snapping of fingers. Some soldiers (*e.g., Cryptotermes*) use their heads, short and truncated in front (phragmotic), to plug the entrance holes of nests.

The higher termites (Termitidae) may supplement or replace mandibular defenses with chemical mechanisms that utilize sticky, possibly toxic, liquids secreted by either the salivary or the frontal glands. The whitish or brownish liquid becomes rubberlike after exposure to air and entangles enemies. The frontal gland of some termites (*e.g., Coptotermes* and *Rhinotermes*) occupies a large portion of the abdominal cavity and opens by means of a frontal pore (fontanelle), through which the liquid is ejected. The liquid from the frontal pore of the minor soldier of *Rhinotermes* flows down a groove in the elongated labrum, rests at its hairy tip, and volatilizes as a repellent gas.

The mandibles of soldiers with exclusively chemical defense (Nasutitermitinae) have become reduced in size and nonfunctional; the head has become elongated into a long snout (nasus), and the frontal gland, which occupies a major portion of the head, opens at the end of the snout. These nasute soldiers can fire a clear, sticky, resinous liquid accurately for many centimetres. A few genera lack a soldier caste; the mechanisms for defense in these groups are not known.

Nutrition. *Cellulose.* The food of termites is mainly cellulose, obtained from wood, grass, leaves, humus, manure of herbivorous animals, and materials of vegetative origin (*e.g.,* paper, cardboard, cotton). Most lower termites and many higher ones feed on wood, either sound or partly decayed. A few termites (known as foragers or harvesters) collect and eat grass, leaves, and straw. Many higher termites (family Termitidae) are humivores, or exclusively humus feeders.

As with other social insects, not all members of a termite colony feed directly. Because reproductives, soldiers, and young nymphs in lower families (all nymphs in Termitidae) cannot feed themselves directly, they must be fed by workers. Workers (or in families without them, the older nymphs) feed for the entire colony and transfer food to dependent castes either by mouth feeding or by anal feeding. Food transferred by mouth may consist of either pastelike regurgitated chewed wood and saliva or a clear liquid. This method is used in all termite families. During anal feeding, present only among lower termites, a pastelike liquid or droplet is discharged from the anus of the worker and licked away by the dependent castes. This liquid food, distinct from feces, consists of hindgut fluid containing protozoans, products of digestion, and wood fragments.

Importance of symbiotic protozoans

Cellulose digestion in lower termite families depends upon symbiotic flagellate protozoa, which live anaerobically (without oxygen) in the termite hindgut and secrete enzymes (cellulase and cellobiase) that break down cellulose into a simple sugar (glucose) and acetic acid. The termites depend entirely on protozoans for cellulose digestion and would starve without them. Newly hatched nymphs acquire protozoa from older, infected termites during anal feeding—a type of feeding necessary to lower termites that harbour protozoans.

Since the protozoans lost at the time of each molt are reacquired only through anal feeding, termites live in groups that allow contact of molting nymphs with infected, nonmolting individuals. It is possible that the necessity for transfer of protozoans was responsible for the evolution of the termite society.

Higher termites lack symbiotic protozoans; only bacteria are present in the gut. Digestion may occur with the aid of bacterial cellulase and cellobiase enzymes, but the termites themselves may secrete the enzymes.

In addition to cellulose, termites require vitamins and nitrogenous foods (*e.g.,* proteins), which probably are supplied by fungi normally present in the decayed wood diet common to most termites. The fungi also may break down wood into components that are digested easily by termites.

Fungus gardens. The Macrotermitinae (family Termitidae) cultivate symbiotic fungi (*Termitomyces*). The termites construct spongelike "fungus gardens," or combs, possibly of fecal matter rich in the carbohydrate lignin. The fungi grow on the combs, and the termites consume both fungi and combs. The fungi break down the fecal matter used to construct the combs into substances that can be reutilized by the termites. Nitrogen other than that from fungi is supplied by controlled cannibalism. The termites consume cast-off skins and dead, injured, and excess members of the colony.

Communication. Among the members of a termite colony there is continuous exchange of information, such as alarm, indication of direction and presence of a food source, and, among reproductives, calling and pairing behaviour. Information is communicated by visual cues, vibrations, physical contact, and chemical signals (*e.g.,* odour).

Many termite species leave their nests to forage for food. Workers (or older nymphs) and soldiers march in columns along the ground and carry grass, pine needles, and seeds for storage in the nest. The foraging trail between the nest and the food source may be indicated by deposits of fecal matter, covered runways over the trail, or pheromones secreted by a sternal gland as the termite drags its abdomen along the ground. The pheromone odour is detected by other termites through olfactory receptors.

Termites communicate alarm by vibrations, odour, and physical contact. Alarmed termites may tap their heads against the ground, quiver and jerk, or run in a zigzag fashion, bumping into other individuals. Although the vertical head-tapping movements produce rattling sounds audible to the human ear, termites cannot hear airborne sounds. It is the substratum vibration that they sense through the vibratory receptors located on their legs. The zigzag and horizontal jerking movements communicate alarm by contact; as an alarmed termite bumps into other termites, they, too, become alarmed. During this excitatory

running, the alarmed termite leaves a scent trail (similar to the foraging trail) of pheromone that communicates direction and serves to recruit workers and soldiers to the point of disturbance.

EVOLUTION AND PALEONTOLOGY

Termites are related to the roaches and probably have evolved from a primitive roachlike ancestor. The most primitive living roach, the subsocial, wood-eating *Cryptocercus punctulatus,* which lives in rotten logs, has affinities with the termites. *Cryptocercus* harbours symbiotic, cellulose-digesting protozoans of the same genera as those found in the hindgut of primitive termites. The genitalia and certain internal structures of *Cryptocercus* have basic anatomical resemblances to those of the most primitive living termite, *Mastotermes darwiniensis,* from Australia. *Mastotermes* has further affinities with other roaches: its hindwing has a folded anal lobe, and its eggs are not laid singly as those of other termites but in clusters held together by a gelatinous material resembling the egg case of roaches.

Evidence of the relationship to primitive roaches suggests that termites evolved in the Late Permian (approximately 230,000,000 years ago), although the known fossil termites date only from the Early Cretaceous (about 130,000,000 years ago). The termite society may be older than any other society; the ant society is 100,000,000 years old.

CLASSIFICATION

Termites of the order Isoptera are small to medium-sized insects that live in social groups, or colonies, and are characterized by their highly developed caste system. The mouthparts are modified for chewing. Antennae are moniliform (beadlike) or filiform (threadlike). Isopterans are very soft bodied insects, usually light in colour. Head structures and the presence or absence of individual caste members are used to distinguish termite families.

Termites, often called white ants, differ from hymenopterans (bees, ants, and wasps) in several ways. Termites have a hemimetabolous (gradual) metamorphosis and pass through a series of nymphal stages. Hymenopterans have the more common holometabolous metamorphosis, with distinct larval, pupal, and adult stages. Termite social castes (reproductives, sterile workers, and sterile soldiers) usually contain members of both sexes in equal numbers, and both males and females develop from fertilized eggs. In the hymenopteran colony, however, the sterile castes contain females only; both sterile and reproductive females develop from fertilized eggs, while reproductive males develop by parthenogenesis from unfertilized eggs. The thorax in termites is joined broadly to the abdomen, without the "waist" characteristic of bees, ants, and wasps. Termites have two pairs of membranous wings, nearly equal in size, that break along a suture when shed, leaving only the wing base, or "scale," attached to the thorax—probably the most distinguishing characteristic of isopterans.

Annotated classification.

ORDER ISOPTERA (termites)
Highly developed caste system, may contain reproductives, soldiers, and workers; reproductives shed wings after mating; distribution worldwide, mostly in tropical rain forests; about 1,900 living, 60 fossil species; may inhabit moist subterranean or hot, dry locations; foods include plant cellulose, often digested by symbiotic protozoans in termite hindgut; all families (except Termitidae) known collectively as "lower termites" contain symbiotic protozoans in hindgut.

Family Mastotermitidae
Primitive; 1 living species (*Mastotermes darwiniensis*) in Australia; 13 Tertiary fossil species worldwide.

Family Kalotermitidae (dry-wood termites)
Wood dwelling, wood eating; survive dry conditions; 292 living, 11 fossil species (some from Baltic amber).

Family Hodotermitidae
Thirty living, 13 fossil species (1, the earliest known termite fossil, from Lower Cretaceous, Labrador); includes rotten-wood termites and harvester termites that forage and store food in nests; *Zootermopsis,* largest termite in North America, found in Rocky Mountains at altitudes of 2,000 to 2,500 metres; *Archotermopsis,* found in Himalayas; *Hodotermes* species, serious pests of African grasslands.

Family Rhinotermitidae (subterranean termites)
Lives under damp conditions; 158 living, 13 fossil species; *Reticulitermes,* widely distributed in North America and other temperate and subtemperate regions and a serious pest; *Coptotermes,* a serious pest in tropical and subtropical regions.

Family Serritermitidae
One living species in South America; specialized family evolved from Rhinotermitidae.

Family Termitidae (higher termites)
Largest termite family (about 75 percent of all termites), 1,413 living, 3 fossil species; 4 subfamilies variable in morphology, social organization, and nesting habits. (K.K.)

Phthiraptera (lice)

Lice (order Phthiraptera) are small, wingless, parasitic insects divisible into two main groups: the Mallophaga, or chewing or biting lice, which are parasites of birds and mammals, and the Anoplura, or sucking lice, parasites of mammals only. One of the biting lice, the human louse, is the carrier of typhus and louse-borne relapsing fever; it thrives in conditions of filth and overcrowding. Outbreaks of louse-borne diseases were frequent by-products of famine, war, and other disasters before the advent of powerful insecticides (see INFECTIOUS DISEASES). Heavy infestations of lice may cause intense skin irritation, and scratching for relief may lead to secondary infections. In domestic animals rubbing and damage to hides and wool may also occur, and meat and egg production may be affected; in badly infested birds the feathers may be severely damaged. One of the dog lice is the intermediate host of the dog tapeworm, and a rat louse is a transmitter of murine typhus among rats.

GENERAL FEATURES

The flattened bodies of lice range from 0.33 millimetre to 11 millimetres in length and are whitish, yellow, brown, or black. Probably all species of birds have chewing lice, and most mammals have either chewing or sucking lice, or both. There are about 2,900 known species of Mallophaga, with many others still undescribed, and about 400 species of Anoplura. No lice have been taken from the duckbilled platypus or from anteaters and armadillos; and none are known from bats or whales. The size of louse populations varies enormously on different individuals, sometimes seasonally. Sick animals and especially birds with damaged bills, probably because of the absence of grooming and preening, may have abnormally large numbers: over 14,000 on a sick fox and over 7,000 on a cormorant with a damaged bill; the numbers found on healthy hosts are usually considerably smaller. Apart from grooming and preening by the host, lice and their eggs may be controlled by predatory mites, dust baths, intense sunlight, and continuous wetting.

NATURAL HISTORY

Life cycle. With the exception of the human body louse, lice spend their entire life cycle, from egg to adult, on the host. The females are usually larger than the males and often outnumber them on any one host; in some species males are rarely found, and reproduction is by unfertilized eggs (parthenogenetic). The eggs are laid singly or in clumps, usually cemented to a feather or hair; the human body louse lays its eggs on clothing next to the skin. The eggs may be simple ovoid structures glistening white among the feathers or hairs or may be heavily sculptured or ornamented with projections that assist in the attachment of the egg or serve in gas exchange. When the nymph within the egg is ready to hatch, it sucks in air through its mouth; this passes down the alimentary canal and accumulates behind the nymph until sufficient pressure is built up to force off the cap (operculum), helped by an armed, platelike structure, the hatching organ, at the upper end of the prelarval skin. The emergent nymph is similar to the adult but issmaller and uncoloured, has fewer hairs, and differs in certain other morphological details.

Metamorphosis in the lice is simple, the nymphs molting three times, each of the three stages between molts (instars) becoming larger and more like the adult (see Figure 20).

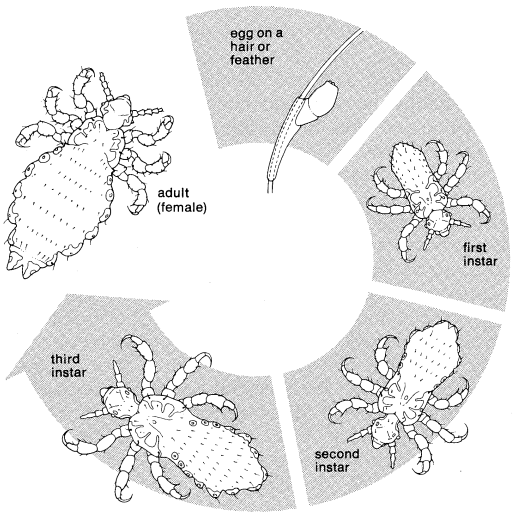


Figure 20: Life cycle of a typical louse.

The duration of the different stages of development varies from species to species and within each species according to temperature. In the human louse the egg stage may last from six to 14 days and the stages from hatching to adult, eight to 16 days. The life cycle may be closely correlated with the particular habits of the host; *e.g.,* the louse of the elephant seal must complete its life cycle during the three to five weeks, twice a year, that the elephant seal spends on shore.

Ecology. Sucking lice live exclusively on blood, their mouthparts being well adapted for this purpose. The delicate stylets, retracted into the head when the louse is not feeding, are used to pierce the skin, and a salivary secretion is injected to prevent coagulation while the blood is sucked into the mouth. The chewing lice of birds feed on the feathers, or on feathers, blood, and tissue fluids, or on fluids only. The fluids are obtained either by gnawing the skin or, as in the poultry body louse, from the central pulp of a developing feather. The chewing lice of hummingbirds have delicate mandibles that can pierce the skin of the host and a modified eversible "throat" (hypopharynx). The feather-eating Mallophaga are able to digest the keratin of feathers. It is probable that the chewing lice of mammals do not feed on wool or hairs but on skin debris, secretions, and perhaps sometimes blood and tissue fluids.

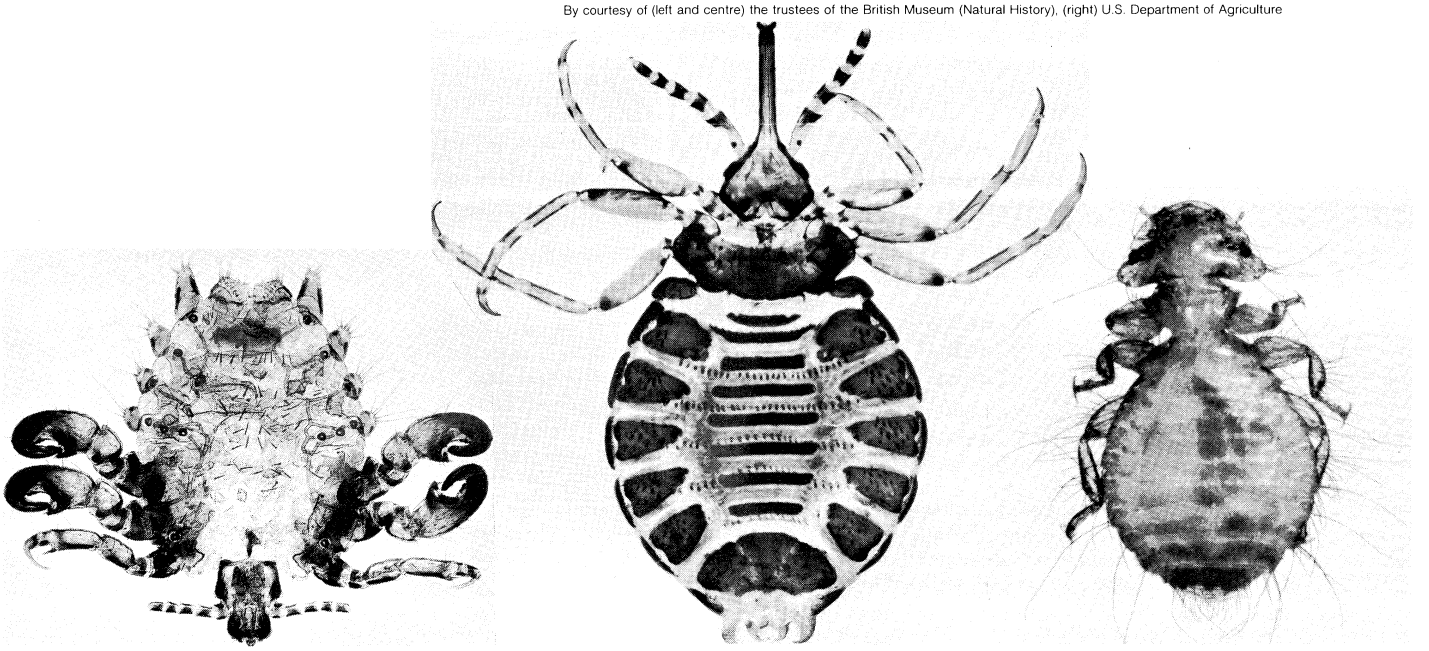
Many birds and mammals are infested by more than one species of lice, most of the birds having at least four or five; these species may inhabit different parts of the body, with adaptations to the particular habitat. Among the avian chewing lice, some species occupy different regions for resting, feeding, and egg laying. A louse is unable to live for more than short periods away from its host, and adaptations serve to maintain its close contact: being attracted by body heat and repelled by light, the louse stays within the warmth and darkness of the host's plumage or pelage. It is also probably sensitive to the smell of its host and the peculiarities of feathers and hairs that help the louse orient itself. A louse may leave its host temporarily to pass to another host of the same species or to a host of another species, such as from prey to predator. Chewing lice have often been found attached to louse flies (Hippoboscidae), also parasitic on birds and mammals, and on other insects by which they may be transferred to a new host. They may not be able to establish themselves on the new host, however, perhaps because of chemical or physical incompatibility with the host as food or habitat; some mammalian lice, for example, can lay their eggs only on hairs of a suitable diameter.

The infrequency of transfer from one host species to another leads to host specificity, or host restriction, in which a species of louse is found only on one species of host or a group of closely related host species. It is probable that some host-specific species have developed through isolation because there is simply no opportunity for the transfer of lice. Domestic and zoo animals sometimes have established populations of lice from different hosts, and pheasants and partridges often have flourishing populations of chicken lice. *Heterodoxus spiniger,* which is parasitic on domestic dogs in tropical regions, was most likely acquired relatively recently from an Australian marsupial.

FORM AND FUNCTION

The louse body is flattened dorsoventrally with the long axis of the head horizontal, enabling it to lie close along the feathers or hairs for attachment or feeding. The shape of the head and body varies considerably, especially in the avian chewing lice, in adaptation to the different ecological niches on the body of the host. Birds with white plumage, such as swans, have a white body louse, while the dark-plumaged coot has an almost black body louse. The

By courtesy of (left and centre) the trustees of the British Museum (Natural History), (right) U.S. Department of Agriculture



Body types of lice. (Left) Pubic or crab louse (*Phthirus pubis*), a sucking louse; (centre) elephant louse (*Haematomyzus elephantis*), a louse intermediate to the Anoplura and Mallophaga; and (right) pigeon biting louse (*Menacanthus latus*), a chewing louse (all magnified about 30×).

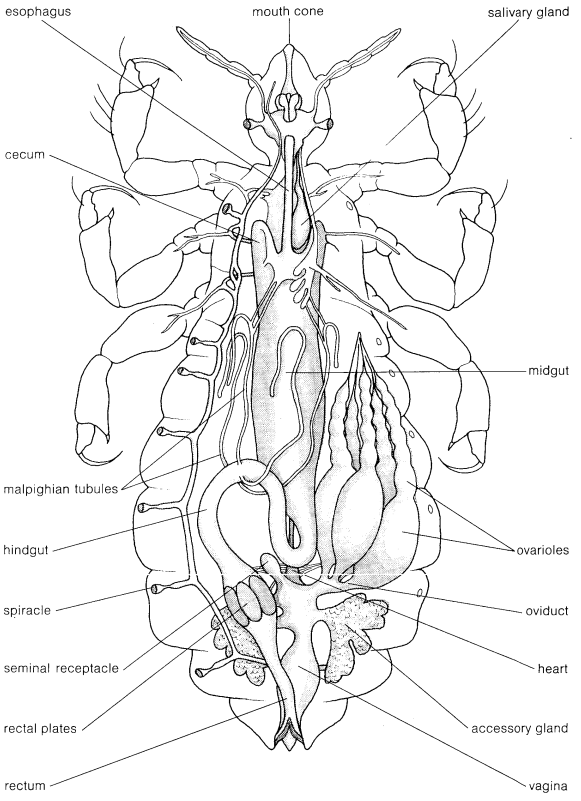


Figure 21: Internal anatomy of a typical female louse
From H. Weber, *Grundriss der Insektenkunde* (1967); Gustav Fischer Verlag, Stuttgart

antennae are short, three- to five-segmented, sometimes modified in the male as clasping organs to hold the female during copulation; the eyes are compound; the mouthparts are biting in the Mallophaga, strongly modified for sucking in the Anoplura. The Anoplura have three stylets enclosed in a sheath within the head, and a small proboscis armed with recurved toothlike processes, probably for holding the skin during feeding. The elephant louse has chewing mouthparts, with the modified mandibles borne on the end of a long proboscis. The thorax may have three visible segments, the mesothorax and metathorax fused, or all three fused into a single segment as in the Anoplura. The legs are well developed, the tarsus one- to two-segmented, with two claws in the avian Mallophaga and a single claw in some of the mammal-infesting families; the Anoplura have a single claw opposed to a tibial process forming a hair-clasping organ.

The abdomen has eight to 10 visible segments. There is one pair of thoracic breathing pores (spiracles) and a maximum of six abdominal pairs. The eversible male genitalia provide important characters for the classification of species. The female has no well-defined ovipositor, but various lobes present on the last two segments of some species may act as guides to the eggs during laying. The alimentary canal in the Mallophaga is composed of the esophagus, well-developed crop and midgut, a smaller hindgut, four malpighian tubules, and a rectum with six papillae. The crop is either a simple swelling between esophagus and midgut or a diverticulum from the esophagus. In the Anoplura the esophagus passes straight into the large midgut with or without a swelling forming a crop; there is also a strong pump for sucking up the blood. Members of the superfamily Amblycera have well-developed, comblike structures at the base of the crop, which prevent undigested feather parts or other particles from passing into the midgut; in the family Philopteridae these combs are smaller and lie at the anterior part of the crop; the Trichodectidae and Anoplura have no crop teeth. Apart from the eyes, which are sensitive to light, the other sensory structures are the tactile hairs and the sense organs in the mouth and on the antenna, some of which function as taste and smell organs.

EVOLUTION AND PALEONTOLOGY

It is generally accepted that the lice are derived from the book lice (order Psocoptera). It is also accepted that the Anoplura are related to the Mallophaga, some authorities believing that they evolved from an ancestral stock before the division into the Amblycera and Ischnocera, others that they diverged from those Ischnocera already parasitic on mammals. The origins of the elephant louse are obscure.

Apart from a louse egg found in Baltic amber, there are no fossils that might provide information on the evolution of the lice; however, their host distribution is in some ways analogous to a fossil history. Mallophagan genera frequently have a number of species that are restricted to one species of bird or to a group of closely related birds, suggesting that the stock ancestral to the bird order was parasitized by an ancestral mallophagan stock that diverged and evolved along with the divergence and evolution of its bird hosts. This relationship between host and parasite may throw some light on the relationships of the hosts themselves. The flamingos, which are usually placed with the storks, are parasitized by three genera of Mallophaga found elsewhere only on ducks, geese, and swans and may therefore be more closely related to those birds than to storks. The louse most nearly related to the human body louse, is that of the chimpanzee, and to the human pubic louse that of the gorilla. However, a number of factors have obscured the direct relationship between louse species and host species, the most important being secondary infestation, which is the establishment of a louse species on a new and unrelated host. This may have happened at any stage during the evolution of host or parasite so that subsequent divergence will have obscured all traces of the original change of host.

CLASSIFICATION

Distinguishing taxonomic features. The important characters used in classifying lice at the subordinal level are mainly based on the mouthparts. Features separating the lower categories are the special modifications of mouthparts, crop, antennae, sutures, and internal thickening of the head capsule; the number and form of claws; the segmentation of thorax and abdomen; the form of body plates; the number of spiracles; the pattern of bristles (or setae); and features of the male genitalia and terminal segments of the abdomen.

Annotated classification. Lice can be included in one order, the Phthiraptera, being separated by the characters of the mouthparts into three suborders: Mallophaga, Rhynchophthirina, and Anoplura.

ORDER PHTHIRAPTERA

Small dorso-ventrally flattened parasitic insects. Eyes reduced or absent, ocelli absent, antenna three- to five-segmented, mouthparts mandibulate or piercing. Obligate permanent ectoparasites of birds and mammals.

Suborder Mallophaga (chewing or biting lice)

Mandibulate mouthparts. Parasites of birds and mammals. About 2,900 species.

Superfamily Amblycera

Antenna four- to five-segmented, third segment pedunculate; articulation of mandibles horizontal; two- to five-segmented maxillary palpus; crop simple.

Family Menoponidae. Widespread parasites of birds, contained in about 60 genera.

Family Boopidae. Confined to marsupials of Australasia, except for one species found on domestic dogs.

Family Laemobothriidae. Contains some of the largest Mallophaga up to 11 millimetres in length. Parasites of birds of prey, rails, and some storks.

Family Ricinidae. Parasites of passerines and hummingbirds.

Family Trimenoponidae. Parasites of New World marsupials and rodents.

Family Gyropidae. New World, parasitic mainly on rodents, with one species on primates and one on peccaries.

Superfamily Ischnocera

Third antennal segment filiform; articulations of mandibles vertical; maxillary palpus absent; crop as diverticulum of esophagus; parasites of birds and mammals.

Family Philopteridae. Parasites of birds, except for one genus *(Trichophilopterus)* on lemurs; contains approximately 130 genera recorded from all orders of birds, except the swifts and hummingbirds.

Family Trichodectidae. Approximately 14 genera parasitic on mammals, rarely on primates, sloths, and rodents; more widespread on land carnivores, hyraxes, horses, donkeys, and artiodactyls, except pigs.

Suborder Rhynchophthirina

Modified mandibles borne at end of long proboscis; filiform five-segmented antennae; meso- and metanotum fused; thoracic spiracle ventral; single tarsal claw; crop absent; one genus with two species: *Haematomyzus elephantis* on the African and Indian elephants and *H. hopkinsi* on the African wart hog.

Suborder Anoplura (sucking lice)

Piercing mouthparts in the form of three fine eversible stylets; filiform four- to five-segmented antennae; all three segments of thorax fused together; thoracic spiracle dorsal; single tarsal claw, at least on second and third legs; crop absent, or if present is a simple enlargement; about 400 species.

Family Echinophthiriidae (seal lice). Parasitic on seals.

Family Haematopinidae (wrinkled sucking lice). Includes two genera parasitic on pigs, cattle, deer, and horses.

Family Hoplopleuridae (small mammal-sucking lice). A large family containing approximately 27 genera parasitic mainly on rodents but also on insectivores, primates, and one on an ungulate.

Family Linognathidae (smooth sucking lice). Parasitic on artiodactyls and hyraxes, except for two species parasitic on carnivores.

Family Neolinognathidae. Two species parasitic on insectivores.

Family Pediculidae (human lice). Two genera parasitic on man, the great apes, and on some of the New World monkeys, the last perhaps being secondarily acquired from man.

Critical appraisal. The classification given above is a currently accepted one. Other classifications rank the Anoplura and Mallophaga as separate orders, with the elephant louse placed in one or the other according to personal opinion. Some workers consider the Phthiraptera as comprising four suborders: the Amblycera and Ischnocera (both now included in the Mallophaga), Anoplura and Rhynchophthirina, dropping the name Mallophaga entirely. This represents the view that the Ischnocera, Anoplura, and Rhynchophthirina are more closely related to each other than are any of them to the Amblycera. There are considerable differences in the number and extent of the mallophagan families recognized and also on the generic limits within the families. In the Anoplura, controversy continues on the relations of the genera to each other and in which families they should be included.

The lice of man are referred to by various names, depending on whether the head louse is considered as a distinct species or as a variety or subspecies of the body louse. At present they are probably best referred to under one name, *Pediculus humanus,* but if separated subspecifically they must be called *Pediculus humanus humanus* (the body louse) and *Pediculus h. capitis* (the head louse).

(T.C.)

Thysanoptera (thrips)

Members of the order Thysanoptera, among the smallest of the winged insects, are abundant in the tropical and temperate regions of the world. The approximately 5,000 species that have been described so far are also known as thrips. Thrips have an economic importance, for some species transmit plant viruses, and by their feeding reduce seed production, disfigure flowers and fruits, and damage plant leaves. In warm areas a number of them cause plant galls and leaf rolls. On the other hand, a few species prey on destructive mites and scale insects, and a number may aid in the pollination of flowers and, indirectly, in the formation of leaf mold.

General features

Most thrips are 1.5 to 3 millimetres in size, with the smallest being about 0.6 millimetre and the largest about 15 millimetres. Because of their minute size they can enter the smallest flowers or tiniest cracks in stems and bark. Their wings, when present, are narrow and fringed, hence the name Thysanoptera (Latin, "fringed wings"). Despite their Lilliputian size, thrips are often elaborately ornamented with hexagonal designs, spines, or body flanges. The bristles (setae) of some are expanded at the tips or elongated, or the body wall may be expanded into prongs or forks. The larvae of thrips tend toward bright colours (red, orange, or yellow); the adults range from whitish to browns or black. Adults have extensible bladders on the tarsi of the legs, as suggested by the common name bladder feet.

NATURAL HISTORY

Life cycle. The development of thrips takes place in the egg, in two larval stages, during which all the immature feeding occurs, and in one, two, or three pupal stages in which no feeding has ever been observed.

Although the changes that occur in pupae of Thysanoptera resemble those that occur in insects that undergo complete metamorphosis, thrips belong to a group of insects that usually undergo simple metamorphosis.

The primary method of reproduction among thrips is by bisexual union. The males frequently are shorter lived than females, and mating often takes place early in the life span, when the adults first appear. Reproduction by unfertilized eggs (parthenogenesis) has been known to occur occasionally.

Egg laying

Thrips in the suborder Terebrantia always lay eggs, as do most thrips in the more advanced suborder Tubulifera. Hatching of eggs within the female (ovoviparity), occurs in some highly evolved members of the Tubulifera. Eggs are elongated, oval to kidney shaped, faintly to distinctly sculptured, and, as is the case in other minute insects, are

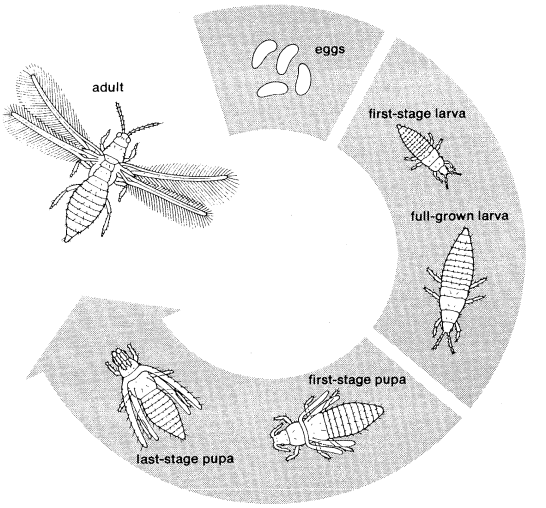


Figure 22: Life cycle of pear thrips.

proportionately large, often occupying two or three abdominal segments of the female before laying. Ordinarily, eggs of the Terebrantia are inserted into living plant tissue in a slit prepared by the sawlike ovipositor of the female. By contrast, eggs of the Tubulifera, whose females have no ovipositor, are pushed under bark, in or between flowers and leaf sheaths, or are glued onto leaf or bark surfaces. Eggs take from three to 16 days to hatch according to the species and conditions.

Although similar in overall form to the adult, the larvae of thrips are weaker and simpler in structure. The long anal bristles characteristic of larvae may hold droplets of excreted liquid that are deposited on leaf surfaces. The presence of dark coloured spots of excreta adjacent to light coloured feeding zones are some typical signs of thrips damage.

Larval stages

The first stage is short (a few days); the second may last up to many months when estivation or hibernation occurs. In the Terebrantia the second stage larva transforms into a prepupa, whereas in the Tubulifera an abbreviated form, the primipupa, is part of the sequence of development and is followed by the prepupal form. The prepupal stage is of short duration (a few hours to several days) and gives

rise to the final pupal stage, of short or long duration, that eventually produces the adult. Pupation takes place in the ground in earthen cells (in the case of many advanced Terebrantia), in cocoons (in many primitive Terebrantia), or on a host plant (in many advanced Terebrantia and the Tubulifera). Some thrips, for example, the ear thrips, have only one generation per year; others such as the onion thrips may have several generations each growing season.

Ecology. In regions having winter cold, thrips hibernate as adults or immature forms in detritus, hollow stems, or in the ground. Those that perish with the cold are replaced each year by spring migrants from warmer areas. In the tropics many thrips exhibit seasonal fluctuations, being more active in the wet than in the dry seasons. Some thrips are sensitive to relative humidity. Larvae are generally found in shady locations, on the underside of leaves or branches. Adults occasionally occur on the upper side of leaves in bright light. Most thrips rest tightly against leaf veins or in crevices. They are primarily active during the daylight hours.

Thrips are preyed upon by many insects (including other thrips), mites, birds, salamanders, and lizards. Heavy rains, winds, and dust are, however, probably as destructive to thrips as are predators.

FORM AND FUNCTION

The larva. In the larvae, the compound eyes are composed of a few facets, simple eyes (ocelli) are absent, antennal segments are fewer than in the adult, the middle portion of the body (thorax) is simple, the lower part of the leg (tarsus) is always one segmented (although also occasionally so in some adults), and the wings and genitalia are not developed. Larval bristles (setae) often differ in form, number, and arrangement from adult setae. Occasionally additional processes such as spines, combs, and plates are found on larvae.

The adult. The head of a thrips is slightly skewed with the mouth cone directed posteriorly. Generally the antennae, basically nine-segmented but frequently reduced by fusion, project forward in front of the eyes. Of the mouthparts only the single left mandible is developed and the maxillae are modified into piercing stylets. Both the maxillae and the labium bear segmented sensory projections (palps).

The middle portion of the body (thorax) bears the legs. The first part of the thorax (prothorax) of thrips bears the fore legs, the inner parts (coxae) of which may exhibit ridges that when rubbed with spurs from the lower parts (femurs) supposedly emit sound. The remainder of the thorax (pterothorax) bears the wings as well as the mid and hind legs. The wings are usually fairly simple and straplike. At rest they are laid over the abdomen but never folded. Typically the wings have long posterior fringes with shorter fringes occurring on the leading edge of the wings.

The abdomen is elongated and generally flattened dorsoventrally, especially in the Tubulifera. There are ten distinctive segments with a rudiment of the 11th. The first and eighth abdominal segments bear breathing holes (spiracles). Often in the Tubulifera, several setae on the back are S-shaped and hook into the wing fringes to hold the wings at rest. In the Terebrantia the terminal segment is rounded in the males and split ventrally up to the eighth segment in the females. In the Tubulifera the terminal segment is tubelike.

The male genital organs are composed of paired appendages and an unpaired copulatory organ (aedeagus), all of which retract into the abdomen. The paired testes and accessory glands with their ducts occur in the posterior half of the abdomen. The female of the Terebrantia usually has an ovipositor composed of two pairs of sawlike blades. These are absent in the Tubulifera. There are eight egg sacs (ovarioles) and spherical, often pigmented, seminal receptacle for storage of sperm acquired during mating.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

The Thysanoptera are part of the Psocoptera or Corrodentia-Hemiptera phyletic line of insects. Their origin may have been correlated with the development of the left mandible as an efficient chisel for piercing pollen granules. Possibly during the period when ancestral types were pollen feeders early thrips proliferated into a distinct order, which has since become diverse. Because of their small size, the development of fringe wings seemingly was favoured to overcome disproportionate drag on the uplift of the wings.

The earliest fossil, *Permothrips longipennis,* seems closest to the family Aeolothripidae, the most generalized and perhaps ancestral type. All extant families are known from fossils embedded in amber. The most advanced groups are the families Thripidae and Phlaeothripidae. The terebrantian families mostly retain their habit of feeding on vascular plants and pollen. Many members of the single extant tubiferan family Phlaeothripidae have adjusted to feeding in moldy debris.

Distinguishing taxonomic features. Although no particular structure is absolutely indicative of a taxonomic level, many features are useful in separating the major groups of Thysanoptera. The form of the antennae and its appendages, the shape and structure of the forewing, abdominal characters including form of the ovipositor in females, and glandular and setal location are important characters for determining the familial relationships. At the generic and species level antennal details, mouthparts, and body sculpturing are useful taxonomic characters, along with colour and bristle arrangement.

Annotated classification. The classification presented here is based on that of H. Priesner (1957, 1960).

ORDER THYSANOPTERA

Insects averaging about 2 mm; unique in that the right mandible is atrophied, resulting in an asymmetrical mouthcone; wings, when present, bear long fringes. Extensible bladders on the tarsi of the adults.

Suborder Terebrantia

The 10th and last abdominal segment, rarely tubelike, always split ventrally, major anal setae arising from subapical region, never from separate platelets; with sawlike ovipositor in female; wings usually with fine hairs (micotrichia) and at least 1 longitudinal vein; larvae with intermediate antennal segments bearing rings and microsetae; pupae with antennal sheaths either extended forward of head or lying over head; eggs without polar thickenings. 2 fossil families, Permothripidae and an undetermined family are, respectively, from Permian and Jurassic periods. Present families include the following.

Family Aeolothripidae

Oligocene (Baltic amber) to present. Worldwide. Antennae 9-segmented; ovipositor may be upturned or straight; forewings broad and rounded at tips, surface with microtrichia and several longitudinal and cross veins; antennal sensors on intermediate segments in form of linear or circular disks.

Family Merothripidae

Oligocene (Baltic amber) to present. Worldwide. Antennae 8- or 9-segmented; ovipositor downturned, often weakly developed; forewings narrow, surface smooth; antennal sensors on intermediate segments disclike.

Family Heterothripidae

Cretaceous (Cedar Lake amber) to present. Western Hemisphere and India. Antennae nine-segmented; ovipositor downturned; forewings narrow, surface with microtrichia; antennal sensors on intermediate segments as flat discs or protruding peglike cones.

Family Thripidae

Oligocene (Baltic amber) to present. Worldwide. Antennae 6- to 9-segmented; ovipositor downturned, rarely weakly developed; forewings narrow, with microtrichia; antennal sensors on intermediate segments as simple or forked sense cones.

Suborder Tubulifera

The 10th abdominal segment tubelike, never split, major anal setae arising from separate plates adjacent to the tube; females without ovipositor; wings without longitudinal veins or fringe; larvae with antennal segments smooth not ringed; pupae with antennal sheaths hornlike or curved around (not over) head; eggs often with a tubercule on anterior pole.

Family Phlaeothripidae

This is the only family currently assigned to the suborder. Representatives occur throughout the world are known from Oligocene (Baltic amber) to the present.

Critical appraisal. The subordinal disposition of the thrips is well established, but the ranking of family groups as given above is disputed by some who would raise the

families to superfamily rank. The order was earlier (1838) designated the Physapoda by H. Burmeister, given in allusion to the swollen tarsi, the so-called bladder feet.

(L.J.S.)

Homoptera (cicadas, hoppers, whiteflies, aphids, scale insects)

The members of this large group of sucking insects, considered a suborder of the Hemiptera by many entomologists but a separate order by others, exhibit considerable diversity in body size and number more than 32,000 species. All of the Homoptera are plant feeders, with mouthparts adapted for sucking plant sap from a wide assortment of trees and wild and cultivated plants. Many homopterans cause injuries or destruction to plants, including fruit trees and grain crops; others are vectors of plant diseases, but a few provide secretions or other products that are beneficial and have commercial value. Most members of the Homoptera fall into one of two large groups; the Auchenorrhyncha, which consists of the cicadas, treehoppers, froghoppers or spittlebugs, leafhoppers, and planthoppers or fulgorids; and the Sternorrhyncha, which includes aphids or plant lice, phylloxerans, coccids, scales, whiteflies, and mealybugs.

GENERAL FEATURES

Size range. Most homopterans range from 4 to 12 millimetres (1/6 to 1/2 inch) in length. There are certain species of cicadas in Borneo and Java, however, that are eight centimetres long with wingspreads of 20 centimetres. The large fulgorid or lanternfly can attain this size also. On the other hand, some of the tiniest scale insects are only 0.5 millimetre in length.

Distribution and abundance. Although Homoptera species are distributed throughout the world, the relative numbers of individual species vary in a given locale. Only one cicadid species is known in Great Britain, and fewer than 12 in all of Europe; however, more than 200 cicadid species are known in North America, and about 180 in Australia.

The abundance of any species in a given environment depends upon the biotic potential of the insect, the abundance of the food plant, and other factors favourable for development of large populations. Certain species never reproduce in excessive numbers, while others, considered pests, produce millions of offspring. Insect species that feed on available crops or other plants present in quantities sufficient to support them normally develop large populations; for example, the oyster shell scale (*Lepidosaphes ulmi*) on fruit trees and ornamentals; the greenbug (*Toxoptera graminum*) on wheat; and the potato leafhopper (*Empoasca fabae*) on potatoes, beans, and alfalfa. Grape leafhoppers (*Erythroneura*) frequently develop large populations that result in heavy plant losses.

IMPORTANCE

Homopterans, because they feed on sap sucked from plants, often cause injuries or destruction to the plants that nourish them. When such plants are cultivated crops (*e.g.*, grains or fruit trees) or valued ornamentals, the economic loss resulting from infestations is severe. In addition, some homopteran species act as vectors of virus- and bacteria-caused diseases of their plant hosts. The check exerted upon insect pests by other insects is an important mechanism of natural control of populations. Predacious insects feed on small, weak species; parasitic insects live on or in a host and feed at its expense. Aphids, for example, are parasitized principally by members of the Hymenoptera; two important aphid predators are ladybird beetles and lacewings. Pests also may be controlled by chemical and biological methods (*e.g.*, development of resistant plants, as with European grapevines).

The homopterans are responsible for injuring numerous plants of economic importance. Cicadas or dogday harvestflies, frequently misnamed locusts, are well-known pests. Characterized by their large size and the strident song of the male, periodical cicadas emerge every 13 or 17 years in large numbers, swarm in trees, mate, and lay eggs in green twigs. Permanent damage to fruit twigs is caused by the egg deposition slits; when the weakened twigs mature into fruit-bearing limbs, they break under the weight of the fruit, and the crop is lost. Failures of this sort can be avoided by not planting young fruit trees in years of cicada emergence.

Plant injury

Leafhoppers cause various types of plant injury by interfering with the normal physiology of the plant. The salivary secretion of the potato leafhopper, for example, causes leaf cell hypertrophy that impairs transport of sugars; the resulting sugar accumulation in the leaves destroys chlorophyll and causes the leaves to turn brown and die. This injury, termed "hopper burn," can result in complete loss of a potato crop if not controlled. Another type of injury is caused by leafhoppers that feed upon plant mesophyll tissue. In addition to removing excessive amounts of sap, these insects also destroy the plant's chlorophyll, resulting in yellow spots on the leaves, which eventually turn yellow or brown. *Erythroneura, Typhlocyba,* and *Empoasca* species cause this injury to apple trees and grapevines. Grape leafhoppers check growth and foliage function and cause formation of grapes that are inferior in size, colour, flavour, and sugar content. Plants also are injured when insects lay eggs in green twigs; for example, the egg punctures of several leafhoppers and treehoppers reduce the flexibility of plant limbs. Plant stunting and severe curling of leaves occur when the leafhopper *Empoasca fabae* punctures the undersurfaces of leaves and veins of bush beans and inhibits growth; this leafhopper attacks alfalfa and causes leaves to turn yellow and drop off. In the same way, aphids and mealybugs cause leaf curling on potatoes and many ornamental plants, and the

From Richard A. Pimentel, *Invertebrate Identification Manual* (1967) Reinhold Publishing Corporation

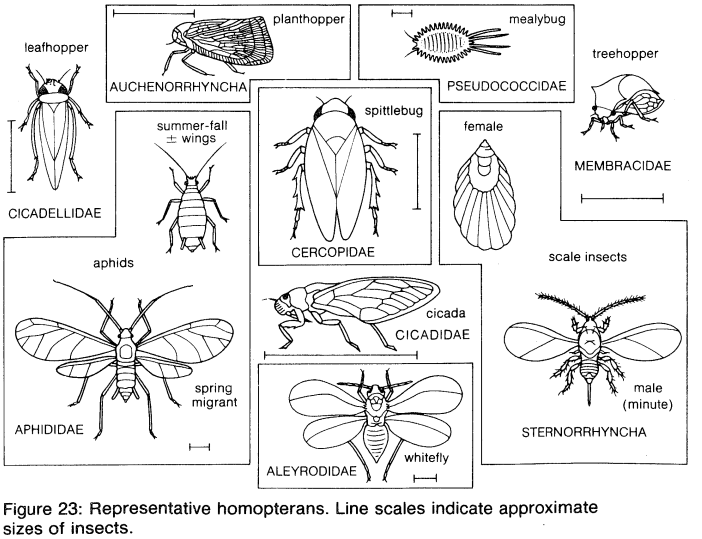


Figure 23: Representative homopterans. Line scales indicate approximate sizes of insects.

potato psyllid feeds on potato leaves and causes curling and yellowing known as "potato yellows."

The froghoppers, often called spittlebugs because immature stages live in spittle-like masses, feed on a variety of plants. One important species, *Philaenus leucophthalmus,* a meadow insect that feeds extensively upon clover and alfalfa, causes severe stunting that can result in loss of up to 50 percent of a crop. Scale insects, unless parasitized, produce enormous populations on green twigs, young limbs, leaves, and fruit; when tree bark or shrubs become encrusted with one or more layers of scales, the entire plant often dies. Damage is caused to apples by the rosy apple aphid. Females of the third seasonal generation remain on the apple leaves until after small apples have formed. Many aphids crawl onto these tiny apples and puncture them causing dimpling of the fruit and normal incision of tiny apples. The cluster of apples, known as aphid apples, are small and gnarled.

More than 100 species of leafhoppers are known organisms causing plant diseases. Some important plant disease viruses transmitted by leafhoppers are aster yellows (transmitted by *Macrosteles fascifrons*); potato yellow dwarf (several species of *Aceratagallia* and *Agallopsis*); and phony peach disease and Pierce's disease of grape (species of *Cuerna, Homalodisca,* and *Oncometopia*). Corn stunt is transmitted by species of *Dalbulus;* curly top of sugar beet by *Circulifer tenellus;* eastern and western x-disease by species of *Colladonus;* and elm phloem necrosis by *Scaphoideus luteolus.* One species of spittlebug is a vector for a yellow virus of peaches. Aphids are vectors for several virus mosaic disease organisms. A membracid species transmits the virus that causes pseudocurly top of tomato and tobacco, and two species of fulgorids are vectors of virus disease organisms of rice. The whitefly *Bemisia tabaci* transmits the virus that causes tobacco leaf curl, and species of mealybugs are vectors of the virus that causes pineapple wilt. The bacteria that cause fire blight disease on pear, apple, and quince trees are transmitted by several types of insects including leafhoppers. The bacterial pathogen, *Neofabraea perennans,* that causes perennial canker of apple is transmitted by the woolly apple aphid.

Because homopterans suck more sap from plants than they need, the surplus is excreted from the tip of the abdomen as sweet droplets known as honeydew. If the insect is feeding on apple foliage and honeydew falls on apples, a sooty fungus grows in each droplet; and the apples become black spotted and are no longer marketable.

Beneficial aspects

Of great economic importance are insects that secrete lac on twigs in tropical and subtropical regions. The lac is refined and used in preparing shellac and varnishes. More than 4,000,000 pounds of lac are refined annually. Other waxes secreted by aphids and scale insects are used in candlemaking and in medicines.

Although few homopterans produce food for man, the tamarisk manna scale, *Trabutina mannipara,* is thought to have produced the biblical manna for the children of Israel. The females produce large quantities of honeydew that solidify in thick layers on plant leaves in arid regions. This sugarlike material, still collected by natives of Arabia and Iraq, is considered a great delicacy. The term manna often refers to plant products also. Certain species of scale insects produce a gum that was used as chewing gum by tribes of North American Indians. Female root-inhabiting scale insects, species of *Margarodes,* enclose their bodies in gold and bronze coloured wax cysts that are used in strings of beads. Certain colour patterns and designs of the forewings of tropical species of leafhoppers and planthoppers have been used in artwork by various peoples. For many generations the Mexican Indians have used a black, white, and red colour design in their art. The design is that of the forewings of a brilliantly coloured *Agrosoma* leafhopper, found on bushes along streams.

NATURAL HISTORY

Life cycle. Generally, homopterans are bisexual, have periods of mating, and produce eggs; however, individual life cycles vary in length and complexity. Metamorphosis is simple or gradual; immature stages resemble adults except that the latter usually have wings. The life cycles of most homopterans are short. A typical example is the common meadow spittlebug, *Philaenus leucophthalmus;* it has one generation a year. Eggs, laid in late summer on stems or sheaths of host plants, hatch the following spring. Adults appear and mate during the summer.

Periodical cicada. The life cycle of three species of periodical cicadas is the longest known; it lasts 17 years. In the temperate zone enormous numbers of orange-winged adults emerge in spring and mate within a week; the females lay eggs a few days later. Using her strong ovipositor, the female cuts deeply into green twigs and through the harder wood of deciduous trees; she inserts 12 to 14 eggs through drilled slots into each of two chambers separated by a thin partition of wood. The female drills slots until she has deposited a total of 400 to 600 eggs. Injury to oaks, hickories, and young fruit trees is severe; and branches usually die beyond the point of egg insertion. Although eggs may be deposited in some 75 different kinds of trees or shrubs, the females prefer hickory, oak, apple, peach, pear, and grape.

Tree damage

The eggs hatch after two to six weeks; the young drop or crawl to the ground, enter the soil using their large digging claws, and begin a subterranean life, feeding on suitable tree and shrub roots for 16 years (periodical cicada). In apple orchards the young feed at depths of 2 to 18 inches; in wooded areas, at depths of 18 to 24 inches. "Harvestflies," common black and green species, appear in summer and feed on tree roots from two to five years. The periodical cicadas that live in central areas of the United States have a 17-year cycle, but three southern species complete their development in 13 years, and there are species that complete the cycle in one year. Since enormous numbers of nymphs feed on tree roots, many trees would die if the metabolic rate of the insect were not low; however, sap is taken from roots very slowly over a period of several years, and most trees survive. Although nymphs are almost full grown in eight years, they continue to feed and develop until the 13th or 17th year when mature nymphs emerge from the soil, climb any convenient tree or post, and attach themselves firmly. The dorsal line of the integument splits, and the adults emerge slowly through this opening. Adults live only a few weeks. Broods of both the 17- and 13-year cicadas have been studied. The largest and most widespread brood of the 17-year form occurs in abundance over much of the northeastern quarter of the United States.

Leafhopper. Many leafhoppers (*e.g., Empoasca maligna, Gyponana mali*) have cycles that involve passing the winter as eggs inserted in apple twigs. Other leafhoppers, however, such as *Empoasca recurvata* and *Erythroneura,* hibernate as adults during the winter. The sugarbeet leafhopper, *Circulifer tenellus,* winters as an adult in desert areas and produces an early spring generation on desert plants. As desert plants become unfavourable for feeding, the leafhoppers migrate to available cultivated plants where from one to four summer generations are produced. When the crop is harvested or the plants become unfavourable for feeding, the leafhoppers return to desert plants. Although definite alternation of desert and cultivated host plants occurs in this life cycle, no specific plant serves as a primary or secondary host. Plant selection by migrating leafhoppers is determined largely by the amount of rainfall and succulence of plants, both wild and cultivated. While most species have one generation a year, a few have two or three. The life cycle of planthoppers or fulgorids is similar to that of leafhoppers, while the pear psylla, *Psylla pyricola,* hibernates as an adult and can produce four generations of nymphs.

Alternation of host plants

Whiteflies. The whiteflies, common on citrus trees and in greenhouse plants, do not survive winter out of doors in the North but produce several generations a year in the South. The metamorphosis of whiteflies varies from the typical gradual form. In the first instar (interval between molts) the young, active, wingless forms are usually called larvae; during three subsequent instars, the sessile and scalelike nymphs undergo internal wing development. The molt from last larval instar to pupa occurs inside the last larval skin, which forms a puparium. Metamorphosis is essentially complete.

Aphids. The aphids or plant lice, soft-bodied insects that develop large populations, have several types of complex life cycles. Generally aphids overwinter in the egg stage on twigs or plant buds, usually designated a primary host. In the spring the eggs hatch into females that reproduce parthenogenetically, giving birth to living young.

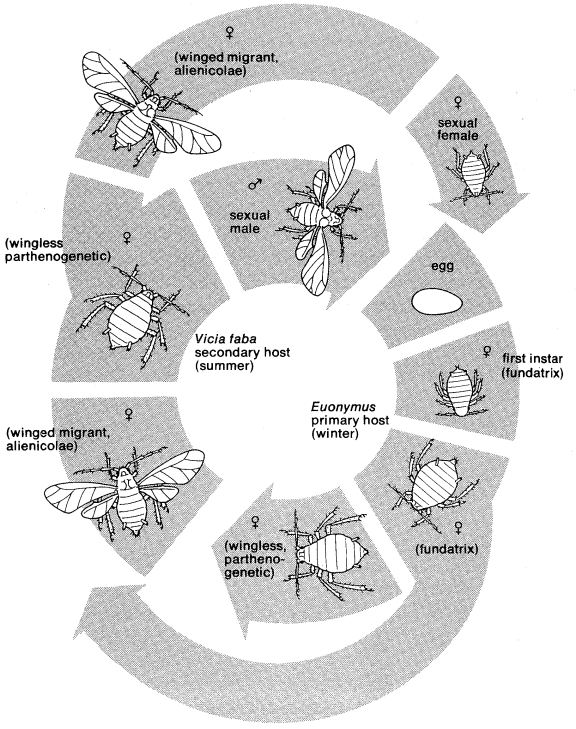


Figure 24: Life cycle of *Aphis fabae* (see text).

Several generations may be produced during the season in this way. Early generations are usually wingless, but by the third generation winged individuals appear. In many species these winged forms migrate to a secondary host plant, usually an annual, and the same type of reproductive process continues. In the latter part of the season, winged aphids of both sexes appear and migrate back to the primary host where mating occurs, and the females lay the overwintering eggs. There are two distinctive characteristics in the aphid life cycle: first, seasonal alternation of food plants involving a primary host (typically a perennial) during the winter and a secondary host (an annual) during the summer; second, there is alternation between sexual and asexual cycles, with eggs resulting from sexual mating and living young, usually females, being produced asexually.

Scale insects. The scale insects have modified life cycles also; for example, the oyster shell scale, *Lepidosaphes ulmi,* typically passes the winter as an egg beneath a secreted scale covering; however, the San José scale *Aspidiotus perniciosus* produces living young. In either case newborn young, active crawlers, leave the scale covering to search for food. After a few days they molt, losing their legs, antennae, and anal spines, and retaining poorly developed eyes. They secrete a hard scale about their soft bodies, insert their mouthparts into a plant, and remain sessile. As the females mature, they increase in size, enlarging the scale covering periodically, but do not change form or develop wings.

Young males also have a crawler stage but become sessile and inactive after the second molt, passing through a more complete metamorphosis beneath the scale covering. The last preadult instar has two external wings and is called the pupa. Adult males have two wings and two small knobs or halteres where the second pair of wings would normally develop. Some males have three pairs of eyes. Adult males seek out wingless females, concealed beneath the scale covering, and mate with them; eggs form in the female. As many as three males may mate simultaneously with one female.

Reproduction and growth. Reproduction is bisexual among the homopterans; but asexual reproduction occurs in the aphids, in a few primitive leafhoppers, and possibly in species whose life cycles are not known in detail. An unusual situation occurs in the normally hermaphroditic cottony cushion scale *Icerya purchasi;* both sexes are present in one individual, and the eggs of any individual may be fertilized by its own sperm.

Egg laying

In the Auchenorrhyncha eggs are laid by the female, who uses an egg-laying structure, the ovipositor, to insert eggs into plant tissue. In the Sternorrhyncha (*e.g.,* aphids) the female places her eggs on the surface of the plant. The eggs of scale insects are retained in the body of the female or remain under the scale covering if separated from the female. In mealybugs and certain "cottony" scales, eggs are extruded from the body and remain in a mass enclosed by waxy plates or shreds. In most homopterans, each female produces a few hundred eggs. Exceptions occur in some scale insects (*e.g.,* cottony maple scale) where a female may lay 5,000 eggs.

Growth is gradual and is accompanied by periodic molting. The nymphal stages, or instars, between egg and adult usually number five in leafhoppers and related species. Wings, if present, develop when the fifth instar molts and the adult emerges.

Sexual dimorphism occurs in most groups of Homoptera. Males and females often are coloured differently; for example, the male leafhopper *Arundanus nacreosus,* a species common on cane, is orange, the female, milk white. Size and form vary between males and females also; the male marsh leafhopper *Hecalus lineatus* is not only a different colour than the female but also only half as long. In treehoppers the pronotum (the dorsal sclerite of the prothorax) often is so different in shape and size between the two sexes that they appear to be two species; examples of this are *Umbonia crassicornis* and *Philya inflata.* Among scale insects most females lack wings, legs, and antennae, while males have all three; males and females are so diverse in appearance that previous knowledge is necessary to associate two sexes of the same species. Most homopterans lack defense mechanisms; however, one scale insect, *Phenacoccus echeveria,* extrudes a honeydew-like material from the posterior ostioles as a defense mechanism.

Ecology. *Habitats.* Every insect lives in a habitat defined by specific physical, chemical, and biological conditions. If these conditions are changed sufficiently, the insect cannot survive and will either migrate to available acceptable conditions or perish. Temperature and humidity are important climatic factors in determining geographical regions and local habitats of specific homopterans. The distribution of homopterans is influenced also by conditions that favour distribution of host plants.

Influence of plant distribution

Plant distribution is determined largely by rainfall-evaporation ratios; insects with specific host relationships occur in the same regions where the plants are found. Other climatic factors may limit the insect to a smaller range within the host plant range; for example, selection of food plants by the desert species of sugar beet leafhopper depends on the abundance of rainfall during one season. Host plants of a given species may be closely related, as legumes on which eggs are deposited and adults live; or the life cycle may be divided between alternate unrelated host plants. The fact that most species are specific in their plant relationships determines habitats such as swamp, marsh, bog, meadow, prairie, desert, deciduous or coniferous forest. Certain species occur only on sagebrush or rabbit brush in the desert, on blueberry bushes in a bog, on white oak in a deciduous forest, or on white pine in a coniferous forest.

Moisture or humidity relationships also affect the habitats of homopterans. The eggs of most auchenorrhynchans are deposited in tender plant stems or in the undersides of leaf ribs or veins. Thus, the incubation period is passed in saturated humidity. After hatching, the nymphs feed on the under surface of the leaf and remain in high relative humidity since most of the stomata, through which transpiration occurs, are on the under surface. A reduction in relative humidity due to reduced transpiration can destroy large field populations. Certain leafhopper and fulgorid

species, although they are not adapted for aquatic life, can live on plants and produce normal populations under conditions of periodic tidal submergence, even in cold waters.

Formation of galls. Insect galls, abnormal growths of plant tissue, are caused usually by the mechanical or chemical stimulus of egg laying in plant tissue and by subsequent activities of the hatching young. The young usually live and feed inside the gall and complete their development before emerging. Some 60 species of homopterans, including aphids, psyllids, and coccids, cause plant galls, although aphids are responsible for a majority of them. Galls frequently seen on foliage include aphid leaf galls, caused by the grape phylloxera *Phylloxera vitifoliae;* the leaf petiole gall of poplar, caused by the aphid *Pemphigus populitransversus;* and the elm cockscomb gall on elm leaves, caused by the aphid *Colopha ulmicola.* Different species cause the formation of different types of plant galls.

Homopteran predators

Associations with other insects. There are insects that attack homopterans as predators or parasites or use them to provision their nests. Colonies of aphids and scale insects are prey for several kinds of ladybird beetles. Female beetles lay their eggs on leaves or twigs where aphids are feeding; when the beetle eggs hatch, the larvae feed upon the homopterans in the colony. One larva of *Hippodamia convergens* can consume 300 aphids in a two week developmental period, while the adult female devours several thousand aphids in three months. Certain species of flower flies or syrphids commonly lay their eggs on leaves or twigs where colonies of aphids are feeding. The hatching larvae thrust their piercing mouth structures into the bodies of the aphids and devour them by extracting visceral and body fluids. Another predator is the aphidlion larva, a chrysopid with mandibles like the ladybird larva; however, instead of chewing the aphids, the aphidlion larva inserts its mandibles into the body of the aphid and sucks fluids through a channel or groove on the inside of its mandible. Green winged adult aphidlions lay eggs in aphid colonies, placing them on stalks, so that when the young larvae hatch, there is an adequate food supply nearby. Most larvae of the chamaemyiids (*i.e.,* aphidflies) feed on aphids, scale insects, and mealybugs; the larvae of Drosophila known as pomace flies, are predacious on mealybugs and other small Homoptera; and the larvae of a few gall midges (*i.e.,* cecidomyiids) prey on aphids and scale insects. Certain Diptera have parasitic larvae that feed on the internal tissues of homopterans including certain scale insects, leafhoppers, and planthoppers. Some moth larvae are parasites of fulgorids, while other larvae are internal parasites of female gall-like coccids of the genus *Kermes.*

Among the Hymenoptera, certain wasps are parasites of planthoppers, leafhoppers, and treehoppers. The larvae of dryinid wasps develop internally in the host although part of the body of the larva protrudes from the body of the host, forming a saclike structure between the abdominal and thoracic segments. Most encyrtid wasps are parasites of aphids, scale insects, and whiteflies. The female eulophid wasp develops as a parasite of scale insects, while the male, developing as a hyperparasite, attacks parasites of the scale insects (often females of its own species). The thamid wasps have habits similar to those of the eulophids in that both parasitize scale insects and whiteflies or are hyperparasites of chalcid wasps that parasitize homopterans.

Another interesting insect association concerns the sand wasps. They paralyze homopterans by stinging them, then store them in burrows, lay eggs, and rear young using the homopterans as food. Best known of these is the large cicada-killer wasp (*Sphecius speciosus*); the female digs a burrow in well drained soil, stings a cicada until it ceases to struggle, places it in the bottom of the burrow, lays her eggs on the cicada, covers the burrow, and dies. The larvae develop on the cicada, remain in the burrow until the following spring or summer, and emerge as adult wasps. Other wasps also burrow in the soil and provision their burrows with one or more kinds of Homoptera, particularly leafhoppers, planthoppers, or treehoppers. Aphid wasps use the same method of provisioning their nests, while squareheaded wasps usually use leafhoppers of one species to provision burrows in decomposed wood.

Many species of adult and young aphids are subterranean and feed on the roots of plants. In some species the alternate food plant is no longer used, and the aphids no longer develop wings. Some entire colonies spend years below the surface of the soil; other species spend most of each year underground; and a few species appear above ground, locate a new host plant, and immediately seek roots. The woolly aphid can live indefinitely on the roots of apple trees but can exist only part of the year on elm, the alternate host. The strawberry root louse has a sexual cycle in which eggs are laid, but these aphids are dependent upon ants for survival; not only do the ants care for the eggs in their nests but they also carry the young aphids from plant to plant. In some subterranean aphids the sexual cycle, and with it the egg-laying stage, has disappeared entirely. Subterranean aphids have no predators and few parasites. Other root feeders are young cicadas, certain young cercopids, some cixiid nymphs of the fulgorids, and immature stages of a few leafhoppers.

FORM AND FUNCTION

External features. *Polymorphism.* Polymorphism is marked in several groups of Homoptera. The Cicadidae are similar in form but vary in size and coloration. The Cercopidae include different types. The small *Clastoptera* are short, ovate, and froglike in appearance and are called froghoppers; the more elongated *Philaenus* are often called spittlebugs. The Membracidae, or treehoppers, have an enlarged prothorax that often covers the head, thorax, and abdomen; it may protrude forward resembling a coarse spine, project anteriorly at the sides resembling a pair of pointed horns, or produce a large keeled hump above the abdomen. The curiously shaped treehoppers are found in tropical American countries; here the prothorax develops into chitinous adornments and processes.

The bodies of the Cicadellidae, or leafhoppers, are dorso-

From H. Weber, *Grundriss der Insektenkunde* (1954); Gustav Fischer Verlag

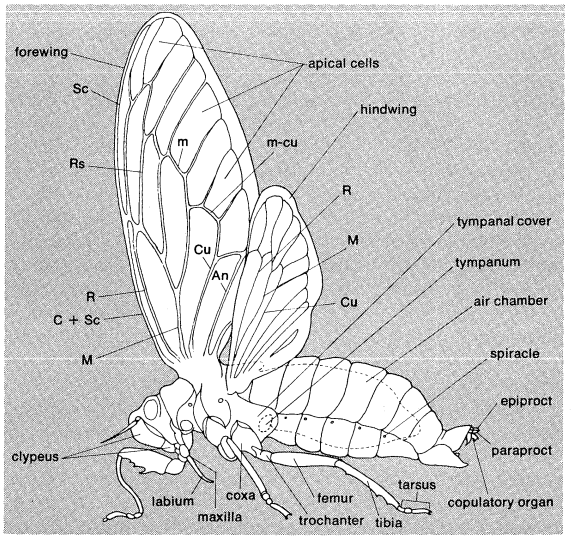


Figure 25: External features of the cicada. Wing veins: Cu, cubitus; M, media; m, medial; R, radius; C, costa; Rs, radial sector; An, anal; m-cu, mediocubital; Sc, subcosta.

ventrally flattened or cylindrical; and the head varies in shape and size from short, broad, and rounded to long, thin, and bladelike. The head size and structure of fulgorid genera vary. Species of *Scolops* have a long, slender, anterior projection of the head that resembles a beak or snout; however, the true mouth structures are beneath the head. In the genus *Apache* the head is flattened laterally and projects as a vertical thin leaflike structure, while in *Cyrpoptus* the head, flattened dorso-ventrally, is horizontal. Dimorphism in the Aleyrodidae or whiteflies occurs only when inactive and sessile immature stages succeeding the first instar pass through a quiescent stage (a pupa) that has no resemblance to the winged adult.

Typically aphids have both winged and wingless adults. Some species ("woolly" plant lice) have waxy plates or fibres over their bodies. Most aphids, however, do not secrete waxy materials. Some aphids are root feeders, some are gall formers, but most are leaf and stem feeders. Polymorphism can result from host-induced variation; for example, progeny of a European fruit *Lecanium* female develop into morphologically different insects depending on the host plant. Generally, homopteran body form is similar to that of other insects. It consists of head, thorax, and abdomen, all covered with a chitinous exoskeleton.

Head. The head is usually fitted with a pair of large compound eyes, but in certain male scale insects three pairs of eyes are present. Simple eyes or ocelli usually occur on the head and probably function as organs of light perception. Cicadas normally have three, while other homopterans have two or none. Vision is variable among the homopterans. Reactions to visual stimuli are greatest in leafhoppers, spittlebugs, planthoppers, treehoppers, and jumping plant lice; reaction is slower in cicadas, much slower in aphids, and practically nonexistent in scale insects and mealybugs. Leafhoppers as well as some spittlebugs, planthoppers, and tropical cicadas are attracted to lights at night. Members of other groups seldom respond to light.

A pair of antennae, arising below and between the eyes, are usually short and bristlelike, varying in length throughout the group. They are probably the most important sensory structures and are of taxonomic significance in species identification of aphids. Mouth structures of homopterans arise at the back of the head. The beak (or proboscis), elongate and segmented, is composed of a sheathlike labium that encloses four piercing stylets, two mandibles, and two maxillae. The stylets alone enter the plant tissue; the mandibles do the piercing; and the two inner stylets, the maxillae, fit together to form a sucking tube composed of two channels, one for conducting food, the other for saliva. In the Auchenorrhyncha the beak arises at the back of the head, while in the Sternorrhyncha it appears to arise between the front coxae; this difference has taxonomic significance.

Legs. Each segment of the thorax bears a pair of legs. In the cicadellids, fulgorids, cercopids, membracids, and psyllids, the hindfemurs are enlarged and adapted for jumping; in the other groups the femurs are normal in size. The type and arrangement of spines on the femur are of taxonomic importance in separating certain families. The mesothorax and metathorax both bear a pair of wings in the adult stage; the wings are usually membranous or of the same texture. Adult female scale insects and most female aphids lack wings. Male scale insects have one pair of wings on the mesothorax; wing rudiments, called halteres, are found on the metathorax. The first pair of wings are hyaline, opaque, pigmented with various colours, or covered with waxy secretions in the form of powder, shreds, or plates. The second pair of wings are always membranous.

Abdomen. The abdomen, typically 11-segmented, appears to have only 7 or 8 segments because the last few segments are modified as specialized genital structures. Genital segments 8 and 9 bear structures associated with external openings of genital ducts. In the male these structures are modified for copulation and transfer of sperm to the female; in the female they are modified for oviposition. Although external genital structures, these are usually enclosed in a genital chamber. The female genitalia in the Auchenorrhyncha consists of an ovipositor, formed by the appendages (gonopods) of segments 8 and 9. The ovipositor, a pair of basal plates and three pairs of elongate bladelike structures, generally is used to pierce or drill slots in plant tissue for oviposition. The variable external genitalia of the male, often complex, are frequently of considerable taxonomic value. In the Auchenorrhyncha they are contained in a genital chamber consisting of portions of the ninth segment. The genitalia consist of the styles and the aedeagus, equipped with a gonopore through which sperm are discharged during mating. When the male mates, the aedeagus and styles are exposed directly to the base of the female ovipositor, where the sperm are transferred.

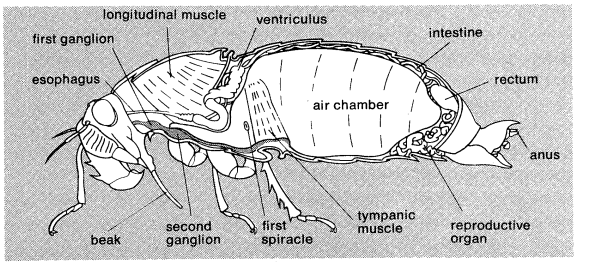


Figure 26: Internal features of the cicada.
From R.E. Snodgrass, *Principles of Insect Morphology* (1935); McGraw-Hill Book Co.

Internal features. In general the internal organs and systems are similar to those of other insects. Although the respiratory systems of homopterans and heteropterans are adapted for terrestrial life, certain species of both groups can live on submerged plants. The circulatory system is open, and blood circulates freely in the body cavity. The nervous system is composed of a ventral nerve cord with ganglionic masses for almost every segment.

The alimentary system is composed of three major parts, the foregut or stomodaeum, the midgut or mesenteron, and the hindgut or proctodaeum. The structure and function of the alimentary canal differ from other insects because homopterans feed entirely upon plant sap and ingest large amounts of it. Little absorption of food can take place in the foregut. The midgut, where digestion and absorption occur, is lined with epithelial cells that produce enzymes and absorb food after digestion. The residue passes into the ileum (small intestine) where, together with the waste products from the malpighian tubules, it passes to the colon for excretion.

Physiology and biochemistry. *Honeydew.* Plant sap contains a large quantity of water. In order to extract sufficient nutrients to survive, a large quantity of sap must be ingested. The filter chamber is a modification of the alimentary tract that allows nutrients to be concentrated in the midgut and small intestine as excess water (containing some sugar and waste materials) bypasses the midgut and small intestine to be exuded from the rectum as honeydew. It attracts ants and other hymenopteran and dipteran insects that feed on sweet nutrients.

Aphids are often called ant's cows. One well-known association is the corn root aphid and the corn field ant. The ants collect eggs in autumn, carry them to their nests, maintain the eggs through the winter, and place the young aphids on the roots of small weeds and grasses in the spring. As soon as newly planted corn seeds germinate, the ants place the aphids on corn roots and obtain honeydew by stroking the aphids with their antennae. The aphids are almost totally dependent upon the ants and are almost helpless in finding their preferred host, the roots of corn plants, without assistance. In a similar manner virgin female *Acropyga* ants carry in their mandibles on their nuptial flight a fertilized female mealybug as a source of honeydew for the new nest.

Spittle. Exuded from the alimentary tract by nymphs of the Cercopidae (*i.e.,* spittlebugs) are spittle masses commonly found on stems of meadow plants. The spittle fluid is voided from the anus after it has been mixed with a mucilaginous substance excreted by epidermal glands of the seventh and eighth abdominal segments. Air bubbles are introduced into the spittle by means of the caudal appendages of the nymph. Immature spittlebugs rest head downward on the plant; as spittle is voided, it covers the nymph and is not easily dislodged, even by heavy rains. Adults do not produce spittle.

Wax and lac

Glandular secretions. Wax, produced by numerous wax glands and secreted by cornicles, is secreted by many aphids and scale insects. Mealybugs, whiteflies, woolly aphids, and cottony scales are named for white wax on their bodies or wings. Probably the best known wax producers are males of the Chinese wax scale *Ericerus pe-la;* they secrete large amounts of pure white wax useful in making candles. The Indian wax scale *Ceroplastes ceriferus* secretes a wax that is used for medicinal purposes.

There are several lac insects; some secrete highly pig-

mented wax. The Indian lac insect *Laccifer lacca* is important commercially. It is found in tropical or subtropical regions on banyan and other plants. The females, globular in form, live on twigs in cells of resin; their bodies are covered with exudations of lac, sometimes so heavy that the twigs become coated to a thickness of 1/2 to 1 1/3 inch. The twigs are cut; lac is melted off, refined, and used in shellac and varnishes.

A group of small scale insects that typically live on desert cacti and resemble mealybugs are known as cochineal insects. *Dactylopius coccus* is the source of a natural crimson or scarlet dye called cochineal dye, originally used by the Indians of Mexico. Mature females are brushed from the cacti and dried; the pigments are extracted from the dried bodies. The Spanish used these dyes as early as 1518, and they were exported to Europe until replaced by the aniline dyes about 1870. The crimson colour of cochineal dye is attributed to cochinealin or carminic acid.

Sound production. At one time it was thought that the familiar call of male cicadas was the only sound produced by homopterans. It is now known, however, that sound production is common among other Auchenorrhyncha (leafhoppers, treehoppers, planthoppers, and spittlebugs) although their songs cannot be detected by the human ear unless amplified. Sound has been observed also in a few aphids and in one psyllid (both Sternorrhyncha).

The auchenorrhynchan Homoptera have evolved the most complex insect sound-producing mechanism known, the timbal organ. A pair of timbals, circular membranes supported by heavy chitinous rings, occur on the dorsolateral surface of the first abdominal segment. Contraction of a large timbal muscle attached to the membrane causes distortion of the timbal, producing a sharp click or pulse. The timbal springs back by its own elasticity when the muscle is relaxed. If the rates of muscle contraction and relaxation are rapid, the sound seems continuous to the human ear. The frequency contractions of the timbal muscles range from 120 to 480 per second. Associated with timbal organs in cicadas are large tracheal air sacs that open to the exterior and have resonant frequencies comparable to timbal vibration frequencies.

Although the timbal organ is similar in all species studied, the songs produced are variable; apparently this variation is caused by actions of tensor muscles and air sacs. The tensor muscles control pulse repetition frequency and sound intensity; abdominal movements control expansion and contraction of the air sacs and the consequent resonance frequency. The timbal is the only sound-producing mechanism evolved in the Homoptera except for a scraping of tibial stout hairs over reticulations of the abdomen of aphid colonies (*Toxoptera coffeae*); the sound produced is a rhythmic and synchronous scraping.

Cicada songs

Each cicada species has a characteristic song, often useful in identification. The analysis of periodical cicada songs has been the basis for morphological separation and determination of geographical range for several 13- and 17-year species. Thirteen-year species sometimes occupy parts of the same geographical areas as 17-year species. The primary song in cicadas, produced by the male, is a mating or pair-forming (aggregating) call. Female cicadas have no sound-producing organs. Males, attracted to calls of other males, stimulate each other to sing in chorus. Most mating takes place under these conditions. Male cicadas are stimulated to sing in the presence of tape recordings of songs of their species. Courtship songs or signals occur after pair formation or aggregating calls. Courtship interruption calls occur also for pair reforming. Calls produced when the insect is attacked, trapped, or in "distress" have been observed and are known as "dying yells."

Sound-producing organs occur in males of some cercopids, membracids, fulgorids, and cicadellids and in females of certain cicadellids. In *Doratura* both sexes have well developed sound-producing organs. The female of *Paropia* has a striated timbal that is poorly developed in the male. The sound-producing organ in the female is probably a primitive condition. Unlike cicadas, several leafhopper males produce calls in darkness and commonly produce mating calls when females are near. Rivalry calls between males also have been observed, usually accompanied by leg movements (kicks) that are attempts to strike and drive away a rival male.

EVOLUTION AND PALEONTOLOGY

Paleontologists do not agree on the exact or relative ages of either the Homoptera or Heteroptera (see below in *Heteroptera*). While some entomologists consider each group a separate insect order, others feel they have a common origin and classify them as suborders of the order Hemiptera.

Proto-homoptera

Although characteristics of the earliest Homoptera are not known, it is probable that the Protohomoptera had three tarsal segments, three ocelli, two pairs of wings about equal in size and shape with complete venation, an alimentary tract lacking a filter chamber, and male genitalia fitted with harpogones and subgenital plates.

Based on the primitive nature of the ovipositor, the primitive sucking pump, and simple alimentary canal, the fulgorids are considered to be different from other Auchenorrhyncha and probably evolved earliest. The ovipositor of *Scolops pungens* is more primitive than that found in many of the Orthoptera. Thus the Fulgoridae are a combination of specialized sucking mouthparts and a primitive ovipositor.

The similarity of the thoracic sterna combined with jumping hind legs places the cicadellids, membracids, and cercopids together and differentiates them from the cicadas, which have different thoracic characters, lack the enlarged hindfemurs, and have a third (median) ocellus on the head. The Cercopidae show some relationship to the Cicadidae by having a complete tentorium in which the anterior tentorial arms are connected with the posterior arms; but in this respect they differ from the Cicadellidae and Membracidae, in which the tentorial structure is reduced. The hind legs of both Cicadellidae and Membracidae bear rows of spines that are absent on the hind legs of Cercopidae. Therefore, although the cicadellids, membracids, and cercopids are related and differ from other Auchenorrhyncha, the Cicadellidae and Membracidae are more closely related to each other than are the Cercopidae to either group. Furthermore, the cicadas and cicadellids, both of which retain different combinations of primitive characteristics, cannot be related through the cercopids, which lack these characteristics entirely. The cicadellids, in their structural and biological diversity, differ from other Homoptera and show a greater array of evolutionary stages in various combinations. Unlike other groups, cicadellids contain groups that stabilized at different evolutionary levels.

The Sternorrhyncha were probably separated from the Auchenorrhyncha as early as the Lower Permian (about 270,000,000 years ago). Although small in size, many fossil psyllids are found in the Upper Permian (about 230,000,000 years ago) strata and onward. If fossil remains have been identified properly, the aleurodids date from the Upper Permian also. They are highly specialized, both biologically and structurally. The aphids exhibit various degrees of polymorphism. The female genitalia are usually reduced; however, two groups, Adeligidae (part of the family Chermidae) and the family Phylloxeridae, have retained a true basic ovipositor. A fossil wing of *Permaphidopis sojaneusis* from the Permian resembles the wing of recent aphids.

Female coccids, wingless and sessile, are unlike the tiny winged males. Although the structural characters that have not been lost provide little information on phylogeny, they do show varying degrees of specialization. Most coccids, for example, have a single tarsal segment, but all species of *Xylococinae* have two; the male genitalia is a simple tubular, heavily sclerotized organ similar to the genitalia of aleurodids. The Peloridoidea represents a primitive form, probably a Paleozoic relic (about 225,000,000 to 570,000,000 years old). Highly specialized in some respects, it still has some primitive characteristics.

The Fulgoridae were the earliest group differentiated from the base of the Auchenorrhyncha stem; the cicadellids probably were next, apparently in Late Permian or early Triassic (about 220,000,000 years ago). The cercopids probably were derived from this stock; certainly the membracids were, later. The cicadas probably arose from

the early cicadellid stem but are not found in fossils until the Cretaceous (about 65,000,000 to 136,000,000 years ago). If fossil forms have been properly placed, cicadellids and cicadas differentiated no later than the Permian; and cercopids and fulgorids must have had an even earlier origin. In the Sternorrhyncha group the psyllids, probably the earliest group to be differentiated, are known from abundant fossils in the Late Permian. The aleurodids also apparently arose in the Late Permian; the aphids date back to the Late Triassic (about 200,000,000 years ago); and the precretaceous fossil, *Mesoccus asiatica* Bekker-Migdisova, from the Permian seems to place the coccids in this geologic age.

CLASSIFICATION

Distinguishing taxonomic features. The beak and the wings are the most distinctive features of homopterans. The beak (or proboscis), fastened rigidly to the head, appears to arise from the ventral margin and consists of two pairs of stylets (mandibles and maxillae) adapted for piercing and sucking. In most homopterans, both pairs of wings are transparent or slightly thickened; the front pair have a uniform structure throughout. When at rest, the forewings are held rooflike over the dorsum with a slight overlapping on the inner margin near the tip. The digestive tract is complex, forming a filter chamber in most groups.

The suborders are distinguished by point of origin of beak, length and appearance of antennae, and number of tarsal segments. Separation of families of the Auchenorrhyncha is based on characters of the ocelli, position of antennae, form of pronotum, and spination of legs. Families of Sternorrhyncha are separated on the basis of number of tarsal segments, structure and venation of wings, and presence or absence of cornicles.

Annotated classification.

ORDER HOMOPTERA

Mostly small (4–12 mm); wings, when present, number two or four; sucking mouthparts; plant feeders; more than 32,000 species; worldwide distribution.

Suborder Coleorrhyncha

Origin of beak at antero-ventral extremity of face; propleura form a sheath for base of beak; hind wings absent; forewings held flat over abdomen when at rest; no flight function; prothorax with paranota; digestive tract lacks filter chamber.

Family Peloridiidae

Most primitive Homoptera; Tasmania, New Zealand, South America.

Suborder Auchenorrhyncha

Beak arises at antero-ventral extremity of the face, not sheathed by propleura; antennae with one to three basal segments, with a terminal seta; forewings rooflike when at rest; filter chamber present; all males apparently produce sound.

Family Cicadidae (cicadas)

Also called dog-day harvest-flies; usually large; three ocelli on face; front wings membranous; male with audible sound-producing organs on ventral base of abdomen, non jumping.

Family Membracidae (treehoppers)

Usually less than 12 mm in length; two ocelli; enlarged pronotum extends over head, thorax and all or part of abdomen; jumping hindtibia; wings largely concealed by pronotum.

Family Cercopidae (froghoppers or spittlebugs)

Less than 15 mm in length; two ocelli; jumping hindtibia with one or two stout spines, and circlet of stout spines at apex; hindcoxae short, conical.

Family Cicadellidae (leafhoppers)

Two ocelli or none; variable in size, 2 to 25 mm; jumping hindtibia with two or more rows of spines; hindcoxae transverse.

Family Delphacidae (planthoppers)

Hindtibia with broad movable apical spur; sexes often dimorphic, very different.

Family Derbidae (planthoppers)

Anal area of wing not reticulate; without cross veins; terminal segment of beak not more than $1\frac{1}{2}$ times as long as wide.

Family Cixiidae (planthoppers)

Head not prolonged in front; carina of head median or absent; tegulae present; claval suture distinct, abdominal terga 6–8, rectangular.

Family Kinnaridae (planthoppers)

Terminal segment of beak at least twice as long as broad; front wings usually not overlapping as apex of clavus; head not prolonged in front; tegulae present; median ocellus usually present; abdominal terga 6–8, chevron shaped.

Family Dictyopharidae (planthoppers)

Head prolonged in front; or frons with two or three carinae or the tegulae absent; claval suture obscure.

Family Fulgoridae (planthoppers)

Second segment of hind tarsus large, apex with row of small spines; anal area of hind wing reticulate with cross veins.

Family Achilidae (planthoppers)

Terminal segment of beak at least twice as long as wide; claval vein extending to apex of clavus; body somewhat flattened; forewings overlapping at apex.

Family Tropiduchidae (planthoppers)

Second segment of hind tarsi with two apical spines, one on each side; apex reduced or conical; front wings longer than abdomen; cross veins between costal margin and apex of clavus.

Family Flatidae (planthoppers)

Costal and/or apical border of wing with numerous cross veins; wings longer than body, in repose held almost vertically at sides of body; clavus with numerous small pustule-like tubercles.

Family Acanaloniidae (planthoppers)

Also called Amphiscepidae; hindtibia without spines except at apex; front wings very broad, costal margin broadly rounded, venation reticulate; wings longer than body, at repose held almost vertically at sides of body.

Family Issidae (planthoppers)

Wings usually shorter than body, if longer than abdomen, usually oval; clavus without numerous small pustule-like tubercles; costal border of wings usually without numerous cross veins.

Suborder Sternorrhyncha

Beak appears to arise either between fore coxae or behind them; antennae usually long, filamentous, without a well differentiated terminal seta.

Family Psyllidae (jumping plant lice)

Beak long; mouthparts well developed in both sexes; tarsi two-segmented with two claws; antennae 5 to 10 (usually 10), segmented; front wings often thicker than hindwings, not exceeding 7 mm in length.

Family Aleyrodidae (whiteflies)

Very small; covered with a white powdery, waxy material; wings opaque; not jumping insects.

Family Aphididae (aphids or plantlice)

Wings membranous, Rs vein present in forewing; cornicles usually present; sexual females oviparous, parthenogenetic females viviparous; females and usually males with functional mouthparts; without abundant wax glands.

Family Eriosomatidae (woolly and gall-making aphids)

Aphididae in part; Rs vein present in forewing; cornicles indistinct or lacking; M vein in forewing not branched; wax glands usually abundant; sexual forms with the mouthparts atrophied and not functional.

Family Chermidae (pine and spruce aphids)

Adelgidae and Phylloxeridae in part; feed on needles, twigs, and leaves of conifers; Rs vein in forewing absent; cornicles absent; all females viviparous; Cu_1 and Cu_2 in forewing separated at base; apterous parthenogenetic females covered with wax.

Family Phylloxeridae (phylloxerans)

Chermidae in part. Rs in forewing absent; cornicles absent; all females oviparous; Cu_1 and Cu_2 in forewing stalked at base; apterous parthenogenetic females not covered with wax.

Family Margarodidae (giant coccids, ground pearls, cottony cushion scales)

Males with compound eyes and ocelli; anal ring reduced, without pores or setae; females wingless and legless.

Family Ortheziidae (ensign coccids)

Male with ocelli only; abdominal spiracles absent; anal ring distinct and flat, bearing many pores and 6 long setae; females wingless.

Family Diaspididae (armoured scales)

Apical segments of female fused, forming a pygidium; female with scale covering separate from body; legs absent; beak 1-segmented; antennae rudimentary.

Family Coccidae (soft scales, wax scales, tortoise scales)
Females flattened, elongate oval; exoskeleton hard, smooth, or wax covered; legs present or absent; antennae absent or much reduced. Females often tortoise-shaped; males winged or wingless; anus covered by two dorsal plates.

Family Aclerdidae (aclerdid scales)
Scales attacking grasses; openings of wax glands rarely 8-shaped; pygidium absent; male with ocelli only; anus covered by a single dorsal plate.

Family Lacciferidae (lac scales)
Females globular in form, legless; antennae 3- or 4-segmented, minute; body enclosed in cells of resin; tropical or subtropical.

Family Asterolecaniidae (pit scales)
Females without pygidium; beak with more than 1 segment; posterior end of body not cleft; abdomen not narrowed posteriorly or produced into an anal tube; wax gland openings 8-shaped, usually in rows; legs vestigial or absent.

Family Pseudococcidae (mealybugs)
Covered with a white powdery secretion; wax gland openings on dorsum, not 8-shaped; anal ring with four or more setae; dorsal ostioles and usually 1 to 4 circuli present.

Family Eriococcidae (scales)
Pseudococcidae in part; anal ring with 4 or more setae; dorsal ostioles and ventral circuli absent; body not covered with powdery secretion.

Family Dactylopiidae (cochineal insects)
Occur on cacti; abdomen not narrowed posteriorly; wax gland openings on dorsum; anal ring absent; wax gland ducts minute, arising from centre of cluster of sessile pores; setae stout and cut off at end.

Family Kermidae (gall-like coccids)
Females spherical, hemispherical or oval; legs absent in adult; antennae 6-segmented; anal ring absent; wax gland ducts not minute, openings not 8-shaped.

Critical appraisal. The Homoptera, along with the Heteroptera, are considered by many entomologists as suborders of the order Hemiptera, mainly on the bases of similar types of piercing–sucking mouthparts and on the general type of gradual metamorphosis. This system, however, places only minor importance on the distinct differences in structure and in details of metamorphosis that have led some workers to propose separate ordinal rank for Homoptera and Hemiptera and abandonment of the term Heteroptera. The mouthparts vary considerably in detail; in the Heteroptera the beak arises from the front of the head and is movable, while in the Homoptera the beak is fastened rigidly to the head, cannot be moved, and appears to arise from the dorsoventral portion. Little considered is the fact that certain Homoptera (*e.g.*, scale insects and whiteflies) pass through a stage in their development resembling complete metamorphosis. The names Heteroptera and Homoptera are derived from their different wings. In the Heteroptera only the apical portion of the wing is membranous and has visible veins; the basal portion of the wing is thickened and leathery. When these wings are at repose, they are held flat upon the dorsal portion of the abdomen; and the apical portions of the wings completely overlap. On the other hand, the forewings of the Homoptera are either membranous or of the same texture and contain visible veins throughout. At repose these wings are held at an angle rooflike over the abdomen, overlapping only slightly on the inner apical margin. Those who classify these two groups as suborders of the Hemiptera place only minor emphasis upon these distinct differences.

(D.M.DeL.)

Heteroptera (true bugs)

Heteropterans, the "bugs" in the strictest sense, are considered a suborder of the Hemiptera by some entomologists, but an order (known either as Hemiptera or Heteroptera) by others. This large group of insects, consisting of about 30,000 species, can be recognized by an X-shaped design on the back, which is formed by the wings at rest. A combination of features—sucking mouthparts adapted to pierce plant or animal tissues and a hardened gula (underside of the head)—separate the heteropterans from all other insect orders. Although most species of Heteroptera are terrestrial, a few are aquatic. Some species, which feed on plant juices, are serious pests of cultivated crops; other species are predacious and benefit man by destroying various pests. There also are heteropterans that act as carriers of disease.

Heteropterans can be divided into three large groups on the basis of general habitat: the water-dwelling Hydrocorisae (water boatmen, back swimmers, water scorpions, giant water bugs, and creeping water bugs); the surface-swimming and shore-dwelling Amphibicorisae (water striders, marsh and water treaders, shore bugs, and velvet water bugs); and the Geocorisae, a large group of land bugs (plant bugs, bedbugs, assassin bugs, anthocorid bugs, lace bugs, ambush bugs, stinkbugs, burrower bugs, stilt bugs, and fire bugs).

GENERAL FEATURES

Size range and diversity. Like any other biologically successful group of organisms, the heteropterans are prolific and diverse and have adapted to a variety of habitats. They range in length from more than 100 millimetres (3.9 inches) to less than 1 millimetre (0.039 inch) and have invaded habitats from dry land to water. One of the few insect groups with aquatic adults capable of obtaining adequate oxygen from the water, heteropterans include the only insect species that spend their lives on oceans far from land. Some heteropterans are nourished by blood of animals ranging from mites to man; others feed from the sap of plants as diverse as giant redwoods and algae. Some suck surface fluids (*e.g.*, nectar); others pierce tissues to suck sap, blood, or even to obtain the nourishment provided by dried seeds. Many forms live on open surfaces and escape enemies by running, flying, or remaining inconspicuously motionless. Some seek food and shelter in natural crevices; others actively burrow into the soil or seek out the nests of animals.

Courtesy of G. Ferris and R. Usinger, *Microentomology*, vol. 4 (1939); Stanford University

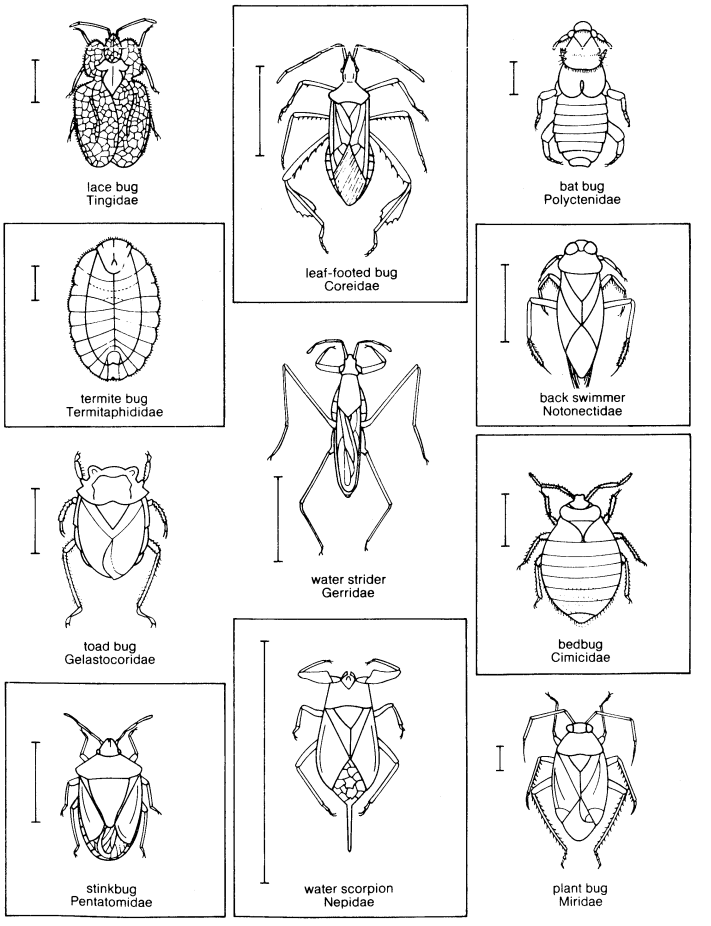


Figure 27: Diversity among heteropterans. Line scales indicate the approximate size of each insect.

Distribution and abundance. The number of individuals in a species depends in part on the availability of food. Growing crops provide concentrated food sources, but large populations, which often occur in areas of no apparent economic importance to man, may result from normal fluctuations in the balance of nature.

Heteropterans are most abundant in the tropics, decreasing in both individuals and species to limits northward beyond the Arctic Circle and southward almost to the Antarctic Circle. Every major land mass harbours different species; migrations to new habitats may be aided by natural agents (*e.g.*, wind, birds, floating debris) or by man. Although heteropterans have been carried throughout the world, only a few species have become established in many lands. Unique among insects are some water striders (Gerridae), which are at home on the open ocean between approximately 40° north latitude and 40° south latitude and may not approach land for several generations.

Approximately 30,000 species of heteropterans are known. Depending upon the availability of favourable habitats, the number of species on a major land mass is directly proportional to its size; the number of species inhabiting small islands is inversely proportional to their distance from large land masses. Age and origin of islands also influence the number of species. Most families containing 150 or more species are represented in every zoogeographic region; two exceptions are the Phymatidae (ambush bugs), with no species known in Australia or the Pacific Islands, and the Plataspidae, with no New World species. Families with the fewest species have the most restricted ranges: Aphylidae, Hyocephalidae, and Lestoniidae are restricted to Australia and contain among them only five species. The Velocipedidae of the Orient and the Vianaididae of the American tropics number four species each. In general, zoogeographic restriction is most evident at the subfamily, tribal, and generic levels.

IMPORTANCE

Heteropterans have complex and important roles in the balance of nature. The majority of them occupy an intermediate position in the ecological food chain; they use food producers (plants) and serve as food sources for parasites and other animals. A few species utilize plant-feeding insects as food.

Beneficial aspects. Heteropterans often serve as food for man and other animals. In some Latin American countries eggs of certain aquatic bugs are collected by providing submerged mats as egg laying sites; the eggs are then dried and made into cakes. Chickens, turkeys, hogs, and other domestic stock consume available heteropterans. Aquatic bugs are an important food source for fish. Wild birds and mammals, from the wrens and shrews to the turkey and grizzly bear, utilize available insects in their diets.

Species that attack unwanted plants and feed on the eggs, immature stages, and adults of injurious insects include most assassin bugs, nabid bugs, anthocorid bugs, stinkbugs, and plant bugs. These heteropterans are used in biological control of noxious weeds and injurious insects.

Harmful aspects. In contrast, some heteropterans (*e.g.*, plant bugs, chinch bug, lace bugs, stinkbugs) damage both wild and domestic plants by sucking sap or injecting tissue-killing fluids. Insertion of a beak or eggs into a plant tissue harms the protective layers and allows bacteria and other disease-causing organisms to infect the plant. Although heteropterans are not the major agents for transmission of plant viral diseases, a few species (Piesmatidae) can carry viral diseases of sugar beets and related crops in North America and Europe.

Role in disease

Heteropterans can affect man in several ways. They are common household pests and can spoil the taste of some fruits (*e.g.*, raspberries contaminated by stinkbugs). But more importantly, some can attack man directly and inflict painful bites as well as introduce disease-causing organisms. The injection of saliva or poison may cause allergic reactions in susceptible persons. Sometimes when they cannot find plants, phytophagous insects probe moist surfaces (*e.g.*, perspiring skin) in search of appropriate food fluids. Transmission of trypanosomes, which cause Chagas' disease in the American tropics, occurs through cone nose bugs (Reduviidae), so-called because of the shape of their head. The insect receives trypanosomes when it feeds on the blood of an infected human. The trypanosome passes part of its life cycle in the insect and again becomes infective to man. Instead of injecting them into a new victim, however, the insect deposits trypanosomes (in its excrement) on the skin of a potential victim. The trypanosomes can enter the bloodstream only through mucous membranes or breaks in the skin (*e.g.*, those that may result from scratching the site of the bite).

NATURAL HISTORY

Life cycle. *General features.* Heteropterans are influenced by seasonal temperature fluctuations in temperate regions. Although the tropical regions, in which they are most abundant, impose no seasonal pattern, heteropterans are influenced by other factors—rainfall, host plant development, internal rhythms.

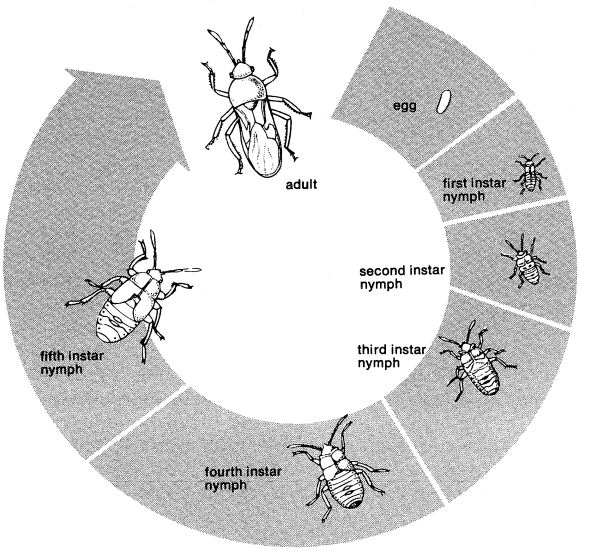


Figure 28: Life cycle of chinch bug.

Dormant stage

Heteropterans generally produce one or two generations a year in temperate regions. The dormant stage, which varies from species to species, is dependent in part on the food source of the species; for example, a heteropteran that feeds on perennial plants often lays eggs that remain dormant through the winter in or on the host plant. The nymphs, which hatch in the spring, then have appropriate leaves and succulent stems to furnish the sap necessary for growth. In contrast, heteropterans that feed on annual plants have a dormant winter adult stage that flies to a new host plant each spring after seeds germinate. Species that have short nymphal and adult stages, which coincide with certain plant characteristics (*e.g.*, short-lived flowers), may pass 10 or 11 months as eggs. Among predatory heteropterans, those whose prey is found on plants or in flowers winter as adults; those whose prey is found on the ground may be dormant as nymphs or adults. Aquatic heteropterans may spend the winter as adults, nymphs, or both; semiaquatic forms are dormant as adults.

Metamorphosis. Heteropterans undergo gradual metamorphosis (hemimetabola). Primary feeding and energy storage, as well as development of adult structures, take place in the nymphal stages. The adults seek mates and potential food sources and are responsible for initiating future generations. The developing embryo, incapable of obtaining food or defending itself, is encapsulated within a protective egg shell that contains adequate food. The nymph, which lacks the locomotor and reproductive ability of the adult, also lacks the hard and rigid extensive exoskeleton necessary for strong muscle action. As a result, however, its body can distend to accommodate large quantities of food. In addition, loss of moisture through the thinner, more membranous body cover is offset by the high water content of the nymph's food.

Reproduction. *Eggs.* Reproduction occurs when eggs

are fertilized as they pass through the female ducts by spermatozoa stored in special receptacles in the female. A notable exception to this method occurs in the bat bugs (Polyctenidae). In this case shell-less, yolkless eggs are fertilized in the female ovariole where they remain; the developing embryo is nourished by special maternal cells through a placenta-like structure consisting of tissues from both parent and offspring. The bat bug nymph leaves the mother's body only after reaching an advanced stage of development. Heteropteran eggs generally are moderate to large in size relative to adults and have a smooth to finely sculptured surface, sometimes with a colour pattern and often with slender projections. Eggs contain sufficient nutrients to permit the embryo to develop into a free-living, sexually immature, wingless nymph. Just before an egg hatches, the shell becomes translucent and reveals the segmentation and strongly coloured structures (*e.g.*, eyes) of the nymph inside.

Egg-laying sites

Eggs may be laid singly or in clusters. Eggs of plant-eating species generally are glued to a surface or inserted into the tissues of a selected host plant. Eggs of predatory forms often are laid near prey. Aquatic bugs may lay their eggs above or below the water's surface. Some bugs have special habits; for example, eggs may be glued in a large cluster to the back of a heteropteran male (Belostomatidae) or to the body of a crayfish as in Corixidae.

Nymphs. A newly developed nymph uses one or both of the following mechanisms to escape the egg: a cuticular spine on the head, sometimes known as an egg burster; or internal hydrostatic pressure created by forcing fluids (sometimes in the head) against the site of egg rupture. The pattern of rupture may be controlled by a line of weakness in the egg; in some cases a flaplike operculum lifts back to allow the nymph to escape. After withdrawing itself from the egg, the newly hatched nymph seeks the organic food essential for its development. Some nymphs leave the egg with a sufficient store of nutrients to allow them to pass through their first molt (or ecdysis). This is especially important to predatory species whose newly hatched nymphs might have to spend considerable time searching for prey.

The wingless nymph must rely on mimicry or disguise to prevent detection by enemies. Relatively inactive nymphs often assume colours and shapes of objects in their immediate environment; others disguise their bodies with a layer of small particles collected on a coat of sticky body hairs; and still others exhibit contrasting colours that break up the body outline. Some active nymphs confuse predators by alternating short quick dashes with periods of no motion; others deceive predators by simulating body shape, colour, or erratic gait of ants, with which nymphs sometimes are associated (see COLORATION; MIMICRY).

Lacking a pupal stage in which to change into an adult, the flightless nymph must sustain itself while it gradually transforms into an adult. As the nymph feeds and grows, it discards through five successive molts (ecdyses) the growth-restricting outer layers of its exoskeleton. Wings appear first as slight outpocketings at posterior regions of the mesothorax and metathorax during the third instar (interval between molts) and enlarge during the fourth and fifth instars; however, the wings are not fully functional until the last ecdysis, when the nymph becomes an adult. At the time of the last molt, other changes occur: the head and thorax assume new shapes; the number of segments in antennae and tarsi may change; scent glands on the top of the abdomen cease functioning and are replaced by metathoracic scent glands; and the external genitalia and the internal reproductive organs become functional.

Transformation to adults

Adults. Flight is common among adult heteropterans; although they are capable fliers, no species has developed the aerial efficiency of some other insect groups (*e.g.*, dragonflies, flies). Predatory heteropterans lie in wait for a potential victim to wander close enough to be captured, instead of overtaking their prey in flight. Winged adults may rely on flight or inaction and the optical deceptions of colour arrangement and body form utilized by the nymphs to escape enemies. Flight also enables the two sexes to come together and provides an effective means for seeking favourable egg-laying sites.

Behaviour. *Sound production and reception.* Heteropterans cannot produce the conspicuous sounds typical of katydids, crickets, and cicadas. The whirring or buzzing noise common to heteropteran flight is caused by the rapidly moving flight mechanism. Sound production by specialized body parts (also called sonification or stridulation) is common among heteropterans but seldom loud enough to attract human attention. Heteropterans produce sounds by moving one roughened member over a roughened area of body surface. Sometimes both roughened parts consist of a series of minute, closely-set, parallel furrows or ridges called a strigose area, strigil, file, or rasp. Sometimes the second part of the mechanism consists of a series (called a plectrum) of minute pegs, setigerous tubercles, or an upturned edge of a sclerite (hard body plate). The plectrum may be on the movable body member or on the fixed body part. One or both sexes, and sometimes even the nymphs, of a species may stridulate. Some sounds appear to have courtship significance. The presence and form of sound-producing mechanisms are useful in the classification of Heteroptera.

Mechanisms for detecting airborne and possibly waterborne vibrations (or sounds) are known in Heteroptera: auditory organs known as Johnston's organs are found in the antennae of all species, and the tympanal (stretched membrane) organs are known in several families of aquatic bugs. A variety of receptors, especially for detecting substrate vibrations, have been described for members of this order.

Feeding. Heteropterans are essentially nonsocial insects. Newly hatched nymphs occasionally remain together for a short time, clinging to the shells of the eggs from which they have hatched; for example, plant feeding (phytophagous) nymphs may remain together because they hatch on their food source and need not search for it. In contrast, however, most predatory heteropterans encounter their prey by chance; in this case, dispersion increases individual opportunities for finding prey and for escaping other predatory heteropterans. Apparent coordinated migrations reported for chinch bugs generally result from a sudden disappearance or failure of a food source.

Ecology. Heteropteran behaviour, governed by natural instincts, includes patterns that enable individual species to live within specific environments. The patterns are so similar for all members of a family that each family can be characterized as essentially phytophagous or predatory, and terrestrial, semiaquatic, or aquatic. Most heteropterans rely on atmospheric oxygen for respiration.

Habitats. Aquatic Heteroptera vary in their independence of atmospheric oxygen. Members of one of the nonswimming families (Nepidae) submerge just below the water's surface and obtain oxygen through a slender, terminal breathing tube that breaks the surface film. Some swimming families (Corixidae, Notonectidae) dive deeply; however, they must surface periodically to replenish air supplies either stored in chambers formed against body surfaces by the folded wings or entrapped by rows of dense hairs folded against the body. These insects swim to overcome the buoyancy of the renewed air supply, grasping underwater objects to anchor themselves. Heteropterans with air storage chambers are independent of oxygen dissolved in the water; therefore, they survive in warm springs or polluted waters almost devoid of oxygen. Another family (Naucoridae) contains species that spend their entire lives below the water's surface; these heteropterans can obtain oxygen from that dissolved in the surrounding water. Water insulates its inhabitants from cold temperature extremes; thus aquatic species exhibit considerable activity the year around. Chilled water holds more dissolved oxygen than warm water, and air is available also between the water's surface and the ice layer during winter months in temperate regions.

Aquatic species

Semiaquatic bugs literally walk on water; they are supported solely by the water's surface film. Their small, slender bodies have little bulk to support. Long, thin middle and hind legs spread the points of contact with the water; the tarsus of each leg causes only a slight dent in the surface film. The shorter front legs may rest on the water or capture and hold prey. Wetting of the body during

submersion is prevented by a thin layer of air trapped by the fine, dense, water-repellent (hydrofuge) pile that covers the insect. Many surface dwellers live in quiet water; others prefer flowing water, even rapids, where considerable strength is needed to overcome the current. In some cases, a bug may anchor the legs on one side of its body to some fixed object in a stream and wait for food carried to it by the rushing water. Adults survive the winter beneath objects on land either close to or some distance removed from a body of water.

Although shore bugs normally frequent shore lines, they occasionally make short forays onto the surface of adjacent bodies of water. Shore bugs generally are heavier bodied than surface dwellers; the strong, and often long, legs of some shore bugs enable them to run quickly across ground, rocks, and other exposed surfaces. Short rapid flights are not uncommon. All shore bug families are predatory. Their front legs often are thickened distinctly and sometimes spined for grasping prey. Adults generally overwinter beneath convenient objects.

Terrestrial species

Terrestrial heteropteran species are more abundant than either aquatic or semiaquatic forms and have a variety of habits and life histories. Some species wander over leaves, flowers, and other surfaces exposed to sunlight; others prefer the shadows (*e.g.,* underside of foliage). Many forms hide under rocks, logs, loose bark, or in crevices. Some species, which are active during the night, hide only during the day; other groups (*e.g.,* Aradidae) live in obscure habitats and leave only if the environmental conditions become unsatisfactory. Members of one family (Cydnidae) burrow by cutting soil in front of them and shoving it behind; these insects live in moving underground chambers. Species in at least two families (Nabidae, Reduviidae) live in spider webs and skillfully avoid entanglement as they eat trapped insects or, in some cases, the spider itself.

Relationships of heteropterans to plants vary. Plants may serve simply as a site on which to stand or move, or as a lure for plant feeding insects on which predatory bugs feed; plants are most important as a source of nourishing sap for phytophagous heteropterans. A few species can cause modifications in plant growth; for example, a feeding insect may cause leaves to roll up around it. In Europe certain lace bugs (Tingidae) insert their eggs into flower buds, which enlarge and house developing lace bug nymphs. A new adult lace bug needs no jaws to gnaw its way out; it leaves when the bud begins to open.

Food habits. Predation is common throughout the Heteroptera, sometimes as a family-wide feeding pattern, sometimes as a deviation within an otherwise phytophagous family. In the suborder Hydrocorisae all families except one (Corixidae) are predatory on small aquatic animals, including small fish, tadpoles, and frogs; the exception has been known to add mosquito larvae to its usual algal diet. The suborder Amphibicorisae, without reported exception, contains forms that prey on plankton, small arthropods, or worms. In the suborder Geocorisae, many families of the series Cimicomorpha (*e.g.,* Reduviidae, Nabidae, Phymatidae), are predatory, generally on other insects; however, a few assassin bugs (Reduviidae), bedbugs (Cimicidae), and bat bugs (Polyctenidae) seek the warm blood of birds and mammals. In addition, some members of plant feeding families (including the Miridae) attack other arthropods; a few families (Pentatomidae, Lygaeidae) of the series Pentatomomorpha, fundamentally plant feeders, contain insect-eating species. The original food habit of this order is not yet clear, but even in some avidly predatory species, the first and second nymphal instars may be plant feeders. In capturing its prey, a heteropteran may use front legs modified for grasping or insert feeding stylets. In either case the victim is paralyzed quickly by an injected salivary secretion.

Predators

Any insectan stage may fall to the hungry predator. Insect eggs are a favourite food for many small (Anthocoridae) or very young heteropteran predators. Caterpillars and other larvae commonly are attacked by predatory Heteroptera, and it is estimated that one stinkbug (Pentatomidae) may destroy more than 200 caterpillars during its lifetime. One stinkbug, long known to feed upon bedbugs, was recommended for use in control as early as 1776.

Other arthropods also may be attacked. One anthocorid bug (Anthocoridae) may destroy an average of 33 mites per day. In some instances size of prey is no barrier. In Sri Lanka a small assassin bug (Reduviidae) only 4 millimetres long can paralyze a giant millipede 150 millimetres long.

Associations. Individual heteropterans may associate accidently on a food source. On quiet pools of water, water striders (Gerridae, Veliidae) often assemble in milling "schools" of a few to hundreds of individuals; they quickly reassemble if dispersed. Sometimes a female will reassemble disturbed batches of eggs. In cases where the female lays her eggs on the back of a male, the parental responsibility is his.

Members of one family (Termitaphididae), unable to live alone, share the nests of certain termites and feed on the fluids of the fungi growing there. In exchange, a substance secreted from pores on the backs of these bugs is eaten by the termites. Reports of heteropteran associations are difficult to assess and may reflect coincidental use of the same environment.

FORM AND FUNCTION

General features. Heteropterans have features typical of all insects and special adaptations that differentiate them from other insect groups. The head tests new environments through light perceptions by compound eyes and simple eyes (ocelli) and through both scent and sound perceptions by the antennae. The mouthparts are located conveniently for quick exploitation of newly discovered food sources. The thorax bears support and transportation

Special adaptations

Adapted from (left) H. Weber, *Grundriss der Insektenkunde,* Gustav Fischer Verlag, Stuttgart; (right) *Introduction to Comparative Entomology* by R. Fox and J. Fox. Copyright © 1964 by Litton Educational Publishing, Inc.

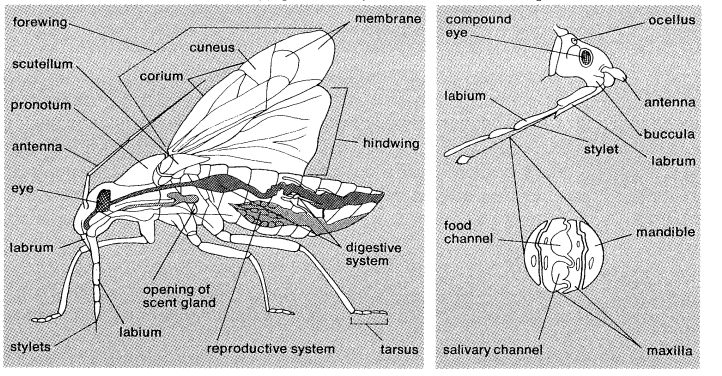


Figure 29: (Left) Heteropteran body plan. (Right) Section through stylet showing relationships of mandibles and maxillae.

mechanisms, legs, and wings. An important function for the third body region, the abdomen, is reproduction. Sutures divide the body surface into sclerites (hard plates), which may be fused or separated by flexible membranes. Hairs, scales, or a variety of glandular secretions may adorn at least some areas of the body. The body surface may be raised into spines of a variety of shapes, knobs, or tubercles; tubercles often are setigerous; *i.e.,* glandular with a straight or curved filament or bristle at the tip. Some of these projecting structures, which have bases associated with specialized cells, probably have a sensory function and respond to vibrations of air, substrate, or touch. Certain long hairs, each arising from a ringed socket, are called trichobothria, and their presence and arrangement on various parts of the body have taxonomic significance.

Distinguishing features. The covering of the head is a sclerotized ring tapering from the flexible neck forward or downward and ending in three lobes: a jugum on each side and a central clypeus (or tylus). The area above and between the eyes (called the vertex) may bear a transverse pair of simple eyes (a median one is never present in Heteroptera). The hard lower surface of the head is called the gula (hypostomal bridge).

Mouthparts. The piercing–sucking mouthparts are composed of a troughlike, four-segmented labium in which lie four stylets; these are modified mandibles and maxillae.

Each of the hairlike maxillae has two major grooves plus minor grooves and ridges along its median surface. When brought together and locked by the minor grooves and ridges, the two major grooves form the left and right halves of two separate tubular canals that extend the length of the maxillae. The anterior canal is the food tube through which fluids are drawn by a sucking pump (described below); the posterior tube is the salivary canal through which digestive and certain other fluids may pass into a food source. These fluids stun or kill the prey, prevent coagulation of blood, and initiate the process of digestion. Along the sides of the maxillary stylets are slender mandibles with serrated tips. The mandibles and maxillae alternately advance into the food tissue until the appropriate fluid is reached by the maxillae; then ingestion begins. Since the labium does not penetrate the food tissue, the stylets must extend beyond the protection of the labial groove. This extension may be accomplished in one of two ways. The labium may be shortened by jackknifing some of its segments, leaving only the apical and sometimes the basal segments to guide the exposed stylets. In some species the extensible stylets are much longer than the labium; when not extended, the stylets either are coiled or looped in a chamber located in the labium or the head, or are found in a long sheath that extends backward into the heteropteran's body cavity.

Sucking pump

The sucking pump is a chamber whose front wall is folded lengthwise into the mouth cavity; the fold is pulled out and springs back to change the size of the cavity. Contraction of certain muscles increases the size of the cavity, decreases the pressure in it, and allows the liquids to flow through the food canal into the pump. When the muscles relax, the size of the chamber decreases gradually and pushes the food toward the true mouth, which opens into the esophagus.

Thoracic features. Each of the three somites of the thorax bears a pair of legs; both meso- and metathorax bear a pair of wings. The large conspicuous part of the thorax between the head and the wings is called the pronotum (*i.e.*, the expanded top part of the prothorax). Projecting from beneath the hind margin of the pronotum is the triangular, sometimes U-shaped, scutellum. Along the side margins of the scutellum may be a groove or fold (the frenum). In repose, the inner edges of the front wings hook onto the frenum. On each side of the metathorax may be a pore (ostiole) that provides an outlet for a highly volatile, repellent fluid secreted by internal scent glands. The ostiole may be surrounded by an elevated ridge (peritreme), and then by an extensive, roughened, evaporative area.

Legs. The legs may be modified in various ways to serve specific functions. Legs may be lengthened or expanded to aid in camouflage. Sometimes the hind femurs and their accompanying leg muscles are adapted for jumping. The middle and hind legs may be flattened for swimming, or their effective width increased by a marginal fringe of long hairs that spread during a power stroke and flatten against the leg on the recovery stroke. The front legs, when specialized for grasping prey, have tibia and tarsus adapted to fold tightly against a thickened femur; stout hairs or spines along one or both of the opposing surfaces of this grasping leg prevent a victim from slipping free. The males of many species have hind legs that are thickened or spined to hold females during mating. Lacking mandibulate mouthparts with which to clean their antennae, many Heteroptera have a comblike series of hairs on the end of the front tibia, across which the antennae may be dragged. Certain genera (Nabidae, Reduviidae) that live suspended slothlike from the underside of spider webs have long, foldable claws.

Wings. The texture of each of the two pairs of wings is distinct in heteropterans. The mesothoracic or forewings (called hemelytra, elytra, or tegmina) are stiff and have an oblique line that abruptly separates the leathery basal half from the membranous apical half; the metathoracic or hindwings are thin, delicate membranes. At rest the forewings are folded over the hindwings; the membranous tips of the forewings overlap, so that the fragile hindwings lie beneath them. In macropterous individuals with wings of normal size, the wings approach or extend beyond the apex of the abdomen; some species have wings that are brachypterous (slightly to moderately reduced), micropterous (very small), or absent. Both pairs of wings are absent in one family (Termitaphididae), apparently an adaptation to the termite tunnels in which they live. In the bat parasite family (Polyctenidae) the budlike character of the nymphal wings is carried into the adult stage with no further development; in effect these insects also are wingless. The front pair of wings of the bedbug family (Cimicidae) are reduced to short, nonfunctioning pads; the hind pair is lost completely. Although these heteropteran families are the only examples of a general pattern of wing loss or reduction, some individual species show a variety of wing developments (alary polymorphism) or aptery (winglessness) in one or both sexes. Species with sexual dimorphism in wing length have females with reduced wings or none at all; apparently the retention of flight in the male assures his ability to move to the female.

Diversity of wings

Wing modifications

Heteropteran wings, especially the front pair, may be modified for specific functions. Their presence above the abdomen provides protection for the soft parts beneath. More effective protection is provided in forms with fused forewings. Some forms can couple their wings by means of a fingerlike projection (jugum) on each forewing, thus enhancing their flying ability. The colour and texture of the forewings may be modified so that an insect resembles bark; bright or contrasting wing colours announce unpalatability to an enemy. The lace bugs (Tingidae), so-called because they have a lacelike pattern of many fine, interconnecting wing veins, resemble the fine veins on the undersides of the leaves on which they feed.

Abdomen. The abdomen contains 11 segments; the tenth and eleventh, and sometimes others, are fused. Sensory appendages (cerci) are never present in this order; the sensory function may be performed in part, however, by several apparently tactile hairs along the underside of the abdomen. The terminal abdominal segments of the male are adapted primarily for transfer of sperm, those of the female for reception of sperm and egg laying. Females that insert eggs into plant tissue have an ovipositor under the tip of the abdomen; the ovipositor consists of blades that slice plant tissue and guide the eggs into the slit. In females that glue eggs to a surface, the ovipositor usually is reduced to a noncutting egg guide.

Internal features. *Digestive system.* The three embryological regions of the alimentary canal become modified by constrictions, dilations, and outpocketings. The stomodaeum becomes a short esophagus that connects to the midgut. The midgut frequently differentiates into four distinct regions, two dilations connected by a short, narrow tube and a second tubular section that gives rise to slender, terminally closed, tubular outpocketings, called mycetomes. The cavities of the mycetomes are filled with bacteria, whose role may be either to supply essential nutrients or to inhibit development of other bacteria. The bacteria, hereditarily transferred from mother to offspring, migrate to the ovarian follicle of the female and invade the eggs. As soon as an embryo begins to develop a gut, the bacteria appear in it.

Role of bacteria in mycetomes

Respiratory system. The respiratory system consists of longitudinal tracheal trunks that branch internally and communicate with the external air through ten pairs of holes called spiracles. Respiration under water presents special problems. Young aquatic nymphs may respire exclusively through the thin body wall. Adult aquatic insects, with their hard body walls, must rely on tracheal respiration; modifications of the spiracles and external skeleton surrounding them serve to exclude water while admitting oxygen. In some species the spiracles are sunken and protected by a circlet of hydrofuge (water repellent) hairs or by a fine-pored membrane; in other aquatic heteropterans the spiracles are closed and only one or two pairs function in association with either diving air storage chambers or the base of a snorkel-like breathing tube that breaks the surface film of the water. Although air storage chambers absorb some dissolved oxygen from the surrounding water, they gradually diminish in size as the nitrogen contained in them dissolves outward into the surrounding water. Unless they are replaced by frequent trips to the surface,

air supplies may be lost completely, thus allowing water to reach and enter the spiracle; drowning would result. A few forms (Naucoridae) have "plastron" respiration. The plastron is actually a modified air storage chamber that consists of air in a series of cuticular grooves radiating from a spiracle. This air, continuous with tracheal air, is held in each groove by hydrofuge hairs, so that its volume does not change. Sustained absorption of dissolved oxygen by the plastron eliminates the need for contact with atmospheric air as long as the surrounding water is well oxygenated.

Other systems. The skeletal, muscular, circulatory, and excretory systems are structurally and functionally typical of a winged insect with the exception of some adaptive variations. The piercing-sucking mouthparts require protractor and retractor muscles for insertion and withdrawal of each stylet, and dilator muscles to enlarge the sucking pump cavity. General blood flow throughout the body is provided by the four- or five-chambered heart, but circulation in legs and wings requires special devices (*e.g.,* the pulsating membranes located within the base of each tibia and sometimes in the tarsus). Rhythmical waving of pulsating membranes causes a pattern of blood flow to the tip of the limb and back. The number of excretory tubes (*i.e.,* malpighian tubules) generally is four.

The nervous system consists of three interconnected parts known as central, visceral (sympathetic), and peripheral parts. The ventral nerve cord exhibits considerable concentration near the head end (cephalization) of the ganglia; in all species the abdominal ganglia have migrated into the thorax. Most members of the order have three ganglia on the ventral nerve cord; the subesophageal, the first thoracic, and a mass that results from fusion of the second and third thoracic ganglia with all the abdominal ganglia. Complete cephalization of the ventral nerve cord occurs in a few Heteroptera where all the ventral ganglia, including the subesophageal, unite into a single ganglionic centre.

Reproductive system

The heteropteran reproductive system consists of a pair of gonads whose ducts unite to form a single tube leading to the exterior near the posterior end of the abdomen. In most species spermatozoa are transferred directly into the female reproductive tract and stored in a spermatheca (outpocketing of female tract) until fertilization occurs (as the egg is laid). In two families (Cimicidae, Anthocoridae) the female has an organ separate from the reproductive tract to receive the spermatozoa. This organ is a rounded internal pouch associated with a slit on the underside of the abdomen and is called the organ of Ribaga. During mating the spermatozoa are deposited in this pouch. They then penetrate the pouch wall, travel through the body cavity, and burrow into the spermatheca, remaining there until needed to fertilize the eggs. Excess spermatozoa are absorbed as nutrients by special cells in the female.

EVOLUTION AND PALEONTOLOGY

The fossil record is too poor to offer significant help in reconstructing the phylogeny of the Heteroptera; their small size, the fragile nature of dead individuals, and a preference for habitats that seldom are conducive to fossilization have resulted in few fossils. Often only wings or parts of wings are represented. Known fossils do reveal, however, that during the Permian Period (more than 225,000,000 years ago), two major lines existed. The order Protohemiptera, which revealed a number of features common to homopterans and heteropterans, including stylet-like mouthparts assumed to be piercing-sucking, actually represents an unrelated parallel development.

The appearance of the Heteroptera as a line distinct from the Homoptera also is represented in Permian fossils. The oldest known member of the order Heteroptera (from the Permian in Australia) represents the only family (Paraknightiidae) with no living species. Fossils from the next geological period, the Triassic, reveal that the major forking of the heteropterans into aquatic and terrestrial branches had already occurred near the beginning of the Mesozoic Era. Several modern aquatic family types are recognizable in the Mesozoic. The bulk of modern terrestrial heteropterans appears in the fossil record by the Tertiary Period of the Cenozoic Era.

CLASSIFICATION

Distinguishing taxonomic features. Modifications of mouthparts and wings are generally accepted as important characteristics by which to divide the class Insecta into orders. One peculiar type of mouthpart, possessed by more than 50,000 insect species, is the piercing-sucking type in which the labium forms a trough containing two pairs of threadlike stylets (highly modified mandibles and maxillae). All modern insects with mouthparts of this type are closely related to each other but not to other insects, none of which possess truly comparable modifications. Some entomologists consider all insects with piercing-sucking mouthparts as members of a single order, the Hemiptera (sometimes called Rhyngota or Rhynchota), with two suborders Homoptera and Heteroptera separated traditionally by texture and resting position of the forewings and by the apparent origin of the beak; other entomologists, while recognizing the proximity of relationship between these two groups, consider that the relationship is of superorder value and that features separating the two groups are of sufficient magnitude to warrant full ordinal status for each group.

Separation of the two groups is reflected by differences in several important functioning body parts. The gula of the Heteroptera is a hard bridge that separates the mouthparts from the prothorax. In the homopterans the gula is small and membranous or absent, allowing the base of the mouthparts to abut or fuse to the prothorax. The forewings of heteropterans are generally divided into basal and apical halves of different texture. Homopteran forewings, on the other hand, are generally of similar texture for their full length. Scent glands are present in all terrestrial and some aquatic nymphs and adults of the Heteroptera. Homopterans lack scent glands.

Reduction and loss of the gula in homopterans, coupled with the migration of the mouthparts to the prothorax, were significant changes in functional specialization. In some homopteran families of the suborder Sternorrhyncha, the mouthparts have fused to the prothorax and remain attached to it if the cranium is removed. These and other basic characters that began to differentiate in the geologically remote Permian time (when many other insect orders were evolving) indicate an important branching of the phylogenetic line of homopterans and heteropterans and substantiate their separation at the ordinal level.

The heteropterans are divided into suborders on the basis of habitat, antennal structure, ability to produce sound (stridulate), and presence and arrangement of long hairs (called trichobothria) on various parts of the body. Division into families is based on structural differences and adaptations for specialized ways of life.

Annotated classification.

ORDER HETEROPTERA (true bugs)

Wings, when present, usually number four that may vary in size; basal portion of forewings thick and leathery, apical portion membranous; hindwings membranous, slightly shorter than forewings; wings at rest held flat over abdomen; piercing-sucking mouthparts arise from front part of head; compound eyes usually well developed; two simple eyes, or ocelli, present or absent; scent glands usually present; gradual metamorphosis; widely distributed, terrestrial and aquatic species number about 30,000.

Suborder Hydrocorisae (or Cryptocerata)

Neither cephalic nor abdominal trichobothria; antennae 4-segmented, shorter than head, usually in grooves on underside of head; semiaquatic (Gelastocoridae, Ochteridae) or aquatic (all other families); swimming members with fringe of swimming hairs on hind legs; aquatic members lay eggs in or on submerged exposed objects, are relatively active under water during winter; probably all males, and some females, stridulate; important food for fish, some water birds.

Family Corixidae (water boatmen). Head overlaps front of prothorax; scutellum small, usually concealed by pronotum; front legs short, 1-segmented tarsus with fringe of long hairs beneath for gathering food; middle legs long, slender, for submerged anchoring; males with stridular pegs on front femurs and tibiae of legs, a polished black strigil above on asymmetrical abdomen; females nonstridulatory, abdomen symmetrical; colour pattern commonly alternating light and dark transverse stripes; inhabit fresh, brackish, or salt (Great Salt Lake) waters; use diving air stores; phytophagous on algae and diatoms, but

may eat insect larvae; more than 300 species; in all major faunal regions.

Family Notonectidae (back swimmers). Head not overlapping prothorax; scutellum large, exposed; all legs long and slender, hind pair longest; male stridulating mechanism involving front legs and labium; inhabits fresh and mineral waters; diving air stores trapped under hairs on underside of abdomen, hence inverted position in swimming; bite painful to man; predatory on insects, small crustaceans, fish, and tadpoles; almost 200 species; in all zoogeographical regions.

Family Pleidae. Very small, stout, convex above; scutellum large, exposed; hindwings reduced or absent; head and pronotum partially or wholly fused; inhabit fresh waters; about 20 species; in all zoogeographic regions.

Family Helotrephidae. Like Pleidae; about 15 species; only in Asia, Africa, and South America.

Family Nepidae (water scorpions). Body stout to very slender; legs long, generally slender, front pair adapted for grasping; abdomen of most nymphs and adults with snorkel-like breathing tube at tip; nymphs and adults stridulate by opposable files on anterior coxae and coxal cavity plates; inhabit fresh water; swim poorly or not at all; predatory; principal food, other insects; nearly 200 species; in all zoogeographical regions.

Family Belostomatidae (giant water bugs and toe biters). Large, some species exceeding 109 millimetres (4 inches) in length; broad, flattened body; middle and hind legs markedly flattened, with swimming hairs; inhabits fresh water; may lie in mud or debris on bottom; can inflict severe bites; predatory; aggressive, strong swimmers; grasping front legs capture insects, fishes, frogs; approximately 100 species; in New World, Asia, and Africa (including Madagascar).

Family Naucoridae (creeping water bugs). Head large, inserted into thorax; scutellum large, exposed; front femurs enlarged, tibiae opposable; middle and hind legs slender, sometimes with fringing hairs; inhabit fresh water, often warm springs; generally use diving air storage chambers or "plastron" respiration; predatory mostly on mollusks and insects; about 150 species; most common in tropical continents, ranging north into Europe and North America; subfamily Aphelocheirinae uses plastron respiration, sometimes elevated to family rank.

Family Gelastocoridae (toad bugs). Broadly oval in shape, roughened surface, elevated protuberant eyes; resemble tiny toads; beak reaching only to coxae of front legs; broadened front femurs and opposable tibiae form effective grasping organ; middle and hind legs slender for running, lack fringing hairs; inhabit shores, burrows into mud or sand to lay eggs; predatory principally on small insects; about 100 species; in all zoogeographic regions.

Family Ochteridae (velvety shore bugs). Similar to Gelastocoridae in general appearance; body surface smooth, with velvety pile; beak reaches abdomen; found among plants growing along margins of ponds and streams; predatory upon other insects; about 25 species; in all zoogeographic regions.

Suborder Amphibicorisae

Trichobothria on head but not on abdomen; members generally live on the water's surface film or on mud or sand; predatory on tiny animals, including injured or freshly-killed insects; nonstridulating except for both sexes of genus *Stridulovelia* (Veliidae); within one species individuals may vary from winged to wingless.

Family Gerridae (water striders; pond skaters). Slender bodied; legs long, slender, front pair short, used for support or grasping prey; middle pair longest, used for propulsion over water surface; hind femurs extends beyond tip of abdomen; tips of tarsi project beyond insertion of claws; body covered below with hydrofuge hairs; on quiet or flowing fresh or marine water; *Halobates* occurs on open oceans hundreds of miles from land; predatory generally on insects; oceanic forms feed on plankton and floating jellyfishes; about 350 species; in all zoogeographic regions and intervening oceans between 40° north and 40° south latitudes.

Family Veliidae (smaller water striders). Habits and structures similar to Gerridae, but generally smaller, stouter bodied, and with hind femurs not reaching beyond tip of abdomen; two genera with last segment of middle tarsus split and with collapsible, fan-shaped tuft of long hairs used for propulsion over water surface; about 300 species; in all zoographic regions, genus *Trochopus* marine, but not in open sea.

Family Hydrometridae (marsh treaders). Body, head, and legs long and slender; claws arise at tip of tarsus; wings present or absent; crawl weakly and clumsily among emergent plants in fresh water; predatory; more than 100 species; mostly tropical, a few species in North America and Eurasia.

Family Mesoveliidae (water treaders). Small; body and legs slender; live on fresh water, especially among emergent plants or on saturated mud; one New Guinea species lives on the forest floor away from water; about 30 species; in all zoogeographical regions.

Family Saldidae (shore bugs). Head large; eyes prominent; labium apparently 3-segmented, long, first segment very short and stout, second very long; legs long and slender for running or leaping; generally inhabit land surfaces near fresh or salt water; some species live on rocks in the intertidal zone, hide in air trapped in crevices when rocks are covered at high tide; predatory on small or crippled insects; over 200 species, in all zoogeographic regions.

Family Leptopodidae. Like Saldidae; all 24 species native to Old World, one Mediterranean species established in New World (California).

Family Hebridae (velvet water bugs). Tiny, covered with hydrofuge hairs suitable to habitat on surface or along edges of bodies of still, fresh water; sometimes crawl into the water; food habits not known; about 120 species; in all zoogeographic regions; some widely distributed species may have been carried in drinking water on early sailing ships.

Family Leotichiidae. Structure suggestive of Saldidae but distinguished by strong median ridge on pronotum; habits unlike Hebridae; one of the two known species (both from southeastern Asia) occurs in caves.

Suborder Geocorisae

Trichobothria either present on abdomen or absent from both abdomen and head; antennae longer than head; difficult suborder to define; perhaps a convenient ecological (rather than a taxonomic) category better replaced by elevation of the two series to suborders; terrestrial; habits and foods varied.

Series Cimicomorpha

Trichobothria absent from head and abdomen; front wings with or without a cuneus (special region on hardened basal portion known as the corium).

Family Miridae (plant bugs). Small, fragile; ocelli absent; labium 4-segmented; front wing with cuneus, membranous area with 1 or 2 closed cells; legs slender to stout, for walking, running, or jumping; predatory, phytophagous, or both; often host specific; includes major crop pests (*e.g.*, "lygus bugs" of the Northern Hemisphere, the "tea blight" [*Helopeltis*] in southeastern Asia), also several important predators of small insects; about 8,000 species; in all zoogeographic regions; found north of the Arctic Circle.

Family Cimicidae (bedbugs). Oval to elongate oval, very flat; eyes small; forewings reduced to small, transversely oval pads exposing most of abdomen; hindwings absent; bloodsucking parasites of mammals and birds, usually present on host only while feeding, otherwise hiding in crevices in nest or lair; eggs glued to walls of crevices; contains the two species of bedbugs that attack man, *Cimex lectularius,* (originally in Europe and North America, now in all parts of the world) and *Cimex hemipterus* (common in southern Asia and Africa); *C. lectularius* probably encountered cave-dwelling men; not yet conclusively proved as carriers of any human diseases even though the bug can be successfully infected with some human parasites; about 75 species; found throughout the world but show marked parallels with ranges of their hosts.

Family Reduviidae (assassin bugs). Head usually protruding; beak short, 3-segmented, strongly curved away from underside of head, with tip in deep stridular groove between front coxae; front legs adapted for grasping; prey on insects and other small animals; genus *Triatoma* and some close relatives suck blood of mammals (including man) and are carriers of the tropical American Chagas' disease; can inflict painful bites; nearly 4,000 species; on all significant land masses.

Family Anthocoridae (anthocorid bugs). Ocelli and cuneus present; membrane of wing with or without veins; male external genitalia asymmetrical; predatory, mostly on small arthropods, including their eggs; some eat ants; others inhabit bird nests; may bite man; about 400 species; in all zoogeographic regions.

Family Polyctenidae (bat bugs). Oval; flattened; eyeless; front forewings, rudimentary, unhinged; hindwings absent; head and certain other parts with transverse rows of spines forming ctenidia (combs); eggs hatch within female, young nymphs retained in body and nourished by mother through a pseudoplacenta; blood suckers live entire life in fur of certain bats; 18 species; found from southern North America to South America, in Africa, and from southern Asia to Java.

Family Aradidae (flat bugs). Extremely flat; surface usually rough; head horizontal; clypeus prominently convex; ocelli absent; legs short; tarsi two-segmented; abdomen much wider than wings when present; found in crevices, as under loose bark or debris on the ground; feed on molds or fungi; nearly

1,000 species in all zoogeographic regions; sometimes part of group separated as family Dysodiidae.

Family Tingidae (lace bugs). Head small, without ocelli; antennae slender to stout, generally with first two segments very short, third long; pronotum and forewings often finely to coarsely reticulate (lacy); tarsi two-segmented; phytophagous, on anything from mosses to flowering plants, often host-specific; several species are common pests on maize, cotton, cacao, and other cultivated plants; almost 2,000 species; in all zoogeographic regions.

Family Nabidae (nabid bugs). Ocelli present; beak four-segmented, long, generally curved away from underside of head; no striated groove between front coxae of legs; front legs more or less thickened for grasping; forewings without a cuneus; predatory on soft-bodied insects; frequently investigated as a possible biological control agent for insect pests of field crops; about 300 species; in all regions of the world.

Family Phymatidae (ambush bugs). Stout; surface roughened by tubercles; beak short, three-segmented, curved away from underside of head, tip resting in stridular groove between front coxae and legs; fourth antennal segment conspicuously thickened; front femurs of legs strongly thickened; scutellum small and triangular or large, broad, and reaching tip of abdomen; predatory, generally, lying among flower petals awaiting prey; more than 200 species; found on all major continents except Australia; not found on oceanic islands.

Family Enicocephalidae (unique-headed bugs). Predatory; about 130 species; in all zoogeographic regions.

Family Dipsocoridae (or *Cryptostemmatidae* or *Ceratocombidae*). Biology poorly known; about 200 species; in all zoogeographic regions, more than half of species found in tropical America.

Family Isometopidae. Phytophagous; less than 50 species; sometimes reduced to a subfamily in the Miridae.

Family Joppeicidae. The sole species occurs in the eastern Mediterranean region.

Family Microphysidae. Predatory; about 25 species; found in Eurasia and North America; one African genus questionably placed here.

Family Schizopteridae. About 100 species; sometimes united with the Dipsocoridae (see above).

Family Termitaphididae. Lives with termites; includes 9 species; found in tropical America, Africa, India, and Australia.

Family Thaumastocoridae (or *Thaumastotheriidae*). Phytophagous; about 12 species; found in Australia, India, North and South America, and the Greater Antilles.

Family Velocipedidae. Predatory; four species; in the Orient.

Family Vianaididae. Lives with ants (myrmecophilous); four species; found in tropical America.

Series Pentatomomorpha (or *Trichophora*)

Trichobothria on abdomen; cuneus always absent; terrestrial; various habits and foods.

Family Pentatomidae (stinkbugs). Head with side margins generally expanded over bases of antennae; ocelli usually present; antennae usually five-segmented; scutellum large to very large, triangular or U-shaped; without claval commissure; tibiae without rows of spines; phytophagous or predatory, sometimes within same species; several important pests—the "southern green stinkbug" (*Nezara viridula*), a serious pest on fruit and vegetable crops throughout the world; the "harlequin cabbage bug" (*Murgantia histrionica*), on North American cruciferous crops; the senn bugs (*Eurygaster* species), on small grains in the Old World—several genera (*Podisus, Zicrona*) feed on caterpillars and other larvae; about 5,000 species; in all zoogeographic regions; taxonomy unsettled, especially assignment of taxonomic levels; definition and relationships of subgroups not in dispute.

Family Plataspidae. Oval to transversely oval; head small, inserted deeply into prothorax; scutellum large, broadly U-shaped, reaching apex of abdomen; forewings much longer than body; wing membrane transversely folded at base and inserted under scutellum; tibiae not spined; phytophagous; about 500 species; found only in the Old World.

Family Cydnidae (burrower bugs). Scutellum triangular or broadly U-shaped; forewings without claval commissure (except genus *Amnestus*); tibiae with rows of spines; colour generally brown or black, sometimes with creamy markings; phytophagous, most species in soil feed on roots of host; more than 600 species; in all zoogeographic regions; subfamily Corimelaeninae (or Thyreocorinae) sometimes given family status.

Family Coreidae. Antennae four-segmented, inserted on or above imaginary line connecting middle of eye to front end of buccula (cheek plates); ocelli present; forewings with claval commissure; wing membrane with numerous interconnected veins; appendages often with leaflike expansions; phytophagous, some as pests of cultivated plants, for example, the "squash bug" (*Anasa tristis*) on cucurbit crops in North America, a rice bug (*Leptocorisa varicornis*) in the Orient; more than 2,000 species; in all zoogeographic regions; subfamilies Alydinae and Rhopalinae (or Corizinae) sometimes given family status.

Family Berytidae, or *Neididae* (stilt bugs). Body slender, narrowly elongate; legs extremely long, linear; labium four-segmented; usually phytophagous, sometimes attack soft-bodied plant-feeding insects; more than 100 species; in all zoogeographic regions.

Family Piesmatidae. Juga (two outer lobes of head) extend as a pair of horizontal, hornlike projections; prothorax with a deep pit below expanded sides; forewings, except apical part of membrane, with numerous coarse punctures; phytophagous, most frequently on plants of the goosefoot family (Chenopodiaceae); proven vectors of viral diseases of sugar beets in Europe and North America; about 20 species; in all zoogeographic regions.

Family Lygaeidae. Antennae four-segmented, inserted on side of head below an imaginary line connecting middle of eye to front end of buccula; ocelli present except in brachypterous forms; forewings with a claval commissure; wing membrane with four or five simple, longitudinal veins; phytophagous (on sap or seeds) or predatory (on mites and small insects); pests include the "chinch bug" (*Blissus leucopterus*) of North America, the Old World "cotton stainer" (*Oxycarenus hyalinipennis*), and several species of "false chinch bugs" in the widely distributed genus *Nysius;* in North America *Geocoris,* considered beneficial, feed upon mites and small insects attacking crop plants; almost 3,000 species found in all parts of the world, including many small islands.

Family Largidae. Head triangular, without ocelli; forewings with claval commissure; female with sixth visible abdominal segment below cleft medially; phytophagous; habits not well known; of no economic importance; about 100 species; in all zoogeographic regions.

Family Pyrrhocoridae (fire bugs or cotton stainers). Head triangular, without ocelli; forewings with long claval commissure; female with sixth visible abdominal segment below entire; phytophagous or predatory; "cotton stainers" (*Dysdercus*), serious pests of cotton crops; some species (*e.g.,* genus *Dindymus*) feed on flies as adults, nymphs prefer termites; more than 300 species; in all zoogeographic regions.

Family Aphylidae. Two species in Australia.

Family Colobathristidae. Phytophagous; about 90 species; inhabit Neotropical, Oriental, and Australian regions; unknown in North America, Europe, and Africa.

Family Hyocephalidae. One species in Australia.

Family Lestoniidae. Two species in Australia.

Family Phloeidae. Three species in tropical America, one in Borneo.

Family Stenocephalidae. About 40 species; except for one species on the Galapagos Islands, all known species live in Eurasia and Africa.

Family Urostylidae. Approximately 50 species confined to Asia and Australia.

Critical appraisal. The classification of the Heteroptera is in a state of flux due to continuous investigations in ontogeny, genetics, ecology, and morphology. Not all results are in agreement; therefore, important changes probably will occur. Wide disparity in superfamily concepts has resulted in dropping them from the above classification of the heteropterans.

The division of the order into three suborders is generally accepted, even if only for convenience; the prefixes reflect the broad ecological habits of the families included. Hydrocorisae reflects the adaptation of most of its families (Gelastocoridae and Ochteridae excepted) for life under water. Amphibicorisae suggests the land and water habits of these insects that live on the surface of the water or along shores. Geocorisae reflects the preferred terrestrial habitats of members of this group. (R.C.F.)

Neuropteran (lacewings, snakeflies, dobsonflies, alderflies)

At one time the term neuropteran referred to many groups of insects that are now placed in other orders, so that, in a strict sense, the order Neuroptera now includes only the

lacewings. Frequently, however, two other closely related insect groups are included also as neuropterans. These are the snakeflies (Raphidiodea), so-called for their body shape, and the dobsonflies and alderflies (Megaloptera). For completeness of discussion, therefore, all three groups, considered here as three separate orders, are included in this section.

GENERAL FEATURES

All three orders may have evolved from an early mecopteran (scorpionfly) ancestral stem, prior to the Trichoptera–Lepidoptera offshoot (see section on Trichoptera). Carnivorous insects of varied structure and habit, both freshwater and terrestrial members of the three orders, are widely distributed with the exception of the snakeflies (which are confined to the Northern Hemisphere). Many members are important in the biological control of other insects and mite pests; others are "flies" of interest to anglers. Some of the most elegant and dainty insects are lacewings. There are more than 500 species of alderflies and dobsonflies, 80 species of snakeflies, and 4,000 species of lacewings.

From *Invertebrate Identification Manual* by Richard A. Pimentel, © 1967 by Litton Educational Publishing, Inc.; reprinted by permission of Van Nostrand Reinhold Company

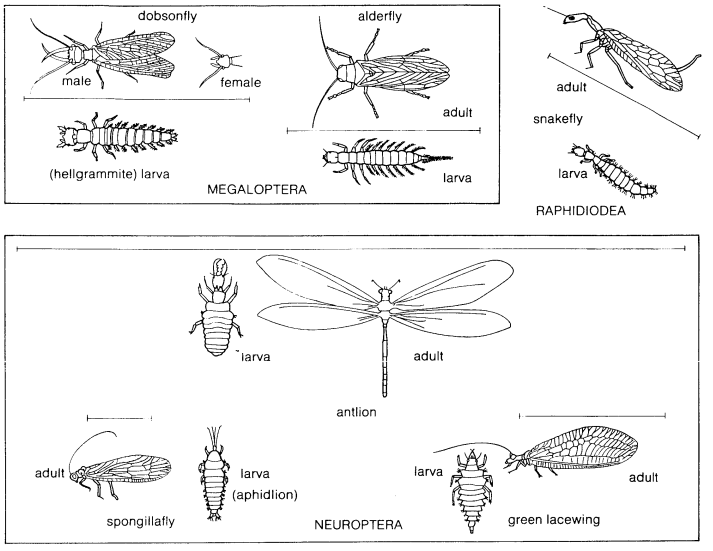


Figure 30: Diversity among neuropterans. Line scales indicate approximate size of each insect.

The size of lacewings varies from 1.5 to more than 35 millimetres (0.059 to more than 1.377 inches) in length and 2 to more than 50 millimetres in anterior wing length. Lacewings are characterized by their many-veined pairs of wings, which give them a lacy appearance. Dobsonflies and alderflies, while similar in general appearance, have wings that appear heavier than those of lacewings; 15 to 30 millimetres in body length, their anterior wing length varies from 20 to more than 50 millimetres. Snakeflies are of medium size, 10 millimetres or more in body and anterior wing length, with wings similar to those of lacewings.

The Megaloptera and Raphidiodea differ from the Neuroptera in their prognathous adult heads (mouthparts anterior) and biting larval mandibles; the neuropterans have hypognathous adult heads (mouthparts ventral) and unique piercing–sucking larval jaws (composed of mandibles plus maxillae).

NATURAL HISTORY

Life cycle of Neuroptera. Neuropteran eggs may be laid loosely in light soil, cemented directly to a surface, or cemented on the end of a stalk. During the cementing process, the female places the tip of the abdomen on a surface and exudes a viscid fluid; she raises her abdomen slowly to draw the fluid into a slender filament. The fluid rapidly hardens, and the egg is attached by its posterior end to the top of the stalk.

Larvae hatch after 5–14 days unless the egg is in the overwintering stage. In some families a thickened portion of skin is used by the larva to break the egg; in others the egg splits. There are generally three larval stages in Neuroptera. The first lasts a few days, the second a few days or for the winter months, while the third varies, depending on the species, from weeks to months.

Neuropteran larvae are carnivorous and free-living with the exception of the aquatic family Sisyridae; these larvae are parasitic on freshwater sponges. Typically, a neuropteran larva sucks out the contents of its prey, leaving only a hollow skin. Although many lacewing larvae are nocturnal and need no camouflage, other species carry debris on bodies adapted for this purpose. In one family debris floats onto the hairs and is caught, while in another the larva takes debris in its jaws and places it on its back. In still another family larvae lie lightly covered in the soil; the lacewing larvae known as antlions dig conical pits in light, dry soil or sand. They vigorously throw out soil with their heads, then lie in the pit, body covered and jaws ready to grasp an ant or any other prey that may fall in. Should a captive attempt to escape by climbing a pit wall, the antlion will use its head to toss soil and cause the prey to fall back into the pit. Larval feeding

The neuropteran larva spins a double cocoon by exuding whitish or yellowish silk through its anus. First, a loosely woven cocoon is spun and fastened to a surface; then the larva spins a second tightly woven cocoon inside the first. This double construction is typical of neuropterans; the walls of the two cocoons may be closely attached or apart, depending on the species. The larva may spend the prepupal stage of several days or months within the cocoon before the pupal molt occurs. The limbs of the pupa are free. In a few species the pupa bursts from the cocoon, but most species use their functional mandibles to emerge. Adults appear either as the pupa leaves the cocoon or after it has reached a suitable position. Some species have two or more broods a year, although the life cycle does not exceed 12 months. Mating may occur anytime during the lifespan of the adult female; sometimes she holds fertilized eggs until weather conditions are suitable.

Life cycles of Megaloptera and Raphidiodea. Megalopteran females lay eggs above water level in masses of 3,000 or more. Larvae crawl into the water where they are very active. Commonly, alderfly larvae are associated with muddy bottoms of ponds and slow-moving streams, dobsonfly larvae with fast-flowing streams or rivers. The larvae are predaceous, usually nocturnal, and may leave the water to search for prey or to molt. Before pupation the larvae leave the water to form cells in damp, coarse soil beneath stones or debris. The pupa crawls from its earthen cell before the adult emerges. Raphidiodean females use slender ovipositors to lay eggs in the cracks of tree bark, and immature stages are found there.

Usually mating occurs at night. Sperm are passed either directly as spermatozoa or in a spermatophore that may project from the female after mating and be eaten by her, Reproduction

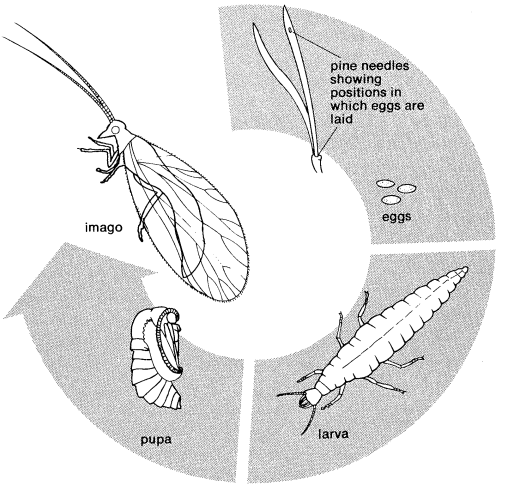


Figure 31: Life cycle of the brown lacewing (*Hemerobius stigma*).

either wholly or in part. While megalopterans lay eggs in masses of several thousands, raphidiodeans lay them singly, and neuropterans lay 600 or 700 singly, separately in groups, or in batches.

Ecology. All of the larvae and adults are carnivorous and predaceous (except the parasitic Sisyridae), devouring enormous numbers of mites, insect eggs and larvae, ants, thysanopterans, psocopterans, and homopterans. For example, chrysopids consume large numbers of aphids and are a potent natural control mechanism. Larval feeding is continuous except during molts. Adults may also feed on dead insects, and females may feed on their own eggs and on spermatophores. The neuropteran family Sisyridae may have evolved from the closely related family Osmylidae whose larvae probe streamside mosses for prey.

Megalopteran larvae, a source of food for freshwater fish, are used as bait by anglers. Spiders capture both adults and terrestrial larvae in their webs. Neuropteran eggs and larvae are preyed upon by other neuropterans, other predaceous insects (*e.g.,* coccinellid, or ladybird beetle, larvae), ants, birds, and bats.

FORM AND FUNCTION

Adults. Antennae are multisegmented in neuropterans and filiform in Megaloptera and Raphidiodea. The head is squarish or transverse in Megaloptera and Neuroptera, elongated in Raphidiodea; prognathous (mouthparts anterior) heads are typical of Megaloptera and Raphidiodea, hypognathous (mouthparts ventral) of Neuroptera. The four wings, which are similar in size, large or elongated, and translucent, contain many veins and crossveins and are sometimes coloured in brown or green shades. The ten-segmented abdomen lacks appendages (cerci); however, terminal claspers may occur. In both sexes, the last abdominal segment may be reduced, modified, or bear a pair of dorso-lateral groups of trichobothria (short, stiff sensory bristles). In females, abdominal terminalia are also variable; Raphidiodea have a long, thin ovipositor.

Larval jaws

Immature stages. The larval head, prognathous in all three groups, may bear as many as seven simple eyes on each side or none. Megalopteran larvae have large mandibles, raphidiodean small, while larval jaws, the outstanding feature of neuropterans, may be short or long, straight or curved. In all neuropteran species each jaw is a sucking tube, mandible above and maxilla below, the two fitting together but separable. Since there is no true mouth, liquid food is sucked by a pharyngeal pump through the two jaws directly into the pharynx. The aquatic megalopteran larvae have lateral gill filaments and either a median caudal filament or anal prolegs; Raphidiodea lack abdominal processes. The terminal abdominal segments of neuropterans are modified into an anal proleg, with or without processes, which also functions as a spinneret (an organ for producing silk from internal glands); in several families the abdomen has special hairs or bristles (macrotrichia) for holding debris for camouflage.

Some pupae resemble adults, particularly the snakeflies. The hooks or ridges, aligned both forward and backward and evident on the dorsum of the abdomen of some pupae, may be used to maintain a position within the cocoon rather than to escape from it; mandibles and limbs are often used to aid in the escape.

Locomotion. Most larvae have walking legs except in those whose legs are modified for burrowing (Ithonidae and Myrmeleontoidea). The aquatic dobsonfly larvae have a pair of terminal abdominal prolegs that are important in moving in strong currents. The legs, abdominal filaments, and median tail filament of alderfly larvae bear long hairs, or setae, which help to propel them through their slow-moving water habitats. Snakefly larvae wriggle rapidly backwards to escape danger. Most neuropteran larvae have the tenth abdominal segment modified into a single anal foot for use in locomotion; a pair of hooked processes also assist in holding. Members of the three groups typically are weak fliers.

EVOLUTION AND PALEONTOLOGY

The closely related dobsonflies, alderflies, and snakeflies may have arisen from a common stem that diverged from the ancestral mecopteran (scorpionflies) stem early in the Carboniferous Period (about 325,000,000 to 345,000,000 years ago); the Neuroptera may have diverged from the ancestral mecopteran stem later in the Carboniferous.

Fossil representatives of the modern families Sialidae and Corydalidae are represented in the Oligocene Baltic amber (about 26,000,000 to 38,000,000 years old). The present-day Raphidiidae are represented in the Oligocene Baltic amber. The present-day neuropteran family Psychopsidae was existent as early as the Triassic Period (about 190,000,000 to 225,000,000 years ago) in Australia. The present-day neuropteran families Coniopterygidae, Osmylidae, Sisyridae, Hemerobiidae, Chrysopidae, and Myrmeleontidae first occurred in the Oligocene Baltic amber. The Ascalaphidae first appeared in the Lower Miocene of France (about 36,000,000 years ago), and the Nemopteridae first occurred in the Miocene of Colorado.

CLASSIFICATION

Distinguishing taxonomic features. The adults of the orders Megaloptera, Raphidiodea, and Neuroptera have similar mouthparts, wings, and genitalia; they differ in shape of head and thorax. The presence of simple eyes (ocelli) and the morphology of antennae, head, wings (including venation), legs, and body, are criteria used to distinguish families.

Larval characters associated with the aquatic habit, such as abdominal filaments, separate Megaloptera and Sisyridae (Neuroptera) from terrestrial Neuroptera and arboreal Raphidiodea. The normal mandibles of Megaloptera and Raphidiodea separate these groups from Neuroptera, whose jaws are combined mandibles and maxillae. The form of the head, neck, mouthparts, legs, and body are used to separate neuropteran families.

Annotated classification. This is a combination of American and English classifications accepting three separate orders.

ORDER MEGALOPTERA

Adults medium to large; head prognathous; biting mouthparts; antennae threadlike (or filiform); 2 pairs of large, similar wings, held rooflike or nearly flat over abdomen in repose. Larvae elongated; biting mandibles; lateral abdominal gill filaments present; aquatic.

Family Sialidae (alderflies)

Wing expanse 20–40 mm; simple eyes (ocelli) absent; 4th tarsal joint of lower leg dilated and bilobed. Larvae with 7 pairs of lateral, segmented, abdominal filaments, median caudal filament; larvae without anal prolegs.

Family Corydalidae (dobsonflies)

Wing expanse 40–100 mm; 3 ocelli; 4th tarsal joint cylindrical; male mandibles sometimes enlarged. Larvae (called hellgrammites) with 8 pairs of unsegmented or imperfectly segmented, lateral, abdominal filaments; terminal filament absent; 2 anal prolegs.

ORDER RAPHIDIODEA (snakeflies)

Adults small to medium; head prognathous, elongated, tapering posteriorly; biting mouthparts; antennae filiform; prothorax elongated, cylindrical; 2 pairs of similar elongated wings, held rooflike over abdomen in repose; female with long, slender ovipositor. Larvae elongated; head and prothorax large, polished; biting mandibles; abdominal filaments and prolegs absent; arboreal.

Family Raphidiidae

Three ocelli; antennal joints smaller at bases.

Family Inocellidae

Ocelli absent; antennal joints cylindrical.

ORDER NEUROPTERA (lacewings)

Adults small to large; head hypognathous; biting mouthparts; antennae varied; 2 pairs of similar wings, held rooflike over abdomen in repose. Larvae elongated or broad; piercing–sucking jaws, elongated mandibles above, elongated maxillae below, form a hollow tube; abdominal filaments absent, except aquatic Sisyridae: paired abdominal prolegs absent; 10th abdominal segment modified as an anal proleg and/or spinneret.

Superfamily Ithonoidea

Adults large, mothlike; larvae with straight jaws, modified for subterranean predatory feeding.

Family Ithonidae (moth lacewings)

Adult wing expanse 30–79 mm; head small, closely set on

prothorax; antennae long, filiform; costal areas of wings not broad. Larvae burrowing; legs adapted for digging, tibia fused to tarsus; mouthparts short, straight; maxillae enlarged.

Superfamily Hemeroboidea

Adults small to large; antennae filiform or moniliform (beadlike); posterior wings shorter than anterior; elongated larvae with jaws straight or curved.

Family Hemerobiidae (brown lacewings)

Adults small; antennae long, moniliform; ocelli absent; wings with few crossveins, hairy; female without projecting ovipositor. Larvae (called aphis wolves) elongated, with short incurved jaws; body smooth, fine hairs; free living on vegetation.

Family Sympherobiidae (brown lacewings)

Adults small, closely related to Hemerobiidae. Larvae spindle-shaped; head small; antennae short; jaws short, stout, slightly incurved.

Family Dilaridae (pleasing lacewings)

Small; antennae of male coarsely comb-shaped; head with 3 ocelli or ocellus-like tubercles; wings with numerous crossveins.

Family Psychopsidae (silky lacewings)

Adults large mothlike species; antennae short, wings broad. Larvae elongated, flat; head broad posteriorly, closely attached to prothorax; jaws incurved, large, sicklelike; often arboreal, under bark.

Family Osmylidae (osmyliidflies)

Adults medium to large; head wider than long; antennae filiform, short; 3 ocelli; wings often with brown markings; claws with many teeth. Larvae elongated; long, slender, straight jaws, slightly upcurved; in margins of freshwater.

Family Polystoechotidae (large lacewings)

Adults medium to large; wing expanse 40–75 mm; antennae short. Larvae with short, sharp, incurved mandibles, maxillae stout, blunt; labial palpi, sensory appendages on labium (lower lip); leg 5-jointed; tarsal claws simple, slightly curved; knobbed structures (called empodia) between terminal elongated claws.

Family Sisyridae (spongillaflies)

Adults small, brownish, closely related to Osmylidae; ocelli absent. Larvae moderately elongated; hairy; antennae long, bristle-like; mandibles and maxillae elongated, with bristle-like piercing structures; labial palpi absent; 1 tarsal claw per leg; aquatic; parasitic on freshwater sponges.

Family Chrysopidae (green lacewings, stinkflies)

Adults small to medium, yellowish, green, or gray, sometimes with markings; wing expanse 31–65 mm; head small, antennae long, filiform; iridescent golden eyes. Larvae (called aphid lions) moderately elongated, broad; long, sharp, incurved jaws; hairy; legs with long, trumpet-shaped empodium between claws; many with dorsal lumps on body, carry packets of debris dorsally; found on vegetation.

Family Apochrysidae (fragile lacewings)

Fragile, closely related to Chrysopidae; antennae filiform.

Family Berothidae (beaded lacewings)

Small, slender; antennae long, filiform; wings hairy, sometimes with seedlike scales; larvae thysanuriform, with short, straight jaws.

Family Trichomatidae

Small, slender, closely related to Berothidae; wings and body with long hairs.

Family Mantispidae (mantispids)

Adults small to large; wing expanse 10–55 mm; antennae short; prothorax greatly elongated; wings narrow; forelegs predatory. Larvae have 2 forms; 1st stage with squarish head; antennae with 3–5 joints; short, straight jaws; labial palpi with 3 or 7 joints; legs with 2 claws and empodium; active hunters; in presence of food reserves larvae become white, parasitic grubs with small heads and atrophied legs.

Superfamily Nemopteroidea

Adults highly specialized, related to Myrmeleontoidea; head prolonged in rostrum; posterior wings narrowed, elongated to twice anterior wing length. Larvae short, broad; with incurved jaws; varied in form, distinct from other neuropterans.

Family Nemopteridae (thread-winged or spoon-winged lacewings)

Adults delicate; head snoutlike; antennae short; posterior wings greatly elongated, ribbonlike or threadlike; often expanded distally to appear spoonlike. Larval antennae long, filiform; jaws incurved; mandibles with or without internal teeth; with or without an elongated neck formed by anterior portion of prothorax; if present, neck may be thin, cylindrical, 3 times length of head; remainder of thorax and abdomen broad, flattened; legs long, with 2 curved claws, empodium absent; free-living in dust, some cave dwelling.

Superfamily Myrmeleontoidea

Adults medium to large in size; antennae varied. Larvae short, broad; jaws incurved; mandibles with inner teeth.

Family Myiodactylidae

Adults resemble Osmylidae; antennae long, cylindrical.

Family Nymphidae (slender lacewings)

Adults slender; long, cylindrical antennae; wings long, slender. Larval neck slender; jaws incurved, mandibles with single internal tooth near midway.

Family Myrmeleontidae (antlions)

Adults with long, slender wings; bodies partly covered with fine hairs; dragonfly-like; antennae short, weakly clubbed or flattened distally. Larvae (called antlions or doodle bugs) short and broad; head small; neck bent to allow rapid upward and backward movement of head; mandibles strong, incurved; 3 internal teeth; posterior legs modified for backward burrowing, tibia fused to tarsus; abdomen hairy, plump, modified posteriorly for backward burrowing; lie in soil surface or at bottom of a self-made pit.

Family Stilbopterygidae (shiny wings)

Adults large, swift fliers; shining wings and bodies; dragonfly-like; antennae short, strongly clubbed; eyes large. Larvae large, black, shiny; found in vegetable refuse.

Family Ascalaphidae (owlflies)

Adults large, dragonfly-like species, 40–50 mm in length; some rapid fliers; antennae long, slender, strongly clavate distally; eyes large, divided; head with long, fine hairs. Larval head squarish; mandibles elongated; 1 or several inner teeth; neck short, narrow; posterior legs with tibia fused to tarsus; body fringed.

Superfamily Coniopterygoidea

Tiny, somewhat separated from other Neuroptera; body, wings covered with white waxy powder; larvae straight-jawed; labrum large, covering the other mouthparts.

Family Coniopterygidae (dustywings)

Adult wing expanse 3–10 mm; antennae long, filiform; wings similar or posterior wings reduced; wing venation reduced. Larvae, 5 simple eyes each side; 5-jointed antennae; jaws short or long, needlelike; legs with padlike empodium.

Critical appraisal. Standard English and American systems differ in their classification of Megaloptera and Raphidiodea as distinct orders.

The standard English classification is order Neuroptera, suborder Megaloptera, superfamily Sialoidea, family Corydalidae, family Sialidae, superfamily Raphidiodea, family Raphidiidae, and suborder Planipennia (true Neuroptera). In other systems, the suborder Megaloptera is termed suborder Sialodea, the combination Raphidioptera is used for Raphidiodea, and the name Neuroptera is used for a superorder containing the orders Megaloptera, Raphidioptera, Planipennia. The modern tendency is to separate the three main groups in the separate orders Megaloptera, Raphidiodea (or Raphidioptera), and Neuroptera (Planipennia), as in the classification above.

Although larvae of most families have been described, further breeding and detailed studies of life histories and morphology of immature and adult forms are required.

(K.A.J.W.)

Trichoptera (caddisflies)

Caddisflies are insects that comprise the order Trichoptera. The mothlike adult, attracted to light at night, often lives near lakes or rivers. Because fish feed on immature, aquatic stages and trout take flying adults, caddisflies are used as models for fishermen's "flies."

GENERAL FEATURES

Adult caddisflies are commonly 3 to 15 millimetres (0.118 to 0.590 inches) in length; their anterior wings range from 4 to 20 millimetres in length, providing wing spans of 8 to 40 millimetres. The wings at rest are folded rooflike over the body. One family (Hydroptilidae), commonly known as microcaddis, are only 1.5 millimetres in length, with anterior wings of 2 to 5 millimetres. Caddisfly wings are either covered with hairs or have hairs on the veins; the posterior pair are often broader than the anterior.

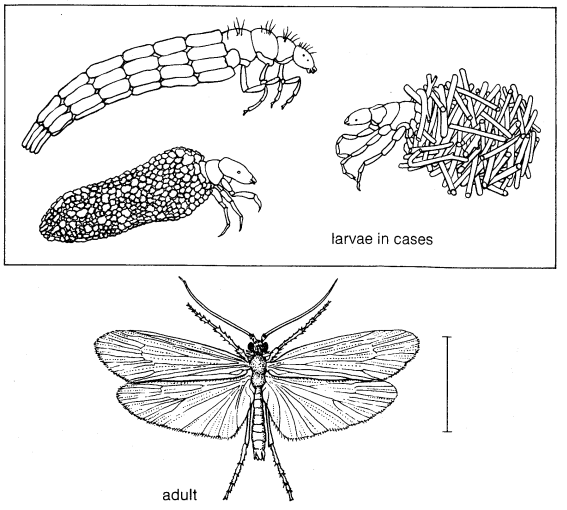


Figure 32: General structure of caddisflies, larvae and adult; line indicates approximate size of adult.

(Top) From *Invertebrate Identification Manual* by Richard A. Pimental, © 1967 by Litton Educational Publishing, Inc., reprinted by permission of Van Nostrand Reinhold Company; (bottom) after McLachlan in A.D. Imms, *A General Textbook of Entomology*, 9th ed. (1970).

Caddisflies are widely distributed in freshwater habitats throughout the world. The larva of one marine species, *Philanisus plebeius,* occurs in intertidal zones of New Zealand and southern Australian coasts. The larva of one terrestrial European species, *Enoicyla pusilla,* lives in damp leaf litter, while one or two other species occur in brackish water. Approximately 7,000 species of caddisflies are known.

NATURAL HISTORY

Life cycle. Caddisflies undergo complete metamorphosis. Males generally have complicated external genitalia. Spermatozoa are transferred either directly or in spermatophores (capsules transferred to females). In the females of some species the terminal segments of the abdomen are modified to form a long ovipositor. Eggs, in masses numbering up to 800, are laid within a jelly that swells on contact with water. A female may wash off a partially extruded egg mass by dipping her abdomen into water during flight, or she may place the mass on stones in the water or on aquatic plants just above the water.

Young larvae, crawlers with strong legs, hatch within a few days and feed on vegetable material or algae. In some species the larvae form webs of debris for protection; others form a funnel-like web between stones in running water to catch food. Some protect their bodies with cases; others spin protective lairs or are free-living. Larvae may be herbivorous, carnivorous, or omnivorous. They produce silk from glands on the lower lip (labium), and many herbivorous species spin tubular protective cases, open at both ends, that enlarge as the larvae grow; sand grains or vegetable debris, added to cases, provide protection and rigidity. In casebearing forms the head and thorax protrude from the case, which is pulled along by the abdomen. Larvae of one species (*Triaenodes bicolor*) swim by means of long, slender, hair-fringed posterior legs, dragging their cases through the water. Undulating movements of the enclosed larvae circulate water to provide oxygen for respiration.

Larval cases

Prior to the last molt before the pupal stage, the larva fixes the case to some solid object and closes both ends for further protection, leaving only water circulation holes. The larva pupates inside the larval case (which then becomes a cocoon) or inside a specially constructed cocoon. After two or three weeks the pupa bites its way out of the cocoon and swims or crawls to the water surface, using its hair-fringed middle pair of legs; then the winged caddisfly adult emerges and lives several weeks. Adults sometimes emerge in large numbers, often forming swarms. Caddisflies usually mate on vegetation or rocks surrounding water. There is generally one complete life cycle per year.

Ecology. Essentially, caddisflies are aquatic insects associated with a wide range of freshwater habitats. Oxygen concentration, associated with water velocity, is important to their larvae, as is the chemical content of the water. In fact, certain species have been used as indicators of pollution. Larvae feed on aquatic plants, algae, diatoms, or plant debris; a few are predatory on other aquatic insects, crustaceans, and mollusks; some are omnivorous. The larvae, often numerous, play an important role in the aquatic community, reducing plant growth and disposing of animal and plant debris; they use vegetation and debris in casemaking and for food. Most adult caddisflies, incapable of feeding on solids, imbibe liquids (*e.g.,* nectar from flowers); one species pollinates an Alpine flower as it feeds.

Sponges and algae grow on the protective cases, and protozoans and mites grow on the larvae of some species. Two hymenopterans parasitize caddisflies; the parasitic wasp seeks its host underwater and lays its eggs inside the caddisfly larval case. The parasitic larva devours the caddisfly larva and remains inside the case to pupate; the adult parasite overwinters in the case and emerges in the early summer. Caddisflies are important as food for other animals. Freshwater fish, particularly trout, and eels feed on larvae and swimming pupae; trout, birds, lizards, frogs, spiders, dragonflies, and bats feed on adults. Day flight of caddisflies includes stationary swarming, usually centring around some obvious object on the shore, or mobile swarming, in which the swarms seem to move haphazardly. Most species fly at night, however, and are strongly attracted to light. Night flight may be bimodal, continuous, or intermittent. Warm moist nights are particularly conducive to caddisfly flight.

FORM AND FUNCTION

Trichopteran adults are characterized by four wings that bear hairs on the membrane or prominent veins, or both. The head and thorax are also usually hairy. The many-jointed antennae, usually about the length of the anterior wing, may be longer than the wingspan. Mouthparts have nonfunctional mandibles but well-developed maxillary and labial palpi (or sensory appendages); the hypopharynx is modified to form a sucking tongue for imbibing fluids. Various sensory and scent organs may occur on the head or wings. Each of the three thoracic segments bears a pair of walking legs. Several of the ten abdominal segments may bear lateral filaments (or cerci), often with complicated external genitalia terminally in males, sometimes with prominent ovipositors in females.

Mouthparts

EVOLUTION AND PALEONTOLOGY

The caddisflies were long classified in the order Neuroptera; the two groups are now thought, however, to represent different evolutionary lines. Ancestral Mecoptera (scorpionflies) probably gave rise to the Neuroptera (lacewings), Trichoptera (caddisflies), and Lepidoptera (moths, butterflies). The closely related Trichoptera and Lepidoptera may have arisen from a common offshoot. Early aquatic trichopterans may have diverged from the terrestrial line in the Late Triassic Period (about 200,000,000 years ago). The earliest known trichopteran fossils are from the Early Jurassic Period (about 185,000,000 years ago). A few fossil wing prints from the Cretaceous Period (about 65,000,000 to 136,000,000 years ago) are known; many fossils in Oligocene Baltic amber (about 26,000,000 to 38,000,000 years ago) belong to or are related to living genera. Later fossil wing prints occur in Miocene deposits.

CLASSIFICATION

Distinguishing taxonomic features. Caddisflies are mothlike insects with two pairs of hairy wings (sometimes a few scales) folded rooflike over the abdomen. Antennae are long and many jointed with a large basal joint. Large compound eyes are present; simple eyes (ocelli) number three or are absent. The number of maxillary palpal joints usually is five but varies between two and six. Posterior wings are broader than anterior; the pattern of veins is generalized and sometimes reduced. Legs are long and slender, tarsi five-jointed; tibial leg segments may bear variable numbers of spurs. Male genitalia, often complex, have been used for specific identification; female genitalia are of increasing taxonomic significance. Another system

relies on the internal characters (*e.g.*, internal head skeletal structures, known collectively as the tentorium). Larvae are essentially aquatic. Campodeiform (elongated and flattened) larvae are free-living or net forming; eruciform (caterpillar-like) larvae are casebearing. In some groups head capsules, functional mandibles, thoracic shields, abdominal gills, abdominal prolegs or hooks, and types of nets or cases are distinct; in all groups antennae are short, and spiracles absent. Pupae have functional mandibles, free limbs, and hook-bearing dorsal plates.

Annotated classification. The number of living families has recently increased. Nineteen were listed in 1953 and 34 (based on a different method of classification) in 1967. The following is based on the later scheme.

ORDER TRICHOPTERA (caddisflies)

Insects with 2 pairs of wings held rooflike over the body in repose; the wing membrane or veins with hairs, or both; wing vein pattern generalized; tibiae usually with spurs, tarsi 5-jointed; larvae and pupae aquatic, with functional mandibles; larvae with terminal abdominal hooked prolegs or hooks, free-living, net spinning, or casemaking, pupate in larval case or cocoon.

Suborder Annulipalpia

Terminal segment of maxillary palpi annulate (with ringed appearance) or subdivided; larvae, net spinners.

Superfamily Hydropsychoidea. Antennae usually anterior wing length or less.

Family Philopotamidae. Ocelli medium to small; larval head elongated.

Family Stenopsychidae. Closely related to Philopotamidae; ocelli large or absent; larval head elongated.

Family Xiphocentrionidae. Closely related to Psychomyiidae; ocelli absent.

Family Psychomyiidae. Ocelli absent.

Family Polycentropodidae. Ocelli absent.

Family Dipseudopsidae. Closely related to Polycentropodidae.

Family Hydropsychidae. Antennae mostly about anterior wing length or slightly longer; ocelli absent; larvae with many-branched ventral abdominal gills on most segments.

Suborder Integripalpia

Terminal segment of maxillary palpi undivided; larvae free-living or case makers.

Superfamily Rhyacophiloidea. Antennae shorter than anterior wing length.

Family Rhyacophiloidae. Ocelli present; larvae free-living, produce a single line of silk.

Family Glossosomatidae. Adults similar to Rhyacophilidae; larvae, saddle case makers with modified terminal prolegs and hooks.

Family Hydroptilidae (microcaddis). Ocelli present or absent; wings reduced, narrow, with long hair fringes; larvae purse case makers or tube case makers in last larval stage; earlier stages free-living.

Superfamily Limnephiloidea. Larvae, tube case makers.

Limnephilid Branch

Antennae about anterior wing length; ocelli usually present; adult mandibles reduced or lost.

Family Limnocentropodidae. Primitive articulating mandibles.

Family Brachycentridae. Maxillary palpi 3-jointed in male.

Family Phryganeidae. Maxillary palpi usually 4-jointed in male.

Family Phryganopsychidae. Male maxillary palpi 4-jointed.

Family Goeridae. Ocelli absent; maxillary palpi of male 3-jointed.

Family Thremnidae. Closely related to Limnephilidae.

Family Limnephilidae. Maxillary palpi of male 3-jointed.

Family Plectrotarsidae. Closely related to Limnephilidae; specialized elongated mouthparts.

Family Lepidostomatidae. Ocelli absent; maxillary palpi 3-jointed in male.

Family Rhynchopsychidae. Closely related to Lepidostomatidae, but mouthparts elongated.

Leptocerid Branch

Antennae about anterior wing length or less, except Leptoceridae; ocelli absent.

Family Pisuliidae. Male maxillary palpi 3-jointed.

Family Calocidae. Maxillary palpi 5-jointed.

Family Pycnocentrellidae. Maxillary palpi 5-jointed.

Family Beraeidae. Maxillary palpi 5-jointed.

Family Sericostomatidae. Male maxillary palpi 3-jointed or less.

Family Helicophidae. Maxillary palpi 5-jointed, both sexes.

Family Philanisidae. Maxillary palpi 5-jointed in both sexes; larvae only in marine littoral zone.

Family Antipodoeciidae. Apparently related to Odontoceridae; male maxillary palpi 3-jointed.

Family Odontoceridae. Maxillary palpi 5-jointed.

Family Molannidae. Maxillary palpi 5-jointed; claws specialized.

Family Philorheithridae. Maxillary palpi of males 5-, 4- and 3-jointed, basal joint with nodule.

Family Helicopsychidae. Maxillary palpi of male 2-jointed; larval cases look like small snail shells; very small in size, traditionally placed in Sericostomatidae; many adult characters reduced or absent; relationships uncertain.

Family Calamoceratidae. Maxillary palpi 5- or 6-jointed.

Family Leptoceridae. Antennae longer than wingspan; maxillary palpi 5-jointed; anterior wings elongated; vein pattern reduced.

Critical appraisal. Many species of caddisflies have been described from male characteristics, particularly genitalia only, family characters being assumed. Work in recent years on female characters, internal adult characters, and larval characters has revealed new sources of taxonomic information. Further progress in classification depends upon these detailed studies. The classification used above reveals the extent of progress thus far. (K.A.J.W.)

From H. Weber, *Grundriss der Insektenkunde* (1966); Gustav Fischer Verlag, Stuttgart

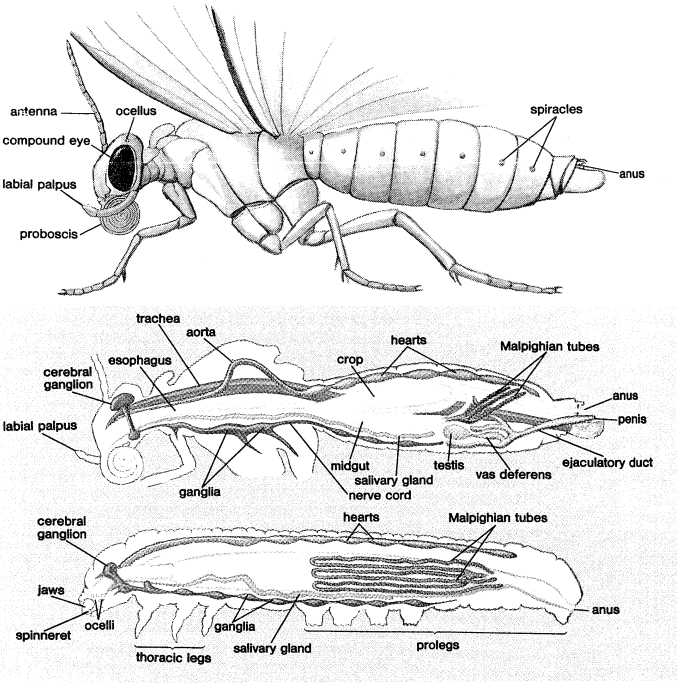


Figure 33: Body plan of generalized male adult butterfly or moth and caterpillar.

Lepidoptera (butterflies, moths, skippers)

The insect order Lepidoptera, with as many as 100,000 species, is second in size only to the Coleoptera, the beetles. In addition to the familiar butterflies and moths the order includes the skippers, a worldwide group intermediate between butterflies and moths. Because of their dayflying habits and bright colours, the butterflies are more familiar than the chiefly night-flying and dull coloured moths, but the latter are far more varied and abundant. With the exception of a few moths, all adult Lepidoptera have two pairs of wings. The name Lepidoptera is derived from the Greek, meaning "scaly winged," and refers to the characteristic covering of microscopic, dustlike scales on the wings.

The life cycle of Lepidoptera is that common to the higher orders of insects and consists of four stages: egg, larva (caterpillar), pupa (chrysalis), and adult (imago). The larva and adult are almost invariably plant eaters. The larvae do most of the eating, the majority of them feeding

on foliage, although many species eat stems, roots, fruits, or flowers. A number of moths and a few butterflies are important pests in agriculture or forestry, invariably as larvae. The adults of many species are important pollenizers, visiting flowers for nectar. In the ecology of most land environments the Lepidoptera are important as transformers of immeasurably large amounts of plant matter into animal matter, which then serves as food for many other groups of animals.

Many members of the order, especially the butterflies, have appealed to the human imagination for thousands of years, as symbols of fragile and ephemeral beauty. References to them abound in literature; and they have been depicted in many paintings, have inspired the designs of jewelry, ornaments and textiles, and even occur in many heraldic devices and on postage stamps. They are the most popular of insects among amateur collectors and students of natural history.

GENERAL FEATURES

The members of the order show great diversity in size and development rates. The smallest moths have wing expanses as small as 4 millimetres, while the largest moths and butterflies may expand to nearly 300 millimetres (about 1 foot). Fast developing species may complete their development in as little as three weeks, while slower ones may take as much as two, or even three, years.

Both moths and butterflies occur on every continent except Antarctica. Though they are enormously more numerous and diversified in the tropics, at least some species occur at the limits of polar vegetation. There are many successful species in nearly every land environment, from arid deserts and high mountaintops to marshes and tropical rain forests. Most of the species have become adapted for life in relatively restricted ecological niches, and may be limited to only one or a small group of food plants, often to only a single part of one plant. They are therefore seldom abundant in more than one type of habitat. A few species, however, with broader food habits, may occasionally reach peaks of abnormal abundance and defoliate large areas of such environments as deciduous forest or grassland. Some of the insects most injurious to man's interests in agriculture, forestry, and households are members of the Lepidoptera.

Geographic distribution

The large family-level groups, and many smaller ones, are worldwide in distribution, with such families as Noctuidae, Geometridae, Pyralididae, and Nymphalidae being dominant elements of the insect fauna of every geographic region. A few families, some subfamilies and a great many genera, however, are characteristically more abundant in one faunal region than in others. Thus, the Neotropical region (mainly Central and South America) is characterized by great diversity of such moth families as the Dioptidae, Pericopidae, and Ctenuchidae and such butterfly subfamilies as the Riodininae, Heliconiinae, Ithomiinae, Morphinae, and Dismorphiinae. The North American (Nearctic) and Eurasian (Palaearctic) faunas show many evidences of close recent connections, chiefly between Asia and western North America. Each region has, however, many distinctive generic-level groups; for example, in the Palaearctic, *Erebia* and other satyr butterflies, and the parnassians (Papilionidae); in the Nearctic, the underwing moths *Catocala* (Noctuidae) and the silverspot fritillaries (Nymphalidae); in the Ethiopian (African) region, the butterfly genera *Charaxes* (Nymphalidae) and *Teracolus* (Pieridae), and many members of the papilionid subfamily Acraeinae; in the Oriental region, *Troides* (Papilionidae) and *Delias* (Pieridae).

Many Lepidoptera occur in isolated colonies as relict (remnant) populations, cut off from relatives elsewhere by geologic or climatic changes. Australia and New Zealand have unusually diverse relict populations of the primitive Micropterigidae and Hepialidae. In North America, Europe, and Asia many relict species have survived since the Pleistocene glaciations on isolated southern mountaintops. They include members of such genera as *Oeneis* and *Erebia* (Satyridae), *Boloria* (Nymphalidae), *Parnassius* (Papilionidae), *Anarta* (Noctuidae), and *Pediasia* (Pyralididae).

IMPORTANCE

Many hundreds of Lepidoptera injure plants useful to man, including some of his most important sources of food, fabrics, fodder, and timber. The great majority of the injurious species are moths, and the most injurious life stage always is the larva. Unlike members of other insect orders, the Lepidoptera do not act as carriers of plant diseases, nor are any of them parasites or predators of man or domestic animals.

The list of valuable plants subject to damage by Lepidoptera is a long one, including many grains, sugar beets and cane, cotton, tobacco, some root crops and leaf crops, many fruits, and timber and shade trees. The damage may involve the leaves, stems, roots, or fruit. Woollens, furs, silk, and even feathers are eaten by tineid moths of several genera (clothes moths). The wax moth (*Galleria mellonella*) causes considerable damage in beehives.

A few Lepidoptera are directly beneficial to man. Nearly all silk is obtained from the completely domesticated silkworm, *Bombyx mori,* originally Asiatic. Such other silks as shantung and tussah are the products of various Asiatic Saturniidae. The larvae and sometimes the adults of a few species are used for food. The larvae of a skipper, *Rhopalocampta libeon,* are collected in large quantities in the Congo, and the caterpillars of giant skippers (Megathymidae), known in Mexico as *gusanos de maguey,* are canned and exported for consumption as hors d'oeuvres.

Lepidoptera beneficial to man

The South American cactus moth, *Cactoblastis cactorum,* has been highly beneficial in weed control, clearing more than 60 million acres of Australia of prickly pear cactus. Doubtless man benefits also from much unrecognized weed eating by caterpillars and flower pollination by adults.

Many Lepidoptera are valuable in biological research, including work in ecology, biogeography, systematics, genetics, and physiology. Much of the present knowledge of endocrine controls of insect development has come from studies of the silkworm moth (*Bombyx mori*) and various saturniid moths. The study of industrial melanism (an increase in the frequency of black individuals in a population, brought on by the environmental changes that accompanied the Industrial Revolution) in the British peppered moth (*Biston betularia*) has profoundly influenced modern ideas about rates of evolutionary change.

NATURAL HISTORY

Feeding habits. The habits of Lepidoptera are extremely diverse, depending on the adaptations of the species or group to climate, environment, type of food plant, way of feeding, and many other factors. The great majority of food plants are seed plants (Spermatophyta), the dominant land plants. Primitive plants such as mosses, liverworts and ferns, and some lichens are eaten by a few groups. Nearly all parts of the plant are eaten by various specially adapted forms. Flowers are eaten by many larvae (*e.g.,* lycaenid butterflies, plume moths, some geometrid moths), and nectar by many adults. Cones and fruits and their seeds are eaten by others (*e.g.,* prodoxid and olethreutid moths, some owlet moths). Some seed eaters have become household pests, feeding on stored grains and cereals. Buds or soft, succulent stems are bored into by members of many families. Several lepidopteran groups (*e.g.,* the pine moths, *Rhyacionia*) specialize on the terminal shoots of conifers. Many groups feed on the turf of grasses and sedges. Three families (Cossidae, Aegeriidae, and Hepialidae) bore in woody stems and rootstocks, the Cossidae in particular tunnelling deep into hardwoods. A good many lepidopterans (especially members of the Tineidae, Blastobasidae, and Pyralididae) feed on dead and decaying plant matter, particularly moldy debris. In comparison with other orders of insects, relatively few Lepidoptera live in plant galls or eat animal matter. Some tineids feed on fur and feather debris, and a few exceptional species are predators or parasites of other animals.

Types of food taken

Life cycle. *Egg.* The number of eggs laid varies greatly from fewer than a hundred in some species to more than a thousand in others. The eggs are almost always laid in a specific way, usually on or in a suitable foodplant. Those of many species are laid singly and widely dispersed; in

others they are laid in masses that may be covered with a hardened secretion of the female's abdominal glands. In a few groups the terminal segments of the female's abdomen are greatly elongated and bladelike, and the eggs are laid in soft plant tissues or in narrow slits or crevices. In some of the primitive hepialid moths and in some butterflies the female may merely scatter the eggs in flight in the general vicinity of a suitable foodplant. Development of the embryo and emergence of the young larva is often controlled by a mechanism of physiologically enforced inactivity (diapause), which has the effect of timing the emergence of the larva to coincide with suitable conditions of weather and the growth of the foodplant. Respiration in the egg is carried on through an aeropyle, a system of air passages in the chorion (shell) that enables oxygen exchange with the environment to occur whether the egg is dry or wet. In a few species (some tineid moths and pierid butterflies) the larvae hatch in the uterus of the female.

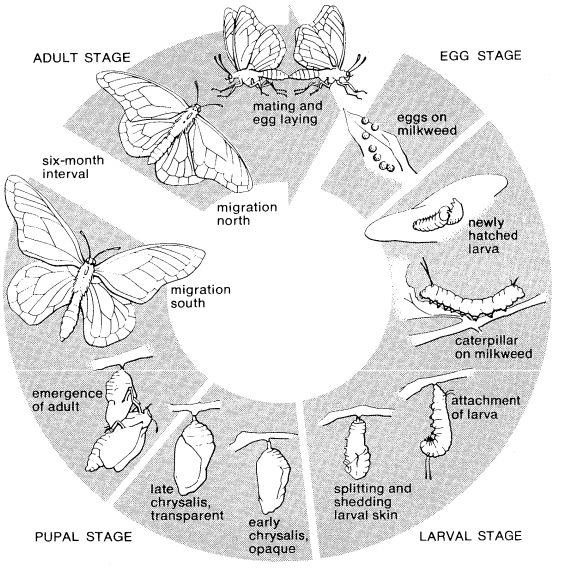


Figure 34: Life cycle of the monarch butterfly (*Danaus plexippus*).

Larva. The larva is the chief, often the only, nutritive stage of the life cycle, its function being simply to transform very large quantities of plant matter into animal matter, and to stay alive during the process. The larval stage is followed by the pupa, a resting stage in which the individual undergoes a major rebuilding of body tissues, to emerge as a sexually active adult. Plant foods being relatively easy to find and eat, it is not surprising that lepidopterous larvae are fundamentally quite uniform, despite their apparent diversity. Most larvae continually spin silk threads that adhere to the surfaces on which they live, thus giving them an almost unshakable grip. Silk is also used in the construction of various nests, cases, and shelters that protect the larvae from the elements and from enemies. Silk also serves in the final cocoon, in which the larva will change into the pupa. The larvae of leaf miners are mostly very small, greatly flattened, and more or less legless, well adapted for life inside a flat leaf. The larvae of borers are relatively plain and unornamented, while those that live in the open, exposed to the attacks of predators, have evolved a multiplicity of protective adaptations of form, colour, texture, and habit. Some possess chemical repellants. The proper timing of the larval development and activities is important in species in which hibernation, or estivation, is necessary. In some species (*e.g.,* nymphalid butterflies of the genus *Limenitis*) the larvae of the summer generation grow right on through without pause, but those of the next year's generation hibernate while still very young. In certain groups larval development is very slow, for reasons not understood, so that there may be only one generation a year. The larvae of the carpenter moths (Cossidae), feeding on poorly nutritious wood, may take two years to develop. Larvae of many Arctic and northern species, regularly have two-year cycles. The usual number of molts (ecdyses) is four or five, but some of the small leaf miners molt only twice. When starved, clothes moth larvae (*Tineola*) have been known to have a dozen molts, sometimes accompanied by a decrease in size.

Production of silk

Larvae seldom travel far from the plants where they began life. In some species there is dispersal of very young larvae, which, hanging from threads, are blown by the wind. Swarms of army worms (*Pseudaletia,* Noctuidae) may travel large distances, driven by crowding and lack of food. Just before pupation many larvae stop eating and crawl some distance before settling down to pupate. Many species pupate in the soil, with little or no cocoon; many others form cocoons in the soil, in trash, or under loose bark. Some cocoons are fastened to twigs or branches or rolled in leaves. The cocoons of leaf miners are usually formed in the "mine" or near it. Cocoons are commonly of silk alone but may also incorporate larval hairs, secretions or waste products, or chewed wood pulp, trash, or leaves. There may be a special seam on the cocoon to facilitate the emergence of the adult moth. Skipper larvae make a flimsy cocoon in the larval nest. Of the true butterflies only the satyr and parnassian butterflies make cocoons; all others pupate naked, with the pupa (chrysalis) hanging from a silk pad by a stalk (cremaster). The chrysalids of some butterflies (pierids, swallowtails, and lycaenids) are additionally supported by a threadlike silk girdle about the body.

Construction of the cocoon

Pupa. The duration of the pupal stage differs greatly in various groups and sometimes in different generations of a single species. It may last only a week or ten days, in small, fast developing species in summer, or for many months, in cases of hibernation or estivation. Pupae have been known to remain alive for three years in abnormal conditions, the adults finally emerging successfully. The danger of desiccation is greatest in small pupae and exarate ones (those in which the appendages are not fixed to the body by a skin or sheath) and least in large or obtect (compact) ones. Most obtect pupae can move only one or two abdominal segments. Those of borers can often wriggle along the tunnel to its exit, aided by rows of backward pointing abdominal spines that give purchase on the walls of the cocoon.

The adult. Although fully formed as an adult, the individual may remain quiescent within the pupal case for a long time, until conditions are right for its emergence. In some groups, the pupa has cocoon cutters, such as the movable mandibles of certain primitive moths, or sawtoothed structures on the head (leaf blotch miner moths). In saturniid moths the adult has bladelike structures on the wing bases. In puss moths (*Cerura*) and others the cocoon is partially dissolved by alkaline secretions.

Once out of the cocoon, the adult crawls upward to where it can hang head up or back down. Small pads on the thorax contain the incompletely developed wings in a compressed state. The insect then forces its body fluids into the thorax by contracting its abdomen and pumps blood out into the small wing pads, expanding them to form the full wings. The adult may be able to fly in a few minutes or may have to hang for several hours until its wings have stiffened sufficiently for flight.

As the larva is the nutritive stage of the life cycle, the adult is the reproductive stage. Its mobility is necessary to bring the males and females together for copulation and to disperse the species into new areas. Nutrition is also an essential adult function in many primitive groups, such as the Micropterigidae. In the majority of lepidopterans it is important, but not essential, although in most of the highly mobile species much of the energy necessary for flight is obtained from nectar or other liquid foods taken in by the adult. The most advanced condition in this respect is found in the groups in which the adult mouthparts have become so reduced that they cannot function; here all the nutrition is unquestionably concentrated in the larval stage.

Behaviour. *Food selection by the adult.* Adults locate their food sources by both sight and scent, the former being especially important in the diurnal species, the latter in many diurnal species and in the nocturnal ones. The

chief source is floral nectar, but sap (especially if fermenting), overripe fruits, homopteran honeydew, fecal matter, and carrion are sometimes used, usually by special groups. There are mutualistic relationships of a broad sort between the species of Lepidoptera with flower-visiting adults and the plants whose flowers they visit and pollenate, but these are seldom specific or obligate, since only rarely are the plant and the lepidopteran mutually dependent. Exceptions are some members of the orchid and morning glory families with very deep, tubular flowers. These appear to be pollinated only by certain hawkmoths with very long proboscises ("tongues"). The mutualism of the yucca moths (Prodoxidae) and yucca plants is obligate; the moth larvae feed only in yucca fruits and the latter can develop only from moth-pollinated flowers. The female yucca moth has special tentacles on the first maxillae, with which it gathers and carries balls of pollen.

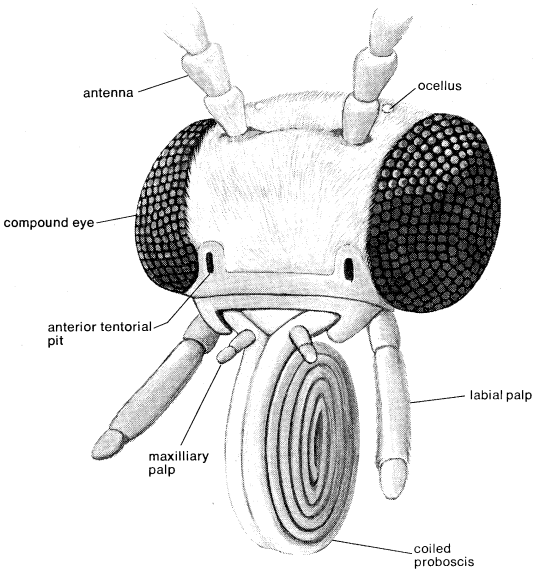


Figure 35: Head and mouthparts of Lepidoptera.
From A. Imms, *Textbook of General Entomology*; Methuen & Co. Ltd.

Use of odours in courtship

Courtship and mating. The sequence of mating activities is usually initiated by the female, who gives off specific odorous substances (pheromones), sometimes even before she emerges from the pupa. These are detected by scent sensilla on the male's antennae. Males with very large, feathery antennae, such as saturniid moths, can locate females from some miles away and may form courting swarms about them. A species may have a particular time of day or night for this release of the female pheromones. Males of some species, such as the European ghost moth, *Hepialus humuli,* may at dusk form dancing swarms, into which the females fly.

Once the male has located the female, a more or less elaborate courtship may ensue. The sight of characteristic colours and patterns may be a requisite for this, as in the bright butterflies. Male pheromones also play an essential part. Distributed from special scent scales (androconia) on the wings, body or legs, the pheromones ensure the receptivity of the female. Finally the accessory genitalic structures must fit together, not only mechanically, but also so as to stimulate sensory nerve organs of the female. Some butterfly courtships are very active (*e.g.,* in sulphurs) the male and female flying high in the air, with elaborate behavioral sequences in which each act serves as a releaser, a stimulus for the next act. Such highly specific sequences are important barriers to hybridization between otherwise similar species.

The males of many species show definite territorial behaviour, defending a particular perch and area against the intrusion of other males, and darting out at passing females. Among butterflies, certain admirals (*Limenitis*), coppers (*Lycaena*), and hackberry butterflies (*Asterocampa*) are noted for this. Congregations of males of many butterflies about hilltops may have a certain territorial and premating function. So may the aggregations, sometimes very large, of many butterflies in wet places. In many instances these consist of young, unmated males that come together because of a mutual attraction, not merely to sip water.

Migration. Many Lepidoptera are famous migrants. The American monarch butterfly (*Danaus plexippus*) is the only species known to perform an annual two-way migration, the same individuals flying southward in the autumn and northward in the spring. The species has also crossed the Pacific Ocean, colonizing Hawaii and Australia, and occasionally reaches Africa and Europe. The cosmopolitan painted lady (*Vanessa cardui*) stages mass flights nearly everywhere it occurs; these are one-way flights with no returns and must be classed as emigrations, rather than true migrations. Many other species in Europe and North America fly northward in large numbers, often reaching regions in which they cannot survive the winters. In Europe the painted lady and many moths reach Britain and Scandinavia from central Europe. In North America the painted lady, the cloudless sulphur butterfly (*Phoebis eubule*), and many noctuid moths often reach Canada. Many spectacular migrations occur in the tropics, in which swarms numbering many millions may fly out to sea and be lost. The best known species are pierid butterflies, but large migrations of the large diurnal moths *Urania leila* and *U. fulgens* have also been recorded. The usual explanation of such mass population movements—*i.e.,* the advantage of possibly extending the range of the species—is far from adequate in many such instances.

Ecology. *Environmental hazards.* As primary consumers of green plant material, the Lepidoptera are enormously important in food chains, not only because of the very large number of species in the order and the diversity of their food habits, but also because of the abundance of individuals. In most land environments they, in turn, are eaten by a host of other animals: predators, parasites, and scavengers. All stages in their life cycles are thus under continual attack.

The major invertebrate predators on Lepidoptera include centipedes, spiders, mantids, bugs, ground beetles, ants, and both social and solitary wasps. Important predators among vertebrates include toads and tree frogs, lizards, birds, rodents, bats, and monkeys. Some of the invertebrates locate their prey by scent, others by sight; most of the vertebrates hunt by sight, except for the bats, which hunt by acoustic echolocation (the so-called bat "sonar").

Predators and parasites

The chief groups of parasites that attack Lepidoptera are tachinid flies (Tachinidae) and many families of wasps, chiefly of the Ichneumonoidea, Chalcidoidea, and Cynipoidea. More precisely called parasitoid predators, they are probably more important than the direct predators. The female parasitoids locate suitable hosts, chiefly by scent, and lay their eggs in, on, or near them. The parasitoid larvae live inside their hosts, gradually feeding on the host tissues and almost invariably consuming them almost completely. Lepidoptera seem to have evolved few defenses against parasitoids, unless some of their toxic or repellent secretions serve to discourage them. The high reproductive rate is important as a counter against loss to parasitoids, as well as against other adverse factors.

Small, red trombiculid mites often ride about on adult Lepidoptera, probably doing them no harm. A few other mites, however—*e.g., Myrmonyssus*—live and breed in the tympanic cavities of noctuid moths, destroying their auditory structures. Curiously, they regularly settle in only one of a moth's two tympanic cavities and thus only half deafen it. It is believed that, by leaving the moth with one good ear, the mite reduces the likelihood of the moth (and hence the mite itself) being captured and eaten by a bat. Lepidoptera are subject to attack by a considerable number of protozoa, roundworms, bacteria, viruses, and fungi that affect the larvae chiefly during peaks of abundance and crowding. Some of these organisms have been used by man as a means of controlling injurious species.

Protection against danger. It is chiefly against the sight-hunting predators that the Lepidoptera have evolved a multiplicity of defense mechanisms that are unequalled by those of any other group of animals.

The adults of many groups, such as skippers, many butterflies, hawk moths and many underwing moths, have fast erratic flight. When escaping they dart or fall to cover and often remain immobile for some time. They have a good chance of escaping, especially if they are coloured like their surroundings. Larvae, especially when small, drop suddenly when disturbed, either dangling from a silk thread or falling to concealment on the ground. The larvae of some noctuids can jump several inches. Dense, loose hairs and scales make many moths slippery and may facilitate escape from sticky spider webs.

Certain noctuid moths and possibly some geometrids and pyralids, subject to predation by bats, are able to receive and identify the navigational sound pulses of the bats. Upon hearing bat pulses, these moths perform violent evasive flight movements (when the bat sound is loud, hence close) or dive to the ground (when the bat pulse is weaker, indicating that the danger is further away).

Targets, such as prominent colour spots or tails on the hindwings, attract the attention and focus the attacks of predators onto parts of the prey less vulnerable to injury. Such spots are likely to be seized and torn off, but this does the moth or butterfly no real harm and gives it time to escape without vital injury.

Concealment and camouflage

Many species manage to hide very effectively from predators. Many cutworms (Noctuidae) and other larvae hide in litter by day, feeding only at night. Many moths hide in crevices, often under loose bark, and some seem to have especially flattened bodies for this purpose. Hibernating butterflies spend the winter in hollow trees or hanging immobile among dead leaves. The larvae of a great many moths, of most skippers and of many butterflies, live in individual nests of rolled, folded, or webbed leaves or grass. Sod webworms live in silk-lined tunnels in turf. Wood borers, especially those in rootstocks and deep tunnels, are relatively secure. Many larvae aggregate in communal nests such as those of the tent caterpillars (*Malacosoma*), the larvae of the ermine moth (*Yponomeuta*), and those of the Mexican social white butterfly (*Eucheira socialis*). The larvae of the bagworms (Psychidae) and casebearers (Coleophoridae) live and pupate in individual, portable cases that are often masked with bits of leaf or twig. Some larvae, such as those of the green geometrid moths (*Synchlora*), attach bits of leaves or flower petals to themselves. Cocoons are frequently masked with leaves or debris; those of the puss moths and some dagger moths (*Apatela*) are hard, woody, and inconspicuous.

A cryptic (camouflaged) appearance conceals something only in appropriate surroundings, and only when accompanied by proper behaviour, usually immobility. Great numbers of larval, pupal, and adult Lepidoptera are thus protected in their natural environments, chiefly on or about plants. Leaf-eating larvae usually blend into leafy environments. Many caterpillars (*e.g.,* Sphingidae, Geometridae) have stripes that simulate leaf veins. The sawtoothed elm caterpillar (*Nerice*) has a jagged outline resembling the edge of an elm leaf. A great many measuring worms (Geometridae) are notably twiglike, with long, slender, stiffly held bodies. Many other caterpillars, especially Notodontidae, have irregular shapes that resemble twisted dead leaves. Likewise, many adult moths that rest during the day among leaves or on bark are cryptically coloured and patterned; moreover, their behavioral mechanisms lead them to select matching backgrounds on which to rest.

Many larvae and adults are disruptively marked; *i.e.,* have bold, contrasting patches or bands of colour that "break up" their outlines into two or more seemingly unrelated masses. Many adult moths and butterflies have "flash coloration," which serves to startle, and thus to delay momentarily, an attacker. Moths with cryptic forewings, and butterflies with cryptic undersides, show only these surfaces when they are at rest. If, when disturbed, they suddenly expose brilliantly coloured and marked hindwings, such as in underwing moths (*Catocala*), or upper surfaces (for example, morpho, hairstreak, and anglewing butterflies), the effect is startling. When the animal lands, the bright surfaces are suddenly hidden, causing it to "disappear." A similar "startle" effect protects larvae that have prominent spots simulating large eyes.

Startling sounds are produced by members of many groups. Larvae of hawkmoths (Sphingidae) and the pupae of many lycaenid butterflies make squeaking or grating sounds when disturbed. The adult death's head moth (*Acherontia atropos*) makes a loud chirping sound. *Ageronia* butterflies (Nymphalidae), when startled into flight, make a loud, clicking sound by means of a structure on the wings. This may have a startling, and therefore a delaying, effect on a predator.

Chemical defenses

Certain butterflies and moths possess repellent or toxic substances that provide protection against predators. Sometimes these are secured directly from the plant on which the larva feeds, such as the toxins (glycosides) that occur in high concentration in many milkweeds eaten by danaid butterflies. More often, the toxin is secreted by the insect itself. The toxin often occurs generally in the blood (*e.g.,* hydrogen cyanide in zygaenid moths) or in the gut, or may be the product of special glands, which release it at the time of an attack. Tiger moths (Arctiidae) give off bubbling drops of repellent from glands on the prothorax. Many groups show autohemorrhization, or reflex bleeding, from leg and body joints when disturbed. The larvae of swallowtail butterflies (*Papilio*) and tussock moths (Liparidae) give off strong smelling, volatile substances from extrusible scent organs (osmeteria). Many notodontid moth larvae spray formic acid from ventral prothoracic glands. Many larvae and some adults introduce toxins that cause severely painful nettling effects by means of hollow, barbed hairs; the slug caterpillar moths (Eucleidae), flannel moths (Megalopygidae), io moths (Saturniidae), and some liparid moths are noted for this. A few adult moths (*e.g., Arctia caja,* some Sphingidae) inject toxins through sharp spines on their hind legs. Besides those already mentioned, the chief groups that are chemically protected are the moth families Ctenuchidae, Pericopidae, and Dioptidae and the butterfly groups Heliconiinae, Ithomiinae, Acraeinae, and the swallowtails (*Battus*) that feed on birthwort. The great majority of these and other protected forms are aposematic; *i.e.,* have markings, shapes, and behaviour that draw attention. They are thus easily recognized and remembered by predators which, after seizing only one or two individuals, will thereafter leave other similar ones alone.

The protective advantage gained by a distasteful or dangerous insect and accompanied by aposematic coloration or acoustic warning may be utilized by harmless and edible insects, through the evolution of close resemblances, known as mimicry. (For background information of this phenomenon, see MIMICRY.) The distasteful insect, known as the model, may even be a member of an insect order different from that of the mimic. Members of various lepidopteran families mimic wasps, bees, and beetles. The clearwing moths (Aegeriidae) are particularly effective mimics of certain stinging wasps, the resemblance being carried to details of the shape and coloration of the wings, abdomen, and legs. Mimicry occurs widely in the moth families Dioptidae, Chalcosiidae, Callidulidae, and Zygaenidae and in the butterfly families Nymphalidae, Papilionidae, Pieridae, and Riodinidae.

Not all warning mechanisms are visual. Inedible arctiid and pericopid moths make high-pitched grating sounds by means of timbal (drumlike) organs. These sounds, above the frequency range audible to man, but within that audible to bats, may function to warn the bat of the moth's inedibility, avoiding capture and tasting that would be injurious to the moth. Some authorities believe that these sounds may interfere with the bats' acoustic orientation system, preventing the detection of the moth.

Polymorphism

The occurrence in a population of two or more distinct hereditary variants, or morphs, is known in many Lepidoptera. Each morph may have a different adaptive value, linked with such physiological features as resistance to cold, or to toxins in the environment. Striking variation in appearance may have great adaptive value by confusing predators, making it more difficult to learn the appearance of the prey. Mixed populations of both light and dark (melanic) individuals may survive better in mixed light and dark environments, and actually be prepared for a shift of the environment to a predominance of dark, through the normal succession of forest growth or through

man-made industrial pollution. In England investigations of the peppered moth *Biston betularia* have abundantly documented the evolution of "industrial" and "natural" melanism and shown major genetic population changes taking place very rapidly.

Striking polymorphisms occur in some mimetic species, notably the African swallowtail, *Papilio dardanus.* The occurrence of different species of inedible butterflies ("models") in various regions has been accompanied by the evolution of correspondingly different mimetic females of the single species of swallowtail. The North American tiger swallowtail, *P. glaucus,* has mostly black females where it coexists with the black distasteful pipevine swallowtail, *Battus philenor.* Where *B. philenor* does not occur, however, the *P. glaucus* females tend to be all nonmimetic yellow like the males, black having no protective significance. Some very striking mimetic polymorphisms occur among Neotropical *Heliconius* butterflies and their various models and mimics.

FORM AND FUNCTION

Within the limitations imposed by their almost entirely plant-eating habits, the Lepidoptera show a great diversity of size, structure, and many other characteristics. Furthermore, some members of the order have retained many primitive features, while others have evolved very advanced and specialized ones. This broad range, from primitive to advanced and from generalized to specialized, is evident not only in the adults but also in the larvae and pupae, and to some degree in the eggs.

The adult. Nearly all external surfaces of the adult animal are covered with scales, which may be broad and flat or very long and hairlike. Each of these is the outgrowth of a single epidermal cell. Similar scales occur in a few groups of other insects, but never to the same extent.

Like most insects, adult lepidopterans have three distinct body segments, or tagmata, the head, the thorax, and the abdomen, each with special functions (Figure 33). The head bears the main sensory organs and those of food-getting and ingestion. The thorax is chiefly concerned with locomotion. The abdomen contains the chief organs of digestion, excretion, and reproduction and bears the external accessory reproductive structures.

Head. The head is relatively small and round or elliptical. Evolutionarily, it is derived from the six anterior somites or metameres (the primitive body segments), but these have become so coalesced that none of the primitive segmentation is evident. The antennae are prominent and many segmented, with many microscopic receptors for detecting odours. In most moths the antennae are slender and tapering; in a few moths the surface area of the antennae is greatly enlarged by many side branches that make them comblike or featherlike. In skippers and butterflies the terminal part is broadened, forming a "club," on which most of the sensilla are concentrated. In many families there is also a cluster of sensory bristles, the chaetosema, on each side of the head, near the eye. On either side of the head is a large compound eye, sometimes consisting of thousands of units (ommatidia). Most moths have, in addition to the compound eyes, a pair of very small, simple eyes (ocelli), which have a very limited function.

Sense organs

The compound eyes are very efficient for distinguishing motion, but not for resolving clear images of distant objects. Diurnal species may be able to distinguish flower shapes, and courting individuals respond to specific wing patterns. Colour perception, especially in the blue-violet end of the spectrum, is acute in some species. Moths are generally able to see by ultraviolet, to which they are often strongly attracted at night.

The sense of smell is acute in some groups, especially in males with large antennal surfaces. The sense of taste is also acute, especially for sugars, through receptors located chiefly on the palpi and on the soles of the tarsi (feet).

Hearing is acute in the moths, with special membranes (tympana) most of which are located in the thorax, some of which react to sound waves above the range of human hearing. Mechanical information, including sound, is also received by chordotonal (stretch) receptors in a number of locations, especially in the antennae.

A relatively primitive set of chewing mouthparts occurs in the adults of the most primitive moths (Micropterigidae) as well as with larvae of nearly all Lepidoptera. Basically these consist of an anterior flap, the labrum; a pair of chewing jaws, the mandibles; a pair of complex first maxillae; and a pair of similar second maxillae joined together behind the mouth to form a structure called the labium. Each of the first and second maxillae bears a jointed, sensory appendage, or palpus, the maxillary and labial palpi respectively. Such mouthparts function for chewing and manipulating solid foods. In the vast majority of adults the mandibles are either vestigial and nonfunctional or absent. Parts of the first maxillae (the galeae) are elongated to form the two halves of a tubular proboscis, or haustellum, through which liquids may be sucked. The segmented palpi of the first and second maxillae are present and function as sensory organs. Not all adults, however, have all of these parts fully formed and functional. In numerous families the proboscis has become considerably reduced and even vestigial, so that the adult cannot feed. In advanced moths and in skippers and butterflies, the maxillary palpi are vestigial or lost, so that only the labial palpi remain functional.

Mouthpart reductions appear to have occurred independently in a number of evolutionary lines, the result being to limit feeding to the larval stage and to enable the adult to concentrate on reproduction and dispersal.

Thorax. The thorax consists of three segments, the prothorax, mesothorax, and metathorax, each derived from a primitive segment (metamere). The prothorax bears the first pair of legs and a pair of respiratory apertures (spiracles). The much larger mesothorax bears the second pair of legs and the pair of forewings. The metathorax bears the third pair of legs and the pair of hindwings. In many moths the metathorax bears a pair of complex auditory organs (tympana), and in many of these it also bears a pair of sound-producing timbal organs.

The wings begin development in the maturing larva as sacs infolding (invaginating) from the epidermis. As the pupa is formed these outfold (evaginate) to lie externally and become large and flat. Within them, branching tubes (tracheae) carry an air supply and also mark the pathways along which will develop the tubular "veins" that support the fully formed wings. The wings are expanded to full size when the adult emerges from the pupa. On the wing the scales lie in rows, overlapping like shingles. The usual scale is a flattened, rigid, air-filled sac (Figure 36) attached by a peglike base. It is usually ribbed longitudinally and toothed terminally. Many males have special, glandular scent scales (androconia) scattered or concentrated in patches (brands or stigmata) on the wings, sometimes forming expansible, hairlike tufts. These have a scent-distributing function essential in specialized courtship. Very primitive moths have the wing membrane aculeate; *i.e.,* more or less covered with microscopic bristles (setae or aculeae).

H.E. Hinton

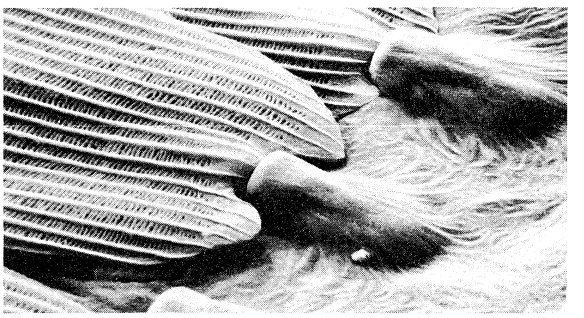


Figure 36: Detail of scales from the wing of a monarch butterfly, (magnified about 2,600 ×).

The scales and hairs of the adults, especially on the wings, are responsible for colours and patterns. There are many different pigments of several chemically diverse types such as melanins, uric acid derivatives, and flavones. In addition, the microstructure of hairs and scales refract light in various ways so as to produce colours as well as

metallic, pearly, iridescent, and white effects. Very fine parallel ridges in scales may produce an iridescence by acting as diffraction gratings. More common are the effects of very thin, superimposed layers (laminae) in the walls of the scales, a colour phenomenon comparable to that produced by a thin film of oil on water. Many of the most striking effects are caused by combinations of pigmental and structural colours.

Coupling of the fore and hind wings

The fore and hind wings on each side are coupled together in various ways. In primitive moths a fingerlike lobe (jugum, or fibula) on the forewing overlaps the base of the hindwing. In most moths a strong bristle or cluster of bristles (frenulum) near the base of the hindwing engages a catch (retinaculum) on the forewing. In some moths and in the skippers and butterflies the frenulum mechanism has been lost, and the wings are coordinated by the friction of the overlapping areas. In the most primitive moths the forewings and hindwings are similar in size, shape, and veins. In most moths and in the skippers and butterflies reduction of the hindwings has taken place; they have become shorter and more rounded, with reduced veining except in the anal cell area (the posterior wing section). The costal (anterior) edge of the forewing is thickened, with stronger veins, while the outer and anal (posterior) margins are thinner and weaker. This accords with the function of the wings as airfoils, the costal margin being the leading edge of the wing.

The wings are characteristically considerably modified in various families. In many moths of the superfamily Tineoidea both wings have become extremely narrow, with reduced, degenerate venation but with long, dense fringes along the margins, maintaining the functional wing area. In one family (Pterophoridae) the wings are deeply cleft into two or three narrow plumes; and in the Orneodidae each wing is cleft into six plumes. In a number of moth families the females are wingless, although the males are fully winged (Psychidae, some Liparidae). In the aquatic pyralid genus *Acentropus* some females are wingless, while the others and the males are winged. There are eight main wing veins, each with a characteristic pattern. These are usually designated according to the modified Comstock-Needham system. The names of the veins (with their symbols in parentheses) and the usual number of branches of each (subscript designations) are as follows, in progression from the costal margin of the wing to the anal, or inner, margin:

Costa (C), along the anterior margin, no branches; Subcosta (Sc), usually with no branches, rarely with two (Sc_1, Sc_2);

Radius (R), typically with five branches (R_1–R_5); Media (M), with three branches (M_1–M_3; four in some other insects);

Cubitus (Cu), primitively with three branches (Cu_{1a}, Cu_{1b}, Cu_2);

Anal Veins, First (1A), Second (2A), and Third (3A), all unbranched.

Abdomen and genitalia. The abdomen has ten segments, although the posterior ones are indistinct. Each of the first eight segments bears a pair of spiracles. The first or second segments bear paired auditory organs in the pyralid and geometrid moths. Segmental appendages are absent except for vestiges that may form parts of the genitalia. Various segments may bear special structures that produce and disperse pheromones. The genitalia of both sexes are often complex and bear characteristic spines, teeth, setae, and scale tufts. These structures are important in complex courtships and matings, preventing hybridization between unsuitable males and females.

In males (Figure 37) a ringlike structure (vinculum) is the base of attachment for a number of dorsal structures (tegumen, uncus, and gnathos) and a pair of lateral clasping organs (valvae). In copulation a median, tubular organ (the aedeagus) is extended through an eversible sheath (vesica) to inseminate the female. These structures evolutionarily derive from the body-wall plates, or sclerites (lateral tergites and ventral sternites), of segments eight and 10 and from vestiges of abdominal appendages.

The female genitalia exhibit a number of different patterns of the internal ducts and the openings, varying from a condition in which there are no special genital openings, insemination and egglaying (oviposition) taking place through a single aperture, shared with the excretory system, to one in which there are two specialized openings, one for insemination and one for oviposition, both distinct from the anus.

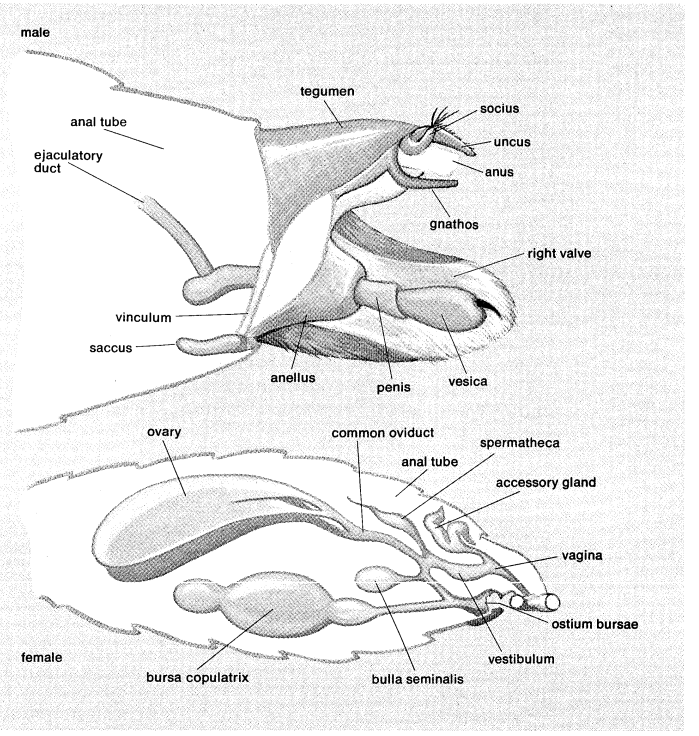


Figure 37: Genitalia and associated structures of the male and female Lepidoptera.

Internal reproductive system

The internal reproductive systems of both sexes contain the organs typical of most insects. The testes of the male are paired in primitive lepidopterans but fused into a single organ in advanced forms. In both cases, the sperm ducts are paired. As in other insects, the sperm pass from the testes down paired ducts (vasa deferentia) for storage in sacs called seminal vesicles. Accessory glands, providing fluids that lengthen the life of the sperm, open into the vasa deferentia.

The female reproductive system consists of paired ovaries, paired accessory glands that provide the yolks and shells of the eggs, and a system of receptacles and ducts for receiving, conducting, and storing sperm. The individual oviducts join to form a common oviduct that leads to the vagina. In copulation, the male deposits a sperm capsule (spermatophore) in a receptacle (bursa copulatrix) of the female. The spermatophore releases the sperm, which swim into the oviduct and thence to the seminal receptacle (bulla seminalis) where they are stored until egg laying, which may be hours, days, or months after mating.

The egg. The egg is enclosed in a protective shell, the chorion, through which a system of tiny canals, the micropyle, permits the entrance of sperm. In some groups the micropyle is at the side, in others on the surface away from the substrate. Formed in one of the paired ovaries, the egg passes along the individual oviduct and through the common oviduct to the vagina. Here, just before it is laid, it comes into contact with a droplet of seminal fluid that has been stored in the female. Fertilization now takes place. Most eggs are more or less spherical, but some (Eucleidae, Lycaenidae) are flat, while others (Pieridae) are long and tapered at the ends. The surface may be strongly sculptured with pits, sharp projections, or raised ridges.

The larva. Compared with the highly specialized adult, the larva is simple and primitive. Many of the primitive characters retained in the larva are important in the classification of the suborders, superfamilies, and families, and in phyletic speculations. The head bears a pair of very short antennae and on each side a cluster of minute, simple eyes (ocelli or stemmata). A short liplike labrum

lies anterior to the mouth. Behind the labrum are paired jaws (mandibles), short, broad and powerful. Next is a pair of small first maxillae, each with a segmented palp; and then, more or less connected with the maxillae, the labium-hypopharynx, a complex structure with a pair of labial palps and a median, tubular spinneret (a protrusion through which silk is extruded). Each of the three thoracic segments bears a pair of short, segmented legs. The remaining ten segments constitute the abdomen. Each of segments three through six and ten bears a pair of fleshy appendages called prolegs, which may be homologous with the primitive segmental appendages. Each proleg has one or two curved rows of minute hooklets (crochets) and an eversible, soft end, the planta. The prothorax and abdominal segments one to eight have each a pair of spiracles.

The larval epidermis bears a number of bristles, known as primary setae, on the head and each body segment. These setae are constant in position and number in each species and important in classification. There are often many secondary setae, sometimes forming a dense, hairy vestiture. Larvae that live and feed as borers, burrowers, or miners are mostly plain. Those that live and feed in the open often show a great diversity of shape, colour, pattern, and ornamentation. Many have wartlike projections (verrucae) that may bear tufts of setae or spiny projections; or there may be prominent hornlike or spiny processes (*e.g.*, Saturniidae, Sphingidae) or fleshy filaments (Danainae).

External ornamentation of the larva

In some families the number of the prolegs has been modified. In the measuring worms (larvae of the Geometridae) the prolegs of segments three, four, and five are missing, and in some noctuid moths one or two pairs have been lost. In Drepanidae and some Noctuidae the last pair has been lost; and in the puss moth caterpillars this pair has evolved into a pair of long, eversible whiplashes (stemmatopoda). The larvae of some leaf miners (Gracillariidae, Lyonetiidae) have lost some or all of the prolegs. Micropterigid larvae have eight pairs of abdominal legs, which are more like true legs.

Internally, the larva is relatively simple, the very large digestive tract being the most prominent organ. The paired silk glands are often very large, extending far back into the abdomen. The ovaries and testes, which begin to develop during embryonic life, continue to develop in the larva, as do the wings. Many special glands secrete repellant or toxic substances, which may circulate in the blood or be extruded from special openings as a means of defense.

Larval vision can detect little more than differences between light and darkness. Taste is acute, depending on delicate sensory receptors in the antennae and palpi. Food discrimination is keen and many larvae will starve rather than eat abnormal food plants. The sense of touch depends on setae well distributed over the outer surface. Some of these appear to react to sound waves of low pitch, well within the limits of human hearing.

The pupa. Lepidopterous pupae show the same sort of evolutionary gradation from primitive to advanced as do larvae and adults. In the primitive Micropterigidae and Eriocraniidae the pupa is exarate, having the appendages free and movable, and decticous, having functional mandibles. In some less primitive groups the pupa, called incomplete, has retained considerable powers of motion of some appendages. In the higher moths and the butterflies the pupa, called obtect, has all appendages tightly fastened and immobile and is able to wriggle only one or two abdominal segments. In a few groups (Sphingidae, Lycaenidae) the pupa has special stridulating rasps for sound production. Nearly all of the external structures of the adult can be seen on the pupa. The wings are prominent, folded down flat along the ventral surface, with the proboscis halves, the legs, and the antennae between them. At the posterior end is a spiny pad or spike, the cremaster, which in many groups attaches the pupa to silk fibres spun by the larva.

Except for a very slight respiratory exchange and a little water loss, the pupa is physiologically almost self-sufficient. Within it most of the cells and tissues of the larva undergo considerable histolysis (breaking down), as the adult structures are built up from the existing rudiments. Some structures begin developing as far back as the first larval stages.

Growth, molting, and metamorphosis. As in other insects, growth and its structural changes are controlled by an interacting set of hormones. These are chiefly secreted by the corpora allata and other parts of the brain, and by paired prothoracic glands. The prothoracic gland hormone is necessary for larval molting, metamorphosis to the pupa, and formation of adult characteristics. On the other hand, a hormone secreted by the corpora allata inhibits metamorphosis until late larval development. A hormone secreted by neurosecretory cells in the pupal brain stimulates the prothoracic glands and thereby brings about differentiation of the adult and the end of the obligatory resting stage (diapause) of the pupa.

EVOLUTION AND PALEONTOLOGY

The Lepidoptera belong to an important group of insect orders, the panorpoid complex, the ancestral stem of which began in the Permian. This stem split into a number of branches, from which evolved the modern orders Mecoptera (scorpion flies), Megaloptera (dobsonflies), Neuroptera (lacewings), Diptera (true flies), Trichoptera (caddisflies), and the Lepidoptera. The nearest living relatives of the Lepidoptera are the caddisflies and, in fact, the very primitive moths, the Micropterigidae, have been grouped with the caddisflies by some students. As the Lepidoptera developed into a distinctive group certain major trends began to be emphasized. The adult mandibles disappeared and were replaced by the proboscis formed from the galeae of the maxillae, broadening the food base exploitable by the species to include not only sap and other plant juices but nectar and fruit juices as the flowering plants evolved. Various groups, however, independently lost the proboscis and concentrated on the larval stage as the sole source of nutrition. The characteristic vein pattern evolved, with the long, veinless discal cell in each wing, and an emphasis on the strength of the forewing and a de-emphasis of the hindwing, forming a particular pattern of aerodynamic efficiency. Related, also, was the replacement of the primitive jugum, for wing coupling, with the more efficient frenulum mechanism. However, some efficient flyers, such as the bombycoid moths, the skippers, and the butterflies, subsequently lost the frenulum; and quite a few groups of small moths, settling into restricted ecological niches where strong flight is not a necessity, evolved a great deal of wing reduction.

Evolution of the proboscis

Concurrent with changes in the adults, the larvae were evolving a multiplicity of different ways of feeding on the evolving seed plants. As both larval and adult differentiations accumulated, it became more and more necessary that the pupal stage be a passive, resting phase, to allow time for the metamorphosis of larva into adult.

Very few fossil Lepidoptera are known. The earliest undoubted ones, small moths related to the Eriocranioidea and Tineoidea, are from the Early Tertiary Period. The order must have existed long before then, however. The closely related caddisflies are known from the Jurassic Period of the Mesozoic Era. The best known fossil butterflies are relatively recent, being from the shales of Florissant, Colorado, generally considered to be Middle Tertiary. Some of these are not only undoubted Nymphalidae but are very close to modern genera.

CLASSIFICATION

Distinguishing taxonomic features. The chief characters used in the ordinal, subordinal, and family-level classification of the Lepidoptera are: the mouthparts, ranging from mandibulate to haustellate (with a proboscis) or degenerate; the venation and shape of the wings, homoneurous (the venation of the forewings and hindwings alike) or heteroneurous (forewings and hindwings different), aculeate (more or less covered with specialized bristles called microsetae) or nonaculeate, and type of coupling (jugum or frenulum); the anatomy of the female reproductive system; the exarate or obtect condition of the pupa; and the larval structure and pattern of the primary setae.

The venation of the wings is perhaps the most important single criterion for establishing both differences and relationships in the classification. It must be considered in terms of the evolution of venation changes from prim-

Wing venation and genitalia

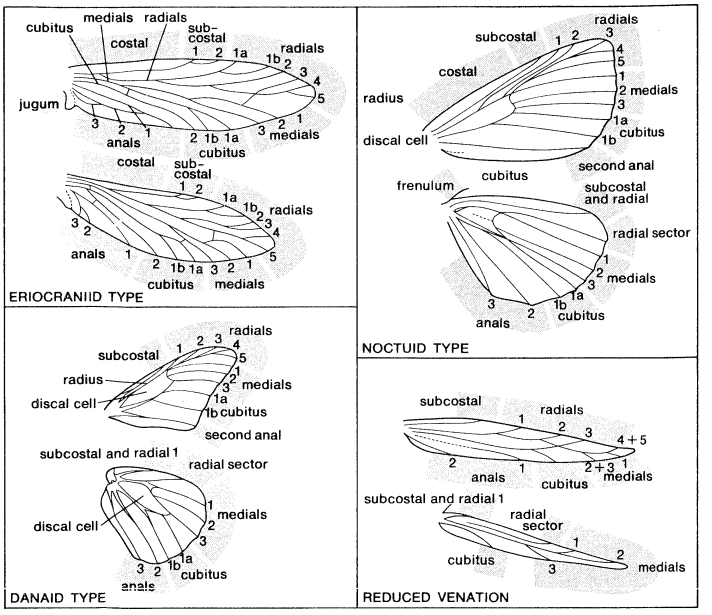


Figure 38: Types of wing venation.

From (top left) J. Comstock, *The Wings of Insects*, Comstock Publishing Co.; (top right) *An Introduction to the Study of Insects*, 3rd ed. by Donald J. Borror and Dwight M. DeLong, copyright © 1964, 1971 by Holt, Rinehart and Winston, Inc., copyright 1954 by Donald J. Borror and Dwight M. DeLong, reproduced by permission of Holt, Rinehart and Winston, Inc.; (bottom right) J. Comstock, *An Introduction to Entomology* (1956), Comstock Publishing Co.

itive to advanced conditions on a number of different phyletic lines. The most primitive groups (Figure 38, top left) tend to have the maximum number of veins and branches in each wing. More advanced groups (Figure 38, top right and bottom left) lack some veins or branches in the forewing and have the hindwing venation considerably reduced, especially by the reduction of free radial branches, the unbranched stem of R_2–R_5, being called the radial sector (R_s).

The structures of the genitalia are extremely important in classification, often serving better than any other characteristics to distinguish the species. The various parts have been homologized with each other, thus enabling comparative studies of families and genera as well as studies of species.

The female genitalia (Figures 37 and 39) show a number of different types of organization of the internal genital ducts and the openings. These are considered so fundamental that the subordinal classification is largely based on them, in correlation with characteristics of the mouthparts, wings, and early developmental stages. These genitalic types are as follows:

1. No special genital openings. Both insemination and oviposition take place through the cloaca, a chamber common to both the reproductive and excretory systems, the aperture being the anus.
2. A separate genital opening for insemination (reception of seminal fluid) but oviposition through the cloaca and anus.
3. Separate genital openings for insemination and oviposition, both close together and both close to the anus.
4. Separate openings for insemination and oviposition, the latter still close to the anus but the former removed some distance ventrally, in the eighth segment.

Type 1 characterizes the extremely primitive suborder Zeugloptera, the mandibulate Micropterigidae. Type 2 occurs in the superfamilies Eriocranioidea, Nepticuloidea and Incurvarioidea of the generally primitive suborder Monotrysia.

Type 3, sometimes called "exoporian," characterizes the monotrysian superfamily Hepialoidea, also quite primitive in other ways. Type 4 characterizes the most advanced suborder Ditrysia, containing the great majority of Lepidoptera.

Annotated classification. The classification given here largely follows that of the British entomologist A.D. Imms, which, in itself, represents a synthesis of the suggestions of several earlier authors. This classification incorporates changes on the family level that have been suggested recently. There is no clear evidence for the phyletic relationships of numerous families, especially in the greatly reduced tineoid moths.

ORDER LEPIDOPTERA

Insects with complete metamorphosis; the wings covered with flat scales; larval mouthparts mandibulate; adult mouthparts mandibulate to haustellate, sometimes vestigial; wings with venation ranging from primitive (complex) to considerably reduced, especially in the hindwings; pupae exarate to obtect; almost all phytophagous (herbivorous).

Suborder Zeugloptera

Female with no special genital openings. Larvae, pupae, and adults functionally mandibulate; wings aculeate, homoneurous, jugate.

Family Micropterigidae (mandibulate moths). Small, a few species in the Palaearctic and Nearctic, more in Australia and New Zealand. The adults eat pollen; the larvae eat mosses and liverworts.

Suborder Monotrysia

Females with 1 or 2 genital openings near anus. Adults at least partly haustellate. Wings homoneurous to heteroneurous, at least partly aculeate. Jugum present in some, frenulum in others. As a group intermediate between Zeugloptera and Ditrysia.

Superfamily Eriocranioidea

Females with 1 genital opening, a long cloaca, and a piercing ovipositor. Wings homoneurous and aculeate. Adults with a short proboscis. Pupae with functional mandibles.

Family Eriocraniidae (eriocraniid moths). A small, widely distributed family, chiefly Nearctic and Palaearctic. Adults nectar feeding, often brilliantly coloured. Related families: Mnesarchiidae (New Zealand), Mesopseustidae (India and Taiwan).

Superfamily Hepialoidea

Females with 2 genital openings. Wings homoneurous and aculeate. Adult mouthparts reduced and nonfunctional, antennae very short.

From A. Imms, *Textbook of General Entomology;* Methuen & Co. Ltd.

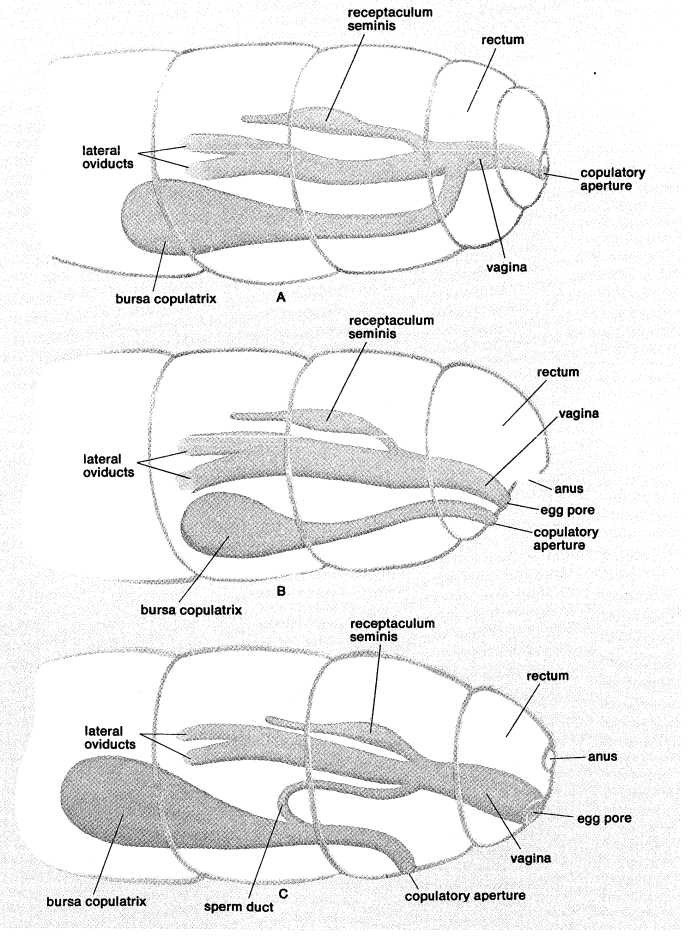


Figure 39: *Relationships of female reproductive ducts and rectum in Lepidoptera.*
(A) Micropterigid type. (B) Hepialid type. (C) Ditrysian type.

Family Hepialidae (swifts and ghost moths). Medium-sized to very large moths, some brilliantly coloured. Larvae mostly borers in turf or wood. About 300 species; worldwide, chiefly in Australia and New Zealand. Related families, less known: Prototheoridae (Africa and Australia), Palaeosetidae (Australia), Anomosetidae (Australia).

Superfamily Nepticuloidea

Females with 1 genital opening and a soft ovipositor. Wings somewhat aculeate, heteroneurous, and somewhat narrowed and with reduced venation.

Family Nepticulidae (midget moths). Very small to minute moths. Antennae with broad "eyecaps" basally. Larvae mostly leaf and bark miners, a few gall makers. Worldwide. Also called Stigmellidae.

Families Opostegidae and Tischeriidae. Worldwide families of small moths with narrow, long-fringed wings. Larvae leaf, stem, or bark miners.

Superfamily Incurvarioidea

Family Incurvariidae (fairy moths). A small but worldwide family of small moths, many being brilliantly coloured diurnal flower visitors. Male antennae often several times as long as forewings. The mutualistic relationships of the yucca moths (Prodoxinae) with their food plants are biologically famous. Family sometimes split into families Incurvariidae, Adelidae and Prodoxidae. The related Heliozelidae (shield bearers) are worldwide.

Suborder Ditrysia

Females with 2 genital openings, the receptive one on the 8th abdominal segment. Wings strongly heteroneurous, nonaculeate, mostly with frenulum but this is degenerative or lost in some. Contains the great majority of Lepidoptera: higher moths, skippers, and butterflies.

Superfamily Zygaenoidea

Wings broad to very broad with short fringes and some primitive venational and larval features. Pupae quite primitively mobile.

Family Eucleidae (slug caterpillar moths). Larvae broad and flat ventrally, with reduced prolegs; move glidingly with head hidden beneath prothorax; many with toxic, irritant setae. Adults with heavy hairy bodies and vestigial proboscises. Also called Limacodidae.

Family Megalopygidae (flannel moths). Larvae like those of Eucleidae, but with prolegs normal, plus traces of some additional ones; setae very toxic and nettling. Almost limited to New World.

Family Zygaenidae (burnet and forester moths). Adults usually diurnal flower visitors, with bright colours and strong proboscises. Strongly protected by high concentrations of hydrogen cyanide in blood. Larvae, leaf skeletonizers. Related families: Aididae and Chalcosiidae (Old World tropics); Pyromorphidae and Dalceridae (New World).

Superfamily Tineoidea

A large group of families of mostly small moths of diverse habits. All have some primitive venation features and life cycles. Wings narrow to very narrow.

Family Tineidae (clothes moths and others). Small, narrow-winged moths with rough, hairy heads and often long, folded maxillary palpi. Larvae often casemakers, feeding on debris and fungi. Worldwide. Clothes moths (*Tineola, Tinea, Trichophaga*) often serious household pests. Related family: Acrolophidae, burrowing sod webworms.

Family Psychidae (bagworms). Larvae live and pupate in often elaborate cases. Adult males with broad, thin scaled wings; females wingless, often greatly degenerate and never leaving larval cases. Worldwide.

Family Epipyropidae (parasitic moths). Small family, chiefly Oriental. Larvae live as external parasites on plant hoppers (Homoptera). Related family: Cyclotornidae, Australian, larvae living similarly when young, then moving to ants' nests.

Family Coleophoridae (casebearer moths). Small, very narrow-winged moths. Larvae mostly mine leaves when young, then live in portable cases with distinctive shapes. Worldwide. Some are pests of fruit trees.

Families Gracillariidae, Lyonetiidae, Elachistidae and Douglasiidae. A group of worldwide families, totalling about 2000 species. Adults with narrow, long-fringed wings often with metallic markings. Larvae mostly leaf miners, sometimes greatly flattened, with degenerative legs and mandibles.

Superfamily Gelechioidea

A group of worldwide families totalling more than 12,000 species. Adults mostly larger and broader winged than Tineoidea. Larvae seldom leaf miners. Pupae obtect, relatively immobile.

Family Gelechiidae (gelechiid moths). Small to minute moths. Larvae diverse, mostly leaf tiers, rollers, and crumplers; stem, fruit, and tuber borers; a few gall makers. Hindwing often concave below apex. Worldwide. Some are serious economic pests; *i.e.,* the pink bollworm of cotton (*Pectinophora gossypiella*) and the Angoumois grain moth (*Sitotroga cerealella*).

Family Cosmopterygidae (cosmopterygid moths). Many adults are very narrow winged with bright, often metallic markings. Small moths. Larvae leaf miners; borers in stems, fruits, or seeds; leaf tiers; or scavengers. Related families; Lavernidae, Momphidae, Walshiidae.

Family Oecophoridae (oecophorid moths). Adults tend to be somewhat broader winged than related groups, and flat bodied; all small. Related families: Ethmiidae, Stenomatidae, Xyloryctidae.

Family Blastobasidae (blastobasid moths). The adults small, often drab; reduced mouthparts. Larvae mostly scavengers, a few predators on scale insects.

Superfamily Yponomeutoidea

A limited, not very distinctive superfamily, the adults with close scaling and venation peculiarities, the larvae with distinctive primary setation.

Family Aegeriidae (wasp or clearwing moths). Almost worldwide; adults diurnal, flower visitors, often brightly coloured with yellow, orange, or scarlet, the wings usually mostly transparent, often very striking mimics of wasps. Larvae borers in stems, twigs, and rootstocks, often injurious; *e.g.,* the peach tree borer (*Sanninoidea exitiosa*). Also called Sesiidae.

Family Yponomeutidae (ermine moths). Adults brightly coloured, especially in tropics. Larvae often webbing leaves or living in shoots, buds, or fruits. Pupae of some make lacework cocoons. Widespread and not uncommon. Related families: Plutellidae, Glyphipterygidae, Heliodinidae, Scythrididae.

Superfamily Cossoidea

Family Cossidae (carpenterworm or goat moths). Medium to large moths. Adults with some primitive venation features, close-scaled wings, very heavy bodies, ranging up to 250 mm wing expanse, the heaviest Lepidoptera. Larvae wood borers, penetrating even solid heartwoods; may require at least 2 years to mature. Some species very destructive; *e.g.,* the coffee borer (*Zeuzera coffeae*) and leopard moth (*Z. pyrina*).

Superfamily Castnioidea

Family Castniidae (castniid moths). A small family of medium sized to large diurnal moths of the New World and Indo-Australian tropics. Adults powerful, heavy bodied and broad winged; with clubbed antennae, bright colours, and often a strong mimetic resemblance to protected butterflies and diurnal moths. The larvae are often stem borers.

Superfamily Tortricoidea

The families included here comprise one of the major, worldwide groups. They have fairly broad, short-fringed wings that seldom span more than 25 mm. Maxillary palpi missing; labial palpi short. Most have cryptic colours and patterns. Larvae mostly leaf folders and rollers, but many bore in fruits and seeds and soft stems.

Family Tortricidae (leaf roller moths). Characteristics of the superfamily. Separated from Olethreutidae (below) by wing venation. Larvae of the green *Tortrix viridana* of Europe defoliate oak forests: those of the spruce budworm (*Choristoneura fumiferana*) are the worst forest pests of North America. Some species of *Peronea* and *Acleris* show most extraordinary polymorphisms of colour and pattern.

Family Olethreutidae (olethreutid moths). Separated from other tortricoid families by wing venation. Many are serious pests; *e.g.,* the pine shoot moths (*Rhyacionia*), the apple codling moth (*Laspeyresia pomonella*), and the oriental fruit moth (*L. molesta*). Chief subfamilies, the Olethreutinae, Eucosminae, and Laspeyresiinae, are often ranked as families. Related smaller families: Phaloniidae (chiefly Northern Hemisphere), Chlidanotidae (Indo-Australian and Neotropical) and Carposinidae (chiefly Indo-Australian, sometimes considered a separate superfamily).

Superfamily Pyralidoidea

An enormous group of about 20,000 species. Worldwide. Veins Sc and R_1 of the hindwings are typically close together or fused to beyond the cell. Most members have a pair of tympanal organs on the 1st abdominal segment. Adults usually slender bodied with long legs; many with narrow forewings and broad, often folded, hindwings.

Family Pyralididae (pyralid moths). Small moths, mostly plain, often abundant. Separated from other families by wing venation. Palps often large and held forward. Family contains many important pest species. Third largest family of the Lepidoptera. The subfamily Pyraustinae is the largest; most of the larvae live in rolled or tied leaves, or bore in soft stems or roots. The small subfamily Nymphulinae has aquatic larvae with tracheal gills, living in still or running freshwaters. The larvae of the Pyralidinae are mostly scavengers, as are those of the Galleriinae, many of which live in bee or wasp nests. Larvae of the Crambinae are either sod webworms or bore in soft stems. The larvae of the large subfamily Phycitinae have very diverse habits: most are leaf tiers or webbers, but many are borers, others are scavengers, others pests in stored foods, and some are predators on scale insects. Other important subfamilies are: Chrysauginae, Schoenobiinae, Endotrichinae, and Ancylolominae. Many subfamilies have been separated as families. Probably related is the small, chiefly tropical family Thyrididae.

Family Pterophoridae (plume moths). Adults with very long, slender legs and bodies, the wings usually deeply cleft into plumes. Larvae spin webs on and eat the leaves of various plants.

Family Orneodidae (feather-winged moths). Small, but worldwide, family of uncertain relationships. Each wing is very deeply cleft into 6 or more narrow plumelike divisions.

Superfamily Bombycoidea

Adults large to very large. Frenulum usually lacking; proboscis reduced and nonfunctional; tympanic organs and maxillary palpi lacking; male antennae pectinate.

Family Bombycidae (silkworm moths). A very small family consisting of the domesticated silkworm (*Bombyx mori*) and a few relatives, all Asian. Related family: Eupterotidae.

Family Lasiocampidae (tent caterpillar and lappet moths). Larvae usually hairy and brightly coloured, some living gregariously in silk nests. Adults medium sized, stout bodied, short winged, exceptionally hairy. Widespread.

Family Saturniidae (giant silkworm moths). Every continent has some well known species—*e.g.,* the North American cecropia (*Hyalophora cecropia*), io (*Automeris io*), polyphemus (*Antheraea polyphemus*), and luna (*Actias luna*); European giant peacock or emperor (*Saturnia pavonia*); and Indo-Australian *Attacus atlas and Coscinoscera hercules,* the last with wing expanses up to 275 mm (11 in.). Males with broadly branched antennae; often with bright colours and striking wing shapes and patterns. Larvae very large and fleshy, often with brightly coloured knobs and spines. Most species spin firm cocoons of brown, green, or silvery silk. Wingspread up to 275 mm (in hercules moth, *Coscinoscera hercules*). Worldwide.

Family Citheroniidae (regal moths). Medium to very large, heavy bodied moths, often brightly coloured. Larvae often with long spines; *e.g.,* the North American *Citheronia* (hickory and pine horned devils). Pupation in the ground, with no cocoon. New World only. Related families: Brahmaeidae (African and Indo-Australian) and Lacosomidae (New World).

Superfamily Geometroidea

Adults with frenulum and abdominal tympana. Some authorities place each of the 3 major families as a separate superfamily.

Family Geometridae (measuring worm moths). Larvae ("measuring worms" or "inch worms") long, slender, lacking most of posterior abdominal prolegs, crawling with characteristic looping gait. Pupation usually in the soil, with no cocoon. Adults and larvae commonly very cryptic, resembling bark, dead leaves, and twigs. Many species destructive to foliage; *e.g.,* cankerworms (*Alsophila* and *Palaeacrita*) and the winter moth (*Operophtera brumata*). The second largest moth family, worldwide and abundantly represented nearly everywhere.

Family Uraniidae (uraniid moths). Chiefly tropical. Some adults—*e.g., Chrysiridia* (Madagascar) and *Urania* (Neotropics)—are large, brilliantly iridescent diurnal moths. The Oriental *Epicopeia* (sometimes separated as the family Epicopeiidae) mimic protected swallowtail butterflies.

Family Drepanidae (hooktip moths). A small, worldwide family, chiefly Indo-Australian. Many of the adults have the forewing apexes strongly hooked. Larvae usually with last pair of prolegs lacking. Related families: Thyatiridae and Epiplemidae, both worldwide.

Superfamily Sphingoidea

Family Sphingidae (hawk or sphinx moths). Adults medium sized to large; powerful, with long, narrow forewings, thick antennae and usually strong proboscises; feed by hovering before flowers and drawing nectar through the extended proboscis. Some are diurnal, most are active at twilight. Larvae large, fleshy, mostly smooth; most with a single, long, posterior, dorsal horn; typically rest with head and thorax reared up, fancifully like the Sphinx in profile. Worldwide.

Superfamily Noctuoidea

Adults with frenulum and a pair of complex tympanic organs on metathorax.

Family Noctuidae (owlet moths). The largest family of Lepidoptera, with abundant members everywhere in a great diversity of size, colour, and habit. Sizes range from a wing expanse of 10 mm (*Hypenodes*) to 275 mm (*Thysania*). Many species brightly coloured, but the majority plain and cryptic. Nearly every type of plant food is exploited: foliage, flowers, fruits, stems, rootstocks, and woodland litter. A few genera (*e.g., Cosmia*) cannibalistic. Family includes the army worms and cutworms, among the most injurious of all moth larvae to man's interests.

Family Agaristidae (forester moths). A small, worldwide group; many adults brightly coloured, diurnal, and aposematic or mimetic.

Family Notodontidae (prominents). Adults mostly dull, cryptic. Larvae very diverse: cryptic, disruptive, or aposematic; some highly protected by toxic secretions; posterior prolegs often modified or aborted; chiefly foliage eaters.

Family Arctiidae (tiger moths). Adults, especially in tropics, mostly brightly coloured, aposematic, with protective and toxic secretions. Timbal organ for making very high pitched sounds on each side of metathorax. Larvae often with much secondary hair. Worldwide. Several other families, chiefly tropical, are closely related, also have timbal organs.

Families Dioptidae and Pericopidae. Similar to Arctiidae, separable by wing venation. Many are boldly coloured and marked, diurnal, members of mimicry complexes, often mimicking wasps. New World only.

Family Ctenuchidae (ctenuchid moths). Similar to Arctiidae; separable by wing venation. Includes many diurnal mimics of protected wasps and beetles. Worldwide. Related Lithosiidae also contain many aposematic species; their larvae feed on lichens.

Family Liparidae (tussock moths). Many larvae have prominent tussocks and pencils of hair, poison spines, and osmeteria (protrusions that emit repellant odours). Adults are broad winged, the females heavy bodied, sometimes wingless. Some species—*e.g.,* the gypsy moth (*Lymantria dispar*) and the nun moth (*L. monacha*)—are very destructive to forests. Closely related Thaumetopoeidae of the Old World are best known for the habits of the gregarious larvae (processionary caterpillars), which move in orderly columns.

Superfamily Hesperioidea (skippers)

The skippers (like the true butterflies) are distinguished from moths by diurnal habits, clubbed antennae, a functional proboscis, lack of ocelli, and lack of a frenulum. Adult skippers are fast flying, with short, usually pointed forewings, broad heads, antennae usually hooked beyond the club, and all veins arising from the forewing cell simple. Larvae usually have a pronounced, necklike collar, often with a dorsal sclerotized area.

Family Euschemonidae (regent skipper). A single anomalous Australian species; the male has a frenulum, like a moth; the female lacks it.

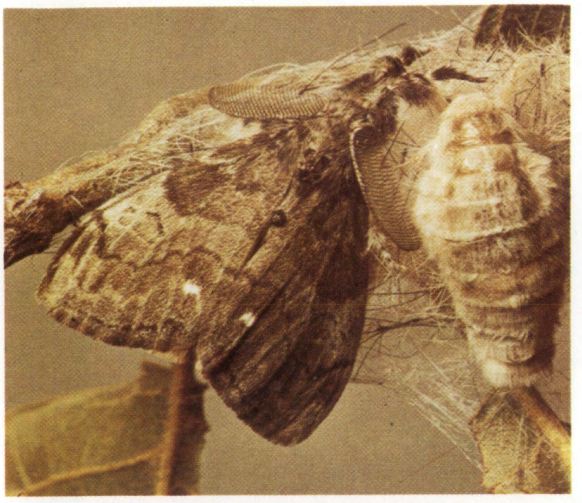
Family Megathymidae (giant skippers). A small, New World family of relatively large skippers, whose larvae bore in stems of agaves and yuccas, constructing silk-lined tubes. Adult wingspan to about 90 mm.

Family Hesperiidae (typical skippers). With more than 3,000 species, the adults range from small, drab species to large, often iridescent, tropical ones. Larvae mostly living concealed in individual leaf nests or in webs among grasses, forming flimsy cocoons. Worldwide.

Superfamily Papilionoidea (true butterflies)

Adults with antennae clubbed, but not hooked, flight slower and not darting; larvae lacking "neck." General evolutionary trend from a primitive condition with the forelegs fully developed in both sexes (Pieridae, Papilionidae) to one in which they are greatly reduced and useless for walking (Nymphalidae). Intermediate conditions occur in Lycaenidae. Pupae ("chrysalids") often brightly coloured and irregularly shaped, not enclosed in cocoons; either hanging head down (Nymphalidae) supported from cremaster, or additionally supported by a silk girdle (Pieridae, Papilionidae, Lycaenidae). All families worldwide.

Family Pieridae (whites and sulphurs). Small to medium sized; mostly white, yellow, or orange, often with dark tips



Tussock moths *(Notolophus leucostigma)* on female cocoon. The wingless adult female lives solely for the purpose of oviposition.



Female tussock moth laying eggs in masses on the surface of her old cocoon. She covers the eggs with a sticky secretion that becomes dry and brittle.

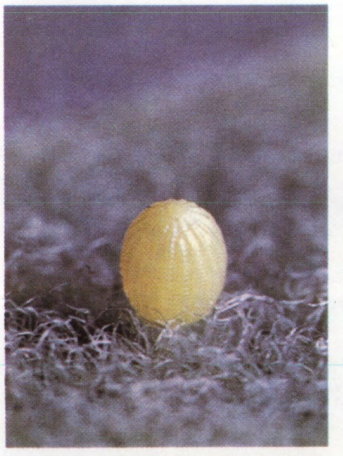


Polyphemus moth depositing eggs. Because the larvae feed on a variety of trees and shrubs, site selection for egg deposition is haphazard.



Eyed hawk moths *(Smerinthus ocellatus)* copulating. The female is at the top.

Mating and egg laying in Lepidoptera





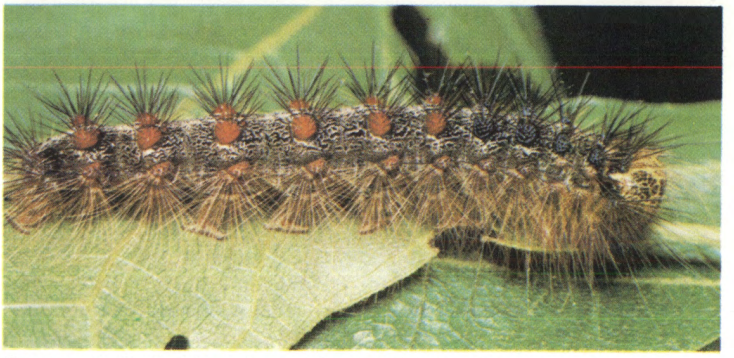
Lepidoptera eggs: (above left) monarch butterfly *(Danaus plexippus);* (above right) clouded sulfur butterfly *(Colias philodice);* (below) handmaiden moth *(Datana ministra).*



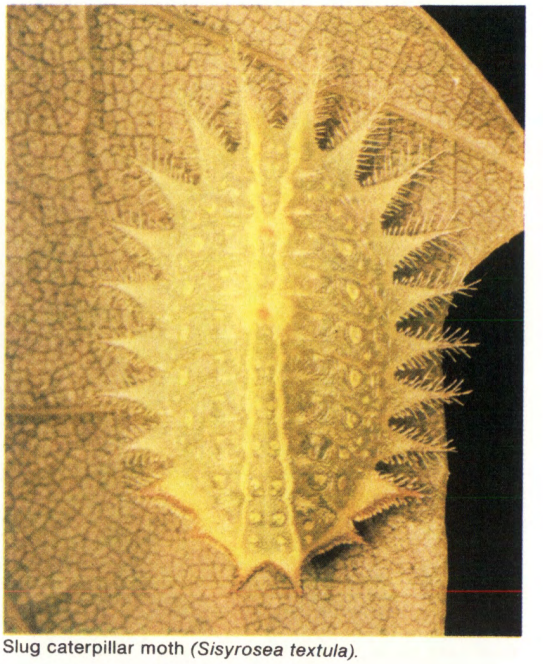
Plate 1: (Top left, centre, centre right, bottom right) Alexander B. Klots, (top right) Toni Angermayer—Photo Researchers, (bottom left) Louis Quitt—Photo Researchers, (centre left) Josephine Cikel—The National Audubon Society Collection/Photo Researchers



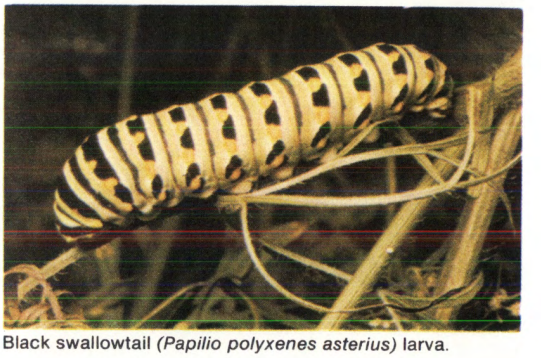
Tussock caterpillar *(Halisidota argentata).*



Gypsy moth *(Porthetria,* or *Lymantria, dispar)* larva, a species injurious to orchards and forests in the United States.



Slug caterpillar moth *(Sisyrosea textula).*



Black swallowtail *(Papilio polyxenes asterius)* larva.

Lepidopteran larvae

Wood nymph *(Eudryas grata)* caterpillar.



Larva of the viceroy *(Basilarchia,* or *Limenitis, archippus).*



Bags of the creosote bush bagworm moth *(Thyridopteryx meadi).* Larvae live inside tubular baglike cases constructed from silk and bits of leaves, twigs, and other debris.

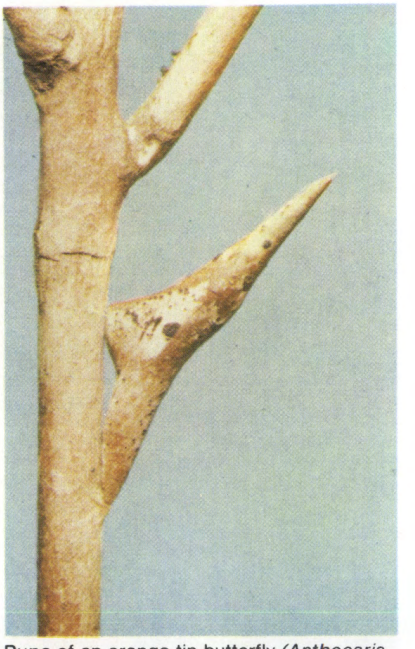




Spicebush swallowtail *(Papilio troilus)* pupa supported head upward by a girdle and cremaster (terminal abdominal spine).



Pupa of the mourning cloak *(Nymphalis antiopa)* suspended head downward by only the cremaster.



Pupa of an orange tip butterfly *(Anthocaris midea)* suspended by both the cremaster and a girdle.

Pupae and aggregations of Lepidoptera



American painted lady *(Vanessa virginiensis)* emerging from its chrysalis.



Cocoon of the Isabella tiger moth *(Isia isabella)* open to show pupa.





(Above) Aggregation of pink edged sulfur butterflies *(Colias interior)*.
(Right) American eastern tent caterpillars *(Malacosoma americanum)* constructing a silken nest for shelter.

Plate 2: (Top left) E.S. Ross, (top right, above centre left, below centre left, centre right) Alexander B. Klots, (bottom left) E.R. Degginger, (bottom right) William E. Ferguson

Plate 3: (Top left) John H. Gerard, (top centre, centre right, bottom right) Louis Quitt—Photo Researchers, (bottom left) John H. Gerard, (top right, centre left) Alexander B. Klots



An orange tip butterfly *(Anthocaris cardamines).*



South African day-flying moth *(Euchromia formosa).*

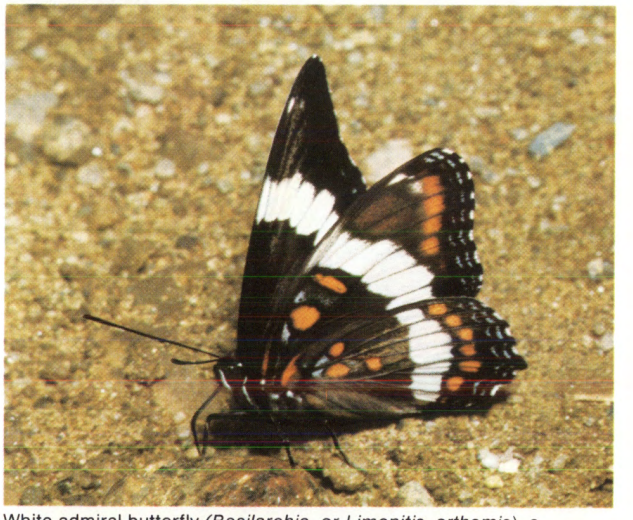


Underwing moth *(Catocala).*

Colourful and unusual Lepidopteran adults



Palaeochrysophanus hippothoe, an iridescent butterfly.



White admiral butterfly *(Basilarchia,* or *Limenitis, arthemis),* a common North American species.







(Above left) *Eriocrania auricyanea,* a primitive moth; (above) calico moth *(Attava punctella);* (left) ermine moth *(Zelleria heimbachi).*

Plate 4: (Top left, above centre left) Hermann Eisenbeiss—Photo Researchers, (above centre right) Ken Brate—Photo Researchers, (top centre) Anthony Bannister from The Natural History Photographic Agency—EB Inc., (top right) E.R. Degginger, (below centre left, below centre right) Alexander B. Klots, (bottom left) William E. Ferguson

on wings. Pupae usually with a frontal horn on head. Many species noted for mass migrations.

Family Papilionidae (swallowtails and parnassians). Medium to large; often brightly coloured. Many have the "swallowtails" on the hindwings. Some females of the Indo-Australian genus *Troides* are the largest butterflies. Larvae often curiously patterned, with eversible, protective scent organs (osmeteria). Adults of many are highly distasteful and much mimicked. Worldwide. Parnassians sometimes placed in a separate family, Parnassiidae.

Family Lycaenidae (blues, coppers and hairstreaks). A large diverse family of small butterflies. Many species are iridescent blue, green, or, in the coppers (Lycaeninae), metallic coppery. Larvae somewhat sluglike, hairy, many secreting honeydew and having mutualistic relationships with ants; some prey on the ant brood or on other insects. Many pupae have stridulatory structures on the abdomen.

Family Riodinidae (metalmarks). Small to medium sized. Closely allied to the Lycaenidae. Family has an unequalled diversity of wing shapes and colors. Many South American species have complex mimicry relationships. Chiefly tropical American.

Family Libytheidae (snout butterflies). Long, protruding palpi. Some are strongly, but unpredictably, migratory. Worldwide, but with only about a dozen species.

Family Nymphalidae (nymphalid butterflies). The largest family of butterflies, so diverse that it is often split into several families, here considered subfamilies. The Nymphalinae is the main group, with many familiar species such as the fritillaries, admirals, checkerspots and anglewings. Many tropical species brilliantly iridescent. Satyrinae contains the familiar wood nymphs, meadow browns, and heaths, usually with eyespots on the wings; the larvae distinctively pointed posteriorly and spinning crude cocoons. The very large Brassolinae and iridescent Morphinae are Neotropical, as are the highly distasteful, aposematic Heliconiinae and Ithomiinae that, with the worldwide Danainae, are models in many mimicry complexes. Most of the pantropical Acraeinae are also highly protected and aposematic models.

Critical appraisal. The terms microlepidoptera and macrolepidoptera are applied, for convenience, to two nontaxonomic divisions of the order Lepidoptera. The microlepidoptera include the suborders Zeugloptera and Monotrysia, and the superfamilies Zygaenoidea through Pyralididoidea of the Ditrysia. The remaining superfamilies of the Ditrysia, consisting of the bombycid, geometrid, sphingid, and noctuid moths, the skippers, and the true butterflies, comprise the macrolepidoptera. Although, in general, the microlepidoptera are smaller than the macrolepidoptera, some of the former, such as certain hepalids and cossids, are among the largest moths, with adult wingspans up to 250 millimetres, while some macrolepidoptera have adult wingspans of only 10 millimetres.

(A.B.K.)

Coleoptera (beetles, weevils)

The members of the Coleoptera—beetles and weevils—comprise the largest order of insects, representing about 40 percent of the known insect species. Among the approximately 250,000 species of Coleoptera are many of the largest and most conspicuous insects, some of which also have brilliant metallic colours, showy patterns, or striking form. Beetles can usually be recognized by their two pairs of wings; the front pair is modified into horny covers (elytra) that hide the rear pair and most of the abdomen and usually meet down the back in a straight line. Coleoptera occur in nearly all climates. They may be divided into four groups: the first three, the Archostemata, the Adephaga, and the Myxophaga, contain relatively few families; the majority of beetles are placed in the fourth group, the Polyphaga.

Beetles attract attention for many different reasons—economic importance, size, abundance, appearance, and remarkable habits. Several groups of beetles (*e.g.*, Lampyridae) are among the few terrestrial animals capable of producing light; members of several other families (*e.g.*, Cerambycidae) can produce sound (stridulate). Most large beetles make a loud noise during flight, and many species, both large and small, are attracted to light at night. Some beetles (*e.g.*, burying beetles of the family Silphidae and whirligig beetles of the family Gyrinidae) attract attention by their bizarre habits; others do so by their grotesque forms (*e.g.*, Scarabaeidae). Many beetles have become adapted to an aquatic environment (*e.g.*, Hydrophilidae); others (*e.g.*, Thorictidae) live in association with ants and termites.

GENERAL FEATURES

Distribution and abundance. Beetles are found in nearly all climates and latitudes, except in such extreme environments as those in Antarctica and at the highest altitudes. They are found on sub-Antarctic islands, close to the northern extremes in the Arctic, and on many mountain tops. Although many species occur in temperate environments, the number of species is greatest in the tropics; in general, individuals of a species are most abundant in temperate areas, and fewer individuals of more species are found in the tropics.

Some species are solitary, others occur in aggregations. Predators like the ground beetles (Carabidae) are more apt to be found alone, as are many long-horned woodboring beetles (Cerambycidae) and weevils (Curculionidae). Ladybugs (Coccinellidae), leaf beetles (Chrysomelidae), pleasing fungus beetles (Erotylidae), darkling beetles (Tenebrionidae), checkered beetles (Cleridae), bess beetles (Passalidae), sap beetles (Nitidulidae), and some species of scarab beetles (Scarabaeoidea) are often found in aggregations of one or several different species.

Solitary and gregarious beetles

Most families contain both widely distributed species and some with very limited ranges. Wide distribution in this sense refers to a zoogeographical or faunal region, limited distribution to a single valley, plain, island, altitude zone, or vegetation type on a mountain.

Size range and diversity of structure. Coleoptera vary greatly in size, from a fraction of a millimetre to more than 200 millimetres (almost 8 inches) in length (*e.g.*, rhinoceros beetle, *Xyloryctes satyrus*) and up to 75 millimetres in width (*e.g.*, goliath beetle, *Goliathus goliathus*).

Diversity of structure among adult beetles is as great as range of size. The ground beetles (Carabidae) have a rather generalized (primitive) form—the flattened, oval body has a relatively even surface, with regular ridges or grooves; antennae and legs are of moderate length and slender.

From R.A. Pimentel, *Invertebrate Identification Manual*, ©1967 by Litton Educational Publishing, Inc.; reprinted by permission of Van Nostrand Reinhold Company

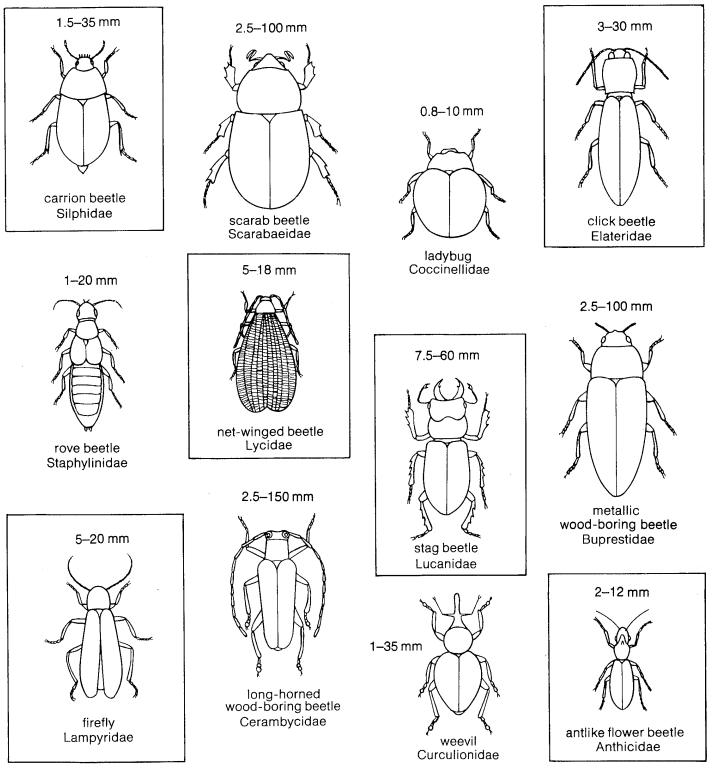


Figure 40: Representative beetles.

The underside of most water beetles (Hydrophilidae) is oval, smooth, and flattened, the antennae either short or very slender, and the forelegs short, the hindlegs long and fringed with hairs that are used as paddles. Rove beetles (Staphylinidae) have very short elytra and a slender abdomen. Staphylinid larvae have a form similar to that of the adults. Soldier beetles (Cantharidae), fireflies (Lampyridae), and net-winged beetles (Lycidae) have an un-beetle-like appearance; *e.g.,* soft elytra.

Click beetles (Elateridae) have a hingelike joint in the body region called the thorax that enables them to snap their bodies and jump high in the air; their relatives, the Buprestidae (metallic wood borers), cannot jump but take flight very quickly. Cleridae (checkered beetles) are usually oblong or cylindrical, fairly active, and often brightly coloured. Nitidulidae (sap beetles) are short, flattened, and have slightly shortened elytra. Coccinellidae (ladybugs, ladybird beetles) are rounded, with a smooth, raised upper surface and a flat underside. The Endomychidae (handsome fungus beetles) often have enlarged, rounded elytra. The Erotylidae (pleasing fungus beetles) are usually slender, smooth, and shiny, as are the Languriidae.

Among the stout or cylindrical lamellicorns (Scarabaeoidea) are a number of bizarre forms. The male rhinoceros beetles (Dynastinae) have one or more horns on the head and sometimes on part of the thorax. Many of the true scarabs (Scarabaeinae) and other dung-feeding groups of the lamellicorns also have horns, including some of the goliath beetles (Cetoniinae). Male stag beetles (Lucanidae) have greatly enlarged mandibles (jaws); some are as long as the rest of the body.

The Chrysomelidae (leaf beetles) vary from simple eggshaped forms to slender, flat, or wedge-shaped ones, with wide elytra in the tortoise beetles (Cassidinae) and often numerous spines or tubercles in the leaf-mining leaf beetles (Hispinae). The Bruchidae (pea weevils) are short and stumpy, with short stout legs. The head, slightly elongated in front, is similar to that of some curculionid weevils. The Cerambycidae (long-horned wood-boring beetles), diverse in form and structure, usually have antennae that are longer than the body. Cerambycids may be slender and medium to large in size or very small.

The extremely diverse Tenebrionidae (darkling beetles) are not always recognized as members of one family. Most arboreal (tree-dwelling) forms in the tropics or subtropics are slender and long legged. Some slender and egg-shaped forms have a metallic sheen; most of the ground-dwelling forms are black and robust. The large, tropical Trictenotomidae resemble some cerambycids (Prioninae) but are not related to them. The Alleculidae (comb-clawed beetles) resemble some of the slender tenebrionids but are usually more active. Lagriidae (lagriids) have a characteristic shape, usually widened behind, and sometimes a metallic sheen. Colydiidae (cylindrical bark beetles), hard-bodied and shiny or roughened, may be cylindrical and somewhat flattened.

Weevils

The weevils range from slender to stout, elongated to egg-shaped; the bodies of some species contain many rounded projections (tubercles), those of others may be smooth or grooved. The mouth is located on the end of a snoutlike projection, which varies in shape from short and stout to long and slender and sometimes exceeds the length of the rest of the body. Some Anthribidae (fungus weevils), usually cylindrical in shape, have slender antennae that may be longer than the rest of the body; they are easily confused with the cerambycids. Brenthidae (primitive weevils) usually are long and slender with antennae projecting from the sides of the snout. Scolytidae (bark beetles, ambrosia beetles) do not have a distinct snout and are usually cylindrical in shape, as are Platypodidae.

IMPORTANCE

As predators. The Coleoptera are of great economic importance; many species harm plant crops, others are beneficial to man. Predators such as Carabidae (ground beetles) and Staphylinidae (rove beetles) help to control the populations of many insects by feeding on caterpillars and other immature insects (larvae), many soft-bodied adult insects, and insect eggs. Most of the Coccinellidae (ladybugs, ladybird beetles) are highly beneficial to man; both larvae and adults feed on plant-sucking insects (Homoptera) such as aphids and scale insects. Only a few coccinellids (*e.g., Epilachna*) feed on plants.

As plant feeders. Most of the beetles and weevils harmful to man are phytophagous (plant feeders). Of primary importance are the leaf beetles (Chrysomelidae) and the weevils and their relatives (Curculionoidea). Leaf-beetle larvae feed on leaves, stems, or roots of plants; and most adults chew leaves. Weevil larvae or adults feed on almost every plant part; especially numerous are species that bore into trunks, stems, and seeds. Both larval and adult forms of Scolytidae (bark beetles) are serious pests; they feed beneath tree bark, harming vital areas of living trees (*e.g.,* cambium, pine needles, leaf stalks). The Platypodidae have similar habits, but the tropical Brenthidae usually attack deadwood.

The long-horned beetles (Cerambycidae) bore into stems, trunks, roots, and cones of living and dead trees and large semiwoody herbs; adults often feed on tender, new bark. Most pea-weevil larvae (Bruchidae) develop in dried seeds of leguminous plants (peas, beans). Buprestidae (metallic wood borers) have habits similar to those of cerambycids, and many kill trees or branches by boring in the cambium. The scarab beetles (Scarabaeidae) include many important pests of plant crops, lawns, and pastures. The larvae of many Melolonthinae (June beetles, chafers), for example, feed on grass roots. The Dynastinae (rhinoceros, unicorn, and elephant beetles) are often pests of palms, killing them by destroying the growing points. Lumber, furniture, and other items made from wood are often severely damaged by several groups of beetles that bore in dry wood; *e.g.,* powder-post beetles (Lyctidae), deathwatch beetles (Anobiidae), and branch borers (Bostrychidae). Some anobiids also feed on various types of dried products; *e.g.,* the cigarette beetle, *Lasioderma serricorne,* feeds on tobacco, various dried foods, and drugs.

As scavengers. Many groups of beetles function as scavengers, breaking down materials such as dead logs, lumber used in houses (in which case they are pests), dead plant and animal matter, excrement, and other waste products. Coleopterans that function as scavengers include Scarabaeidae, Tenebrionidae (darkling beetles), Silphidae (carrion and burying beetles), and Dermestidae (dermestid or hide beetles). Some dermestid species cause serious damage in museums by feeding on dried animal materials. The larvae of several species of small dermestids damage carpets, upholstery, and clothing. However, some dermestids are valuable as scavengers; some of the carrion-feeding species (*e.g., Dermestes caninus*) are used by zoologists to clean skeletons of animals.

In transmitting plant diseases. Little is known concerning the role of beetles in transmitting plant diseases. Since beetles do not suck plant juices as do plant-sucking insects (Homoptera), there is less likelihood of disease transmission. Transmission of diseases may occur, however, if beetles carry fungal spores on or in the body; *e.g.,* the fungus that causes Dutch elm disease is transmitted by the European elm bark beetle, *Scolytus multistriatus.*

NATURAL HISTORY

Reproduction and life cycle. *General features.* Reproduction is almost always bisexual in beetles, although some species consist of females only, and parthenogenesis (reproduction without fertilization) sometimes occurs. The male reproductive organ is a hardened tubelike structure called the aedeagus. The aedeagus enters a structure in the tip of the abdomen of the female (bursa copulatrix), and the sperm are stored in a saclike structure in the female (spermatheca) until they are needed to fertilize eggs. The females of most species lay eggs. After the larva has hatched, it feeds until its skin (cuticle) becomes too small and splits; the larva crawls out of the old skin (exuviae), and a new skin forms and hardens. The process, called molting, is repeated, usually from three to five times, until the larva is mature. During a nonfeeding period (prepupal stage) the insect enters the pupal stage. The pupa, which forms beneath the final larval skin and emerges when it splits, resembles the adult, except that it is soft and pale;

in addition, the appendages are curled or loosely attached to the body, and the small wings are in flat bags called wing pads. The adult emerges when the thin skin of the pupa is shed; the wings stretch out to full size, and the new outer skeleton hardens and becomes coloured. No further growth of the hardened skeleton occurs; the abdomen of the gravid (pregnant) female may enlarge, however, by stretching of membranes between the abdominal segments. The four developmental stages of Coleoptera—egg, larva, pupa, adult—constitute complete metamorphosis. The length of each stage in the life cycle depends on several factors; *e.g.,* climate, nature of habitat, available food.

Egg-laying sites

Eggs. Eggs vary in form, may be laid singly or in groups, and usually are laid at a site that allows proper development of the larva—on a leaf of a host plant (leaf-eating species), in bark, or in tree trunks (woodborers). Eggs also may be laid near roots, in flowers, in fruits, in tree injuries, on water plants, or under rocks.

Larvae. There are several types of coleopteran larvae (see Figure 41). Carabid larvae have a tapering, flattened,

By courtesy of (*Phyllophaga*) Ohio Agricultural Research and Development Center, Wooster, Ohio; (*Anthrenus scrophulariae*) Cornell University Agricultural Experiment Station, Ithaca, New York; (*Chrysobothris trinerva*) the U.S. Department of Agriculture; (*Oberea tripunctata, Hydrophilus triangularis, Psephenus herricki, Dineutes*) from A. Peterson, *Larvae of Insects*

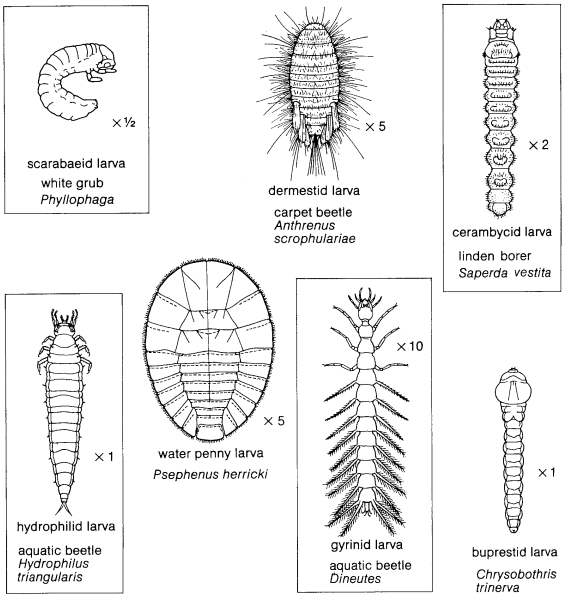


Figure 41: Diversity of beetle larvae.

smooth body, as do those of staphylinids (rove beetles) and silphids (carrion beetles); larvae of the Dytiscidae (diving beetles), although somewhat similar to those of carabids, have a lobed air-float at the end. Larvae of click beetles (Elateridae) are cylindrical or flat and slender and have a hard surface; some click beetle larvae, called wireworms, feed on newly planted seeds and roots of plant crops (*e.g.,* maize, cotton, potatoes); others feed in deadwood or on wood-boring beetle larvae (Cerambycidae). Larvae of Buprestidae (metallic wood borers), which are soft-bodied and slender, bore under the bark of trees or burrow beneath the surface of leaves.

Dermestid larvae, somewhat tapering and cylindrical, have whorls of short bristles and some longer ones and resemble hedgehogs. Coccinellid larvae, flattened, broad in the middle and tapering at the back, sometimes have a few low projections (tubercles) bearing short hairs and are often strikingly coloured with red or yellow and black. Larvae of the plant-feeding epilachnines often are yellow with black bristles. Scarabaeid larvae are soft-bodied, thick, C-shaped, somewhat flattened beneath and round above. Cetonine larvae, similar to those of geotrupids and lucanids, are often short, less bent than most scarabaeids, hairy, active, and capable of locomotion on their backs through movement of body segments. Passalid larvae, white and slender, are found with their parents. Chrysomelid larvae are short, flattish or fat, and sometimes have lobes at the sides or appendages at the hind end. Cerambycid larvae are long and slender, with swellings at the sides of the segments; pale, almost hairless, and fairly soft, they either have minute legs or none at all, and the eyes are poorly developed. Weevil larvae, usually white and soft, are fatter in the abdominal region than at the head end. The head capsule may be hardened, brown in colour, and have strong mandibles.

Some beetles undergo hypermetamorphosis, in which the different larval instars (the stages between molts) represent more than one type of larva. The early larval stages usually are active, and the later stages are parasitic on other organisms. The active, young larvae of most Meloidae (blister beetles), called triungulins, for example, hatch from eggs laid on flowers, become attached to bees visiting the flowers, and thus are carried to a bee nest, where they become parasitic on bee larvae.

Pupae. Pupae of beetles usually have the form of the adult, except that the elytra are represented by pads on the sides; the colour, generally white, is sometimes pale brown or patterned. As the time for emergence of the adult approaches, the pupa may darken, especially on the mandibles and eyes. After emerging from the pupal skin, the adult rapidly assumes its final adult form and coloration, although metallic colours may take some days to develop their final appearance.

Larvae that bore in wood, cones, or seeds and those that live in the ground or in excrement chew or dig a cavity, or pupal cell. In some cases, the pupa lies on a cushion of frass (chewed or torn wood fibres) or other material; in others, it is enclosed in a cocoon of frass or other material (*e.g.,* a smooth, white, hard covering similar to the shell of a bird's egg). Sometimes material is used only to seal off the open end of the tube, gallery, or cell for protection from ants and other predators. After the body of the adult has hardened, the adult breaks or dissolves the barrier and emerges. Many wood borer adults must chew through solid wood, although the larva usually chews close to the surface of the tree, or to the bark, before pupating.

Typical life cycle. The adult *Aspidomorpha furcata,* a tortoise beetle of South China, feeds on the leaves of the host plant *Ipomoea* (sweet potato), where the entire life cycle takes place. Eggs, laid in small groups, are cemented together in a thin paperlike egg capsule whose thin brown layers separate and camouflage them. The larva, which hatches in four to six days, burrows directly through the egg capsule and feeds on the leaf epidermis for about three days before it molts. Later it eats through the entire leaf. When the fifth instar larva is fully grown, it goes through a resting period, the prepupal period, before pupation occurs. During each molt, the old larval skin is pushed back and attached to new processes, which form at the hind end. The dried and shrunken skins plus extruded feces combine to camouflage the larva. When pupation takes place, the combination of exuviae (molted skins) and feces becomes attached to a paired process at the hind end of the pupa, thus camouflaging and shielding it. The last larval skin is used to attach the end of the abdomen of

Egg capsule

By courtesy of the U.S. Department of Agriculture

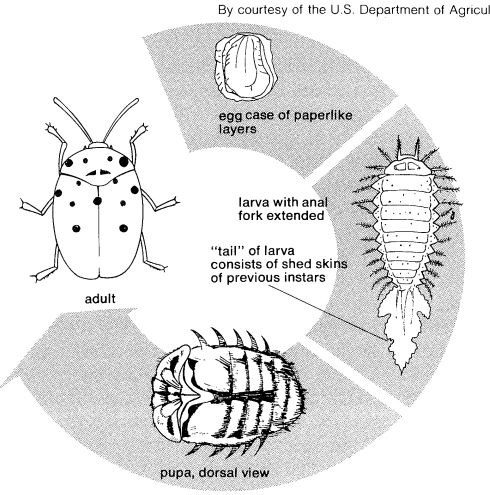


Figure 42: Life cycle of a tortoise beetle.

the pupa to the leaf surface. The pupa usually rests with its camouflage flat over its back, although it may erect the camouflage and turn it back to discourage an enemy.

The pupal stage lasts four days or longer. The life cycle from egg to adult requires 21 to 27 days in mild weather, longer in winter; adults may live more than 230 days. Females lay from 63 to 228 egg cases, with an average of about 3 eggs per case. There may be several generations per year. (See Figure 42 for the stages in the life cycle of a tortoise beetle.)

In cooler temperate areas life cycles may occupy much longer periods, even up to four years or more. In general, wood-boring beetles and root feeders have longer life cycles, leaf-feeding species have short ones, with several generations per year possible with subtropical and tropical species.

Ecology. Ecology, the relationship of organisms to each other and to their environment, represents fundamental interactions in nature. The habits of some of the small families of beetles have not yet been established.

Beetles as prey. In general, beetles are very well armoured insects and thus are reasonably protected against enemies; most, however, have parasites. Some tachinid flies, for example, lay their eggs on adult beetles, and the larvae feed inside their bodies. Beetle larvae also often have hymenopterous parasites; *e.g.,* wasps. Wood-boring long-horned beetle larvae (Cerambycidae) are parasitized by wasps that lay their eggs directly on or in beetle larvae. Beetles are probably attacked by fewer predators than are many other insects; birds that often feed on various kinds of insects may not eat some kinds of beetles. Swifts and other birds, insectivorous mammals including bats, reptiles, frogs, and other insects may act as beetle predators. Some beetle predators feed particularly on beetle larvae, although many beetle larvae that feed on plants and in the ground probably are distasteful to birds and other predators.

Feeding habits and habitats. Beetles are found in almost any habitat occupied by insects and feed on a variety of plant and animal materials. Many are predatory, some are scavengers, many are plant feeders (phytophagous), others feed on fungi, and a few are parasitic on other organisms. Beetles may live beneath the ground, in water, or as commensals in the nests of social insects such as ants and termites. Plant-feeding species may eat foliage, bore in wood or fruit, and attack roots or blossoms; any part of a plant may be a food source for some type of beetle. Many beetles eat stored plant or animal products, including various types of foods and clothing.

Adult carabids are nocturnal hunting predators whose prey ranges from miscellaneous small insects to fairly large caterpillars and land snails. Many small carabids live in moss on tree trunks and feed on small insects hidden in the moss. The predatory larvae remain more hidden than the adults. Both larval and adult dytiscids (diving beetles) feed on various water insects, small frogs, and small fish. Gyrinids (whirligig beetles) feed both on water insects and on other insects that fall into the water. Staphylinids (rove beetles) are usually predatory, both as larvae and as adults. Hydrophilids are water scavenger beetles and eat various dead organisms and live algae in freshwater. Silphids, which dig under small dead animals so that they settle into the ground, lay their eggs on the carrion, on which the larvae feed. Dermestids feed on dead animal skins, dead insects, birds' nests, other animal nests, and accumulations of debris. Ptinidae (spider beetles) feed on dead insects and animal skins, as do certain Anobiidae (*e.g.,* the cigarette beetle, which also feeds on tobacco and other dried products). Heterocerids and histerids prey on fly larvae or those of beetles living in excrement or in carrion. Some Elateridae (click beetles) are predatory as larvae, feeding on the larvae of wood-boring insects; others feed on roots of various crops. Most larvae of Cantharidae (soldier beetles), which prey on worms and larvae of other insects, occur under bark, in rotting wood, or in soil; the adults are usually found on flowers. The lampyrids (fireflies), which are often luminous both as larvae and as adults, are primarily predators upon snails. Clerid larvae are largely predators of wood-boring beetle larvae, although some of the adults are flower feeders.

Dung feeders

Among the lamellicorns, the true scarabs (Scarabaeinae) and several related groups (*e.g.,* Aphodiinae, Geotrupinae) are dung feeders. Some make balls of animal excrement (*e.g.,* of cattle) and roll them to protected spots for burying. Eggs are laid on the buried balls, and the larvae feed on the excrement.

The Melolonthinae (June beetles, chafers) are phytophagous, the larvae usually feeding on roots of grasses or other plants, the adults feeding on leaves. The larvae of Rutelinae (shining leaf chafers), which have similar habits, sometimes feed in humus or in rotten wood. The adults of Cetoniinae feed on pollen and tree juices; the larvae, often called white grubs, feed on organic matter in the soil and may damage plant roots. The Dynastinae (rhinoceros, unicorn, and elephant beetles) feed in rotting wood, decaying vegetation, or humus. Buprestidae bore into living and dead trees, generally feeding on the cambium layer. Some of the small buprestids are leaf miners; that is, they feed on leaf cells between the upper and lower surfaces of a leaf. Moderate numbers of larvae and adults of Passalidae (bess beetles), which feed on rotting logs, generally are found together; apparently the adults protect and feed their young, an unusual habit among beetles. The lucanids feed in logs and stumps as larvae; the adults often feed on juices from damaged trees, particularly at openings of larval tunnels of Cerambycidae (long-horned wood-boring beetles).

Tenebrionidae (darkling beetles) live in various habitats as scavengers or predators. Those in rotten wood and under bark may be predatory on other larvae; those in damp places may feed on rodent excrement, on other wastes, or on fungi; and those in stored products feed on grain, meal, and other staples. Lagriidae (lagriids) and Alleculidae (comb-clawed beetles) feed in rotten logs. Melandryidae (false darkling beetles) usually feed on fungi or in old wood. Pythids usually are scavengers in burrows of other beetles, including weevils. Rhipiphoridae (wedge-shaped beetles), which usually are parasites in wasps' nests and undergo hypermetamorphosis, are related to another group of insects, the Strepsiptera, which are mostly parasitic in the bodies of wasps, leafhoppers, grasshoppers, and other insects. Some larvae of Mordellidae (tumbling flower beetles) may live in dead or dying deciduous wood or attack the heartwood of weak trees; others may be found in pith or herbaceous weeds. The adults frequent flowers and are good fliers. Meloidae (blister beetles) undergo hypermetamorphosis and usually live in bee nests, feeding on eggs and stored food. The larvae of Pyrochroidae (fire-coloured beetles) live under bark and in rotting wood. Most cerambycid larvae are wood borers; a few live in large herbs or in cones, and many feed in roots. The adults, which chew new bark or other plant parts, at times cut rings in branches, killing them so that the larvae can bore in. Some chrysomelid larvae eat leaves; many others are root feeders. Members of one group of chrysomelid larvae (Sagrinae) bore in stems of leguminous vines, causing gall-like swellings; others (Donaciinae) feed on and in the stems of freshwater plants, often below the water level, or hide (Hispinae) in unopened parts of monocot plants such as palms, grasses, bamboos, and gingers; still others (Alticinae) may feed on both roots and leaves or act as leaf miners. Bruchidae live primarily in the seeds of legumes; hence their common name, seed beetles (or pea weevils).

Weevils also are diverse in habits. Many larvae bore into solid wood of living or dead trees and stumps; some feed on or in roots or in stems of semiwoody plants; and others feed in seeds, pods, grain, meal, fruits, nuts, and other parts of plants; some are hidden inside the plants. Most adults are able to fly; nonflying forms, however, are more abundant on small oceanic islands than are flying forms. Some species, which develop in rotting stumps, soil, or palm trunks, make a rough cocoon from frass. Brenthids develop in dead trunks and stumps; anthribids live in various parts of plants, often in seeds; and scolytids feed largely under the bark of living trees, although some bore in other parts of plants, seeds, cones, and needles.

Special ecological relationships. Certain beetles have

special ecological relationships. Beetles that live in nests of ants and in nests of termites have become modified in form because certain structures no longer needed have degenerated (*e.g.,* wings, wing covers); some have almost lost the power of locomotion. Some species have evolved glands that produce secretions attractive to the host ants or termites.

An association of a different type is known involving beetles in the mountains of New Guinea; called epizoic symbiosis, the association occurs on the backs of large leaf-feeding weevils found on *Nothofagus* and other trees in the moss forests. Various kinds of algae, fungi, lichens, liverworts, mosses, and diatoms develop on the backs of the weevils. Among them live protozoans, rotifers, nematodes, phytophagous mites, and parasitic mites. The phytophagous mites, known only from this association, feed primarily in the fungal growth on the backs of the beetles. Plant spores, which may be carried from one weevil to another either by the mites during mating of the weevils or by air dispersal, are trapped with a sticky fluid that may be produced by the weevils for this purpose.

Ambrosia beetles (Scolytidae) associate with fungi in the host tree. Certain adult scolytids and platypodids have specialized structures called mycetangia, which are used to carry the fungi when the beetles seek out new host trees.

FORM AND FUNCTION

Adult features. Many structural modifications are found among the beetles. So varied is the structure that it is difficult to make general statements; for example, a few beetles have no elytra, and some have no wings.

Primary body regions

As in all adult insects, the segmented body consists of three primary body regions: head, thorax, and abdomen (see Figure 43). In beetles, however, two of the three thoracic segments (mesothorax and metathorax) are attached to the abdomen; the third (prothorax), isolated as the region between the head and trunk, is covered by a dorsal plate, the pronotum. The body covering (exoskeleton) varies from very horny and rigid to soft and flexible, but it usually consists of hard plates (sclerites) separated by flexible membranes.

The antennae are usually 11-segmented but vary widely in form. The jaws (mandibles) may be relatively large, as long as the rest of the body, or almost completely absent; usually they are triangular in shape and suitable for biting or chewing. The paired maxillary and labial palps are usually small and are used for feeding or handling food, but in some beetles one or the other pair may be greatly enlarged. The compound eyes are usually prominent, sometimes reduced or absent, and occasionally divided. Simple eyes (ocelli) are rarely present. A neck is sometimes evident, but in many beetles the head is recessed into the prothorax or under the pronotum.

From H. Weber, *Grundriss der Insektenkunde*, 4th ed. (1966); Gustav Fischer Verlag

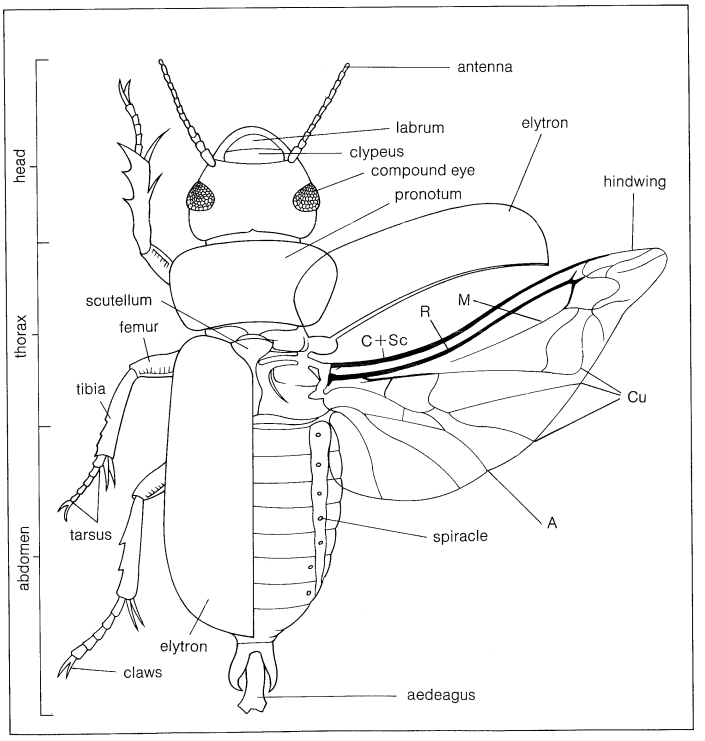


Figure 43: *Coleopteran body plan.*
The wing veins shown (with their abbreviations in parentheses) are anal (A), cubitus (Cu), media (M), radius (R), and costa + subcosta (C + Sc).

The prothorax is generally very distinct; the mesothorax and metathorax are hidden under the elytra along with most of the abdomen. The pronotum may be four-sided very wide or very long, and sculptured with lateral spines or grooves and pits. The front legs emerge from cavities in the underside that may be confluent or separated by other parts. The mesothoracic spiracle (respiratory opening) is often visible just behind the base of the front legs. The mesothorax bears the elytra (wing covers) and the second pair of legs. The metathorax bears the flying wings (hindwings) and the third pair of legs.

The legs are modified in various ways, for running, swimming, jumping, digging, or clasping. In some beetles, wings are not capable of producing flight, but in many others, they are powerful and sustain strong flight.

The abdomen is composed of nine or ten segments, but often some of these are not externally visible. From five to eight segments can usually be seen, with short apical appendages evident in some beetles. Each abdominal segment has a pair of spiracles, the openings into the air-tube (tracheal) system. Predatory beetles generally have short digestive tracts. Differences in salivary glands occur, depending on the food source.

Larval features. Coleoptera larvae differ in appearance from adults. This is characteristic of insects with complete metamorphosis (Endopterygota), in which the wings develop internally until they become apparent in the pupal stage. The differences in body form of the larvae are closely associated with larval habitats and modes of feeding. The numerous predatory types, for example, have slender or gradually tapered bodies, with large, slender, mandibles, and relatively long legs; thus, these larvae are adapted to be rapid runners that capture and hold prey with their mandibles. A mouthpart structure (epipharynx) may be adapted for imbibing body fluids that exude from wounds caused by the mandibles. Typical predatory larvae are found in the dytiscids (predacious water beetles) and carabids (ground beetles). The larvae of tiger beetles (Cicindelidae), although similar to the dytiscid type, live in holes in the ground or in branches and capture passing insects. The enlarged head of the tiger beetle larva fills the burrow opening, and its legs are modified for attachment and leverage. Haliplid, staphylinid, and gyrinid larvae are also similar to those of dytiscids, except that the gyrinids have a gill on the side of each abdominal segment, rather than at the tip of the abdomen, as do most of the others. Some predatory larvae are generally less tapered, sometimes less armoured, and have shorter appendages than do the dytiscids; *e.g.,* larvae predatory on other larvae in deadwood or in the ground. Larvae such as those of the Histeridae (hister beetles), which usually live in special environments such as dung and tunnels in wood, have short appendages but slender mandibles.

Dependency on habitats

Another larval type includes many of the scavengers; *e.g.,* silphids, hydrophilids. These larvae have short legs and mandibles.

Wood-feeding borer larvae (cerambycid and buprestid type) have soft white bodies that may be cylindrical or flattened. The thoracic region of buprestids is flat and broad, the head is dark and retracted in the thorax, and the mandibles are short and stout. Larvae of wood-boring beetles usually have yeast associated with the digestive tract; they help to convert wood to digestible compounds.

Larvae that feed on leaves, stems, and roots (chrysomelid type) are short and oval shaped. Coccinellid larvae, similar in form except for longer legs, however, prey on soft-bodied insects like aphids.

The lamellicorn larval type is C-shaped, has a soft body,

and a hard, dark, nonretractile head. These larvae usually are found in protected habitats, where they feed on roots, rotten wood, or excrement. Weevil larvae have a similar form, although the head may be a little smaller, the body less arched. Another larval type is that of the elaterids and many tenebrionids, which have very slender usually brown bodies, with a hard outer skeleton.

Adaptations. *Protection.* Coleoptera protect themselves against enemies in various ways. Some closely resemble their surroundings; the upper surface of one African species (*Petrognatha gigas*), for example, resembles dead velvety moss, and its irregular antennae are very much like dried tendrils or twigs. Many weevils fall and feign death at the least alarm and, folding their limbs closely around the body, look like seeds or particles of soil, thus escaping observation.

Certain beetles, especially those living in ants' nests, resemble ants, and the common wasp beetle of Europe, *Clytus arietis,* both in its movements and coloration, closely resembles a wasp.

Some beetles obtain some measure of protection possibly from their repellent appearance or from their evil-smelling or distasteful secretions, either in the form of exudations of blood from definite parts of the body or as the product of special fetid glands. The so-called bombardier beetles of the Carabidae have the property of secreting an evil-smelling defensive fluid from the anal end of the body. In some cases, this fluid volatilizes explosively into a gas at high temperature when it comes into contact with the air; it acts as a repellent to other insects or enemies. A number of beetles secure protection by virtue of their agility; many ground beetles and tiger beetles run rapidly, and the latter also take to wing with great readiness. The flea beetles (a group of the Chrysomelidae) have remarkable powers of leaping.

Sound production. Many beetles produce sound, usually by rubbing one part of the body (a scraper) against another part (the file). These stridulating organs are generally present in both sexes and probably serve for mutual sex calling. Some beetles have a filelike area on the head that is rasped by the front margin of the prothorax. Among the cerambycids, sound is produced either by rubbing the rear margin of the prothorax over a grooved area on the mesothorax or by rubbing the femurs of the hindlegs against the margins of the elytra.

Stridulation in larvae

Stridulation, however, is not confined to adult beetles; it occurs also in certain larvae. Some larvae of the Scarabaeoidea, for example, have a series of ridges, or tubercles, on the coxal segment of the middle pair of legs, and the hindlegs are modified in various ways as rasping organs. In the larvae of some chafers (Melolonthinae), a ridged area on the mandible is rasped by a series of teeth on the maxillae. Stridulation in larvae is independent of sex and may be used to warn neighbouring larvae to avoid getting in each other's way.

Light production. Some beetles emit a bright light, whose source is in special luminous organs that consist of an outer, light-producing layer and an inner reflector layer. The outer layer is supplied with oxygen by means of air tubes (tracheae), and the reflector layer contains crystals that apparently act as a background, scattering the light and preventing its dispersion internally. The light is produced during a reaction involving a compound (luciferin) and an enzyme-like substance (luciferase) in the outer layer of the luminous organ (see also BIOLUMINESCENCE). Luminous beetles include the Lampyridae (fireflies), Phengodidae, Drilidae, and certain Elateridae (click beetles). A familiar example of a firefly is the common European glowworm (*Lampyris noctiluca*), whose wingless female emits a bright light near the hind end of the body; the winged male emits a much feebler light. Both sexes of the elaterid genera *Pyrophorus* and *Photophorus* are winged and luminous.

EVOLUTION AND PALEONTOLOGY

Coleopterans are very ancient insects; they date from the Permian Period (about 225,000,000–280,000,000 years ago), after the appearance of gymnosperm plants. Although the beetles have a number of similarities to another ancient group of insects, the cockroaches (Blattaria), they probably evolved from ancestors of the present-day Neuroptera. This theory is based largely on the nature of the life cycle of beetles and on their larval structure. Although many beetle fossils are known, they consist mostly of isolated elytra, which reveal little about the history of the order. Complete fossil specimens are closely related to living forms. The evolution of elytra may have been associated with the habit of living under bark of trees, where protection for flying wings is required. Most of the insects that live under bark are beetles.

CLASSIFICATION

Distinguishing taxonomic features. One distinctive feature of coleopterans is wing structure. Most beetles have two pairs of wings. The front pair, which may be thickened, leathery, or hard and brittle, are called elytra and usually serve only as protective covers. A few beetles have greatly reduced wings. Variations in the structure of the first abdominal segment is one criterion used to separate the various suborders of Coleoptera; the hindcoxal leg segments (by which the legs are attached to the body) may divide the abdominal segment partially or completely. Sometimes the abdominal segments are fused, the articulations marked by form sutures.

Variation in length, texture, and appearance of elytra, as well as the number of abdominal segments exposed by short elytra, are used to distinguish the various superfamilies. Characters associated with the size and shape of the coxae also are used as distinguishing features. Structure of antennae and legs are important considerations for taxonomic criteria, as are larval structure, head structure (including mandibles, or jaws), pattern of veins in wings (see Figure 43), habitats, and behaviour.

Annotated classification. About 135 families of beetles are known, of which 120 occur in the Western Hemisphere.

There have been a number of different classifications of Coleoptera. Many were based on the suborders Adephaga and Polyphaga; the latter, which contains about 90 percent of the beetles, included a number of divisions (*e.g.,* clavicorns, serricorns, lamellicorns, phytophagous beetles, and weevils). Sometimes these divisions are considered as superfamilies or series, and sometimes (particularly weevils and relatives) as suborders. The classification below is based on that of R.A. Crowson (1955); it includes four suborders.

ORDER COLEOPTERA (beetles, weevils)

Largest insect order; more than 250,000 species; size range from less than 1 mm to more than 12 cm (5 in.); modified front wings, called elytra, usually meet in a straight line down the middle of the back, covering membranous hindwings; hindwings usually longer than front wings, folded under front wings when at rest; mouthparts adapted for chewing; form of antennae variable; large compound eyes; hard outer skeleton; complete metamorphosis; found in almost all types of habitats; many plant feeders; many species of economic importance, either cause damage or benefit man; worldwide distribution.

Suborder Archostemata

Hindcoxae rarely fused to metasternum; distinct notopleural suture between notum and pleural sclerites.

Family Cupesidae (*Cupedidae;* reticulated beetles). Small and little-known; found under bark; about 20 species widely distributed.

Family Micromalthidae. Rare; 1 to 2 species; most complex life cycle among coleopterans.

Suborder Adephaga

Larval structure primitive; legs specialized for predatory life; hindcoxae of legs immovable fixed to metasternum; distinct notopleural suture between notum and pleural sclerites; wing with base of Rs (radial sector) vein distinct.

Family Cicindelidae (tiger beetles). Voracious and fierce, especially larvae; *Cicindela;* active; often brightly coloured; about 2,000 species; mostly tropical and subtropical.

Family Rhysodidae (wrinkled bark beetles). Small, slender, brownish beetles; about 100 species, mostly tropical.

Family Paussidae. Sometimes a subfamily of Carabidae; about 400 species; tropical or subtropical; associated with ants.

Family Carabidae (ground beetles). Usually dark, shiny, flattened; larvae and adults predatory; *Calosoma* feed on caterpillars; *Brachinus,* bombardier beetles, eject fluid from anus; about 20,000 species; worldwide distribution.

Family Trachypachidae. A few species in Europe and North America.

Family Haliplidae (crawling water beetles). About 200 small aquatic species; wide geographical range.

Family Amphizoidae (trout-stream beetles). About 5 species (*Amphizoa*) in Tibet, North America; feed on drowned insects.

Family Hygrobiidae. A few species (*Hygrobia*) widely distributed; aquatic; produce sound.

Family Gyrinidae (whirligig beetles). About 700 species; surface swimmers; sometimes gregarious.

Family Noteridae (burrowing water beetles). Similar to Dytiscidae; small; larvae burrow.

Family Dytiscidae (true water beetles). Worldwide distribution; about 4,000 species; found in flowing and still water; *Siettitia* eyeless; *Dytiscus* best known.

Suborder Myxophaga

Wing with base of Rs vein absent; prothorax usually with distinct notopleural suture.

Family Calyptomeridae. Sometimes placed in Clambidae.

Family Lepiceridae. A few Central American species.

Family Sphaeriidae (minute bog beetles). Less than 1 mm in length; 1 genus; a few widespread species; sometimes placed in Staphylinoidea.

Family Hydroscaphidae (skiff beetles). Size about 1.5 mm; found in algae on rocks in streams; sometimes placed in Staphylinoidea; generic example: *Hydroscapha;* widely distributed.

Suborder Polyphaga

Includes the majority of beetles; wing with base of Rs vein absent; prothorax never with distinct notopleural suture.

Superfamily Hydrophiloidea

Head usually with Y-shaped line on front; antennae short, hairy and club-shaped at end; habits mostly aquatic; maxillary palp usually longer than antennae.

Family Hydrophilidae (water scavenger beetles). About 2,000 species, numerous in tropics; aquatic and terrestrial; most live on decomposing vegetable matter; *Hydrophilus* is predatory; other examples: *Cercyon, Helophorus.*

Superfamily Histeroidea

Antennae geniculate (elbow-shaped) with last 3 segments club-shaped; wing with medio-cubital loop reduced; elytron truncate leaving 1 or 2 segments of abdomen exposed.

Family Sphaeritidae. One genus, about 4 species.

Family Synteliidae. One genus, a few species in Mexico and the Orient.

Family Histeridae (hister beetles). Small, dark, shiny; found in decaying organic matter; predatory on small insects; about 2,500 species; wide distribution; examples: *Hister, Niponius.*

Superfamily Staphylinoidea

Very large group; antennae with last 3 segments rarely club-shaped; outer skeleton rarely very hard, shiny; wing veins M (media) and Cu (cubitus) not connected; elytron truncate, usually more than 2 abdominal segments exposed.

Family Limulodidae (horseshoe crab beetles). One mm or less; in United States and Australia; ride on ants; 5 genera, about 30 species.

Family Ptiliidae (feather-winged beetles). Among the smallest beetles; live in dung, rotting wood, fungi; about 350 species; temperate and neotropical regions.

Family Leptinidae (mammal-nest beetles). Feed on eggs and young of small arthropods in small-mammal nests; wingless; widely distributed.

Family Anisotomidae (*Leiodidae;* round fungus beetles). Small, shiny; habitats vary—caves, fungi, mammal nests; about 1,000 species.

Family Dasyceridae. About 10 species; Holarctic.

Family Silphidae (carrion beetles, burying beetles). Relatively large, bright-coloured; usually feed on carrion; some predatory, some plant feeders; examples: *Silpha, Nicrophorus;* about 230 species; widely distributed.

Family Scydmaenidae (antlike stone beetles). Under stones, logs; in ant nests; very small, hairy; widely distributed; about 1,200 species; example: *Scydmaenus.*

Family Scaphidiidae (shining fungus beetles). In fungi, dead leaves, rotting wood; small, very shiny; about 250 species; example; *Scaphidium.*

Family Staphylinidae (rove beetles). Short elytra; size variable; active; strong mandibles; in ant nests, predatory on other insects in decaying matter; about 20,000 species; widely distributed; examples: *Stenus, Dinarda.*

Family Pselaphidae (short-winged mold beetles). Very small; diverse in form; live in ant nests; about 3,500 species; worldwide distribution but most abundant in tropics; example: *Claviger.*

Superfamily Scarabaeoidea (Lamellicornia)

Antennae 10-segmented with last 3 to 7 segments forming a lamellate (platelike) club; body stout; larvae without cerci (appendages at end of abdomen); males and females often differ in appearance; outgrowths on head and thorax produce bizarre forms; produce sound (stridulate).

Family Lucanidae (stag beetles, pinching bugs). Large; variable in size; males with enormous mandibles (jaws); about 900 species; widely distributed; example: *Lucanus.*

Family Passalidae (bess beetles). Large, dark, flattened; about 800 species, mostly in moist, warm forests; live in decaying wood.

Family Trogidae (skin beetles). About 150 widely distributed species; example: *Trox;* dung or carrion feeders.

Family Acanthoceridae. About 100 tropical species; associated with rotten wood.

Family Geotrupidae ("dor" beetles). Large; about 300 species; widely distributed; habits variable; examples: *Lethrus, Geotrupes.*

Family Scarabaeidae (scarab beetles, June beetles, tumblebugs, leaf-chafers). Variable in colour, size, habits; most feed on dung, carrion, other decaying matter; about 20,000 species; widely distributed; examples: *Cetonia, Melolontha.*

Superfamily Dascilloidea

Forecoxae projecting; abdomen with 5 visible segments; wing with radial cell short; anal cell of wing, if present, with 1 apical vein.

Family Clambidae (fringed-wing beetles). Small, hairy; in decaying plant material; about 30 species; worldwide distribution; sometimes placed in Staphylinoidea.

Family Eucinetidae. About 25 widely distributed species; in rotten wood; example: *Eucinetus.*

Family Helodidae (marsh beetles). Small, oval; on vegetation in swampy places; aquatic larvae; about 350 species; widely distributed; example: *Scirtes.*

Family Dascillidae. About 200 moderate-sized species; found on vegetation in moist places.

Superfamily Byrrhoidea

Forecoxae large; antennae more or less thickened at tip; body short, with legs and antennae retractable into grooves on under surface.

Family Byrrhidae (pill beetles). Small, oval; found under debris, in sand, at grass roots; about 270 species; widely distributed; example: *Byrrhus.*

Superfamily Dryopoidea

Forecoxae sometimes slightly projecting; antennae usually filiform (threadlike) or short and broad; head with distinct labrum (upper lip); last tarsal segment of leg often as long as preceding 3 or 4 together.

Family Psephenidae (water-penny beetles). Larvae flat, almost circular; a few species, mostly in India, North America.

Family Eurypogonidae. A few species in Northern Asia, North America.

Family Ptilodactylidae. About 200 tropical species; aquatic or in rotten wood.

Family Chelonariidae. About 50 species in tropics of Asia and America.

Family Heteroceridae (variegated mud-loving beetles). About 500 widely distributed species; example: *Heterocerus.*

Family Limnichidae (minute marsh-loving beetles). Similar to Dryopidae; a few widely distributed species.

Family Dryopidae (long-toed water beetles). Small, downy; crawl on stream bottoms; few species; widely distributed.

Family Elmidae (riffle beetles). Varied habitat; several hundred widely distributed species.

Superfamily Rhipiceroidea

Antennae flabellate (fanlike); noselike projection between mandibles; about 180 species; widely distributed; 2 families, Rhipiceridae (cedar beetles), Callirhipidae; example: *Sandalus.*

Superfamily Buprestoidea

Antenna short, serrate; abdomen weakly hardened. *Family Buprestidae* (metallic wood-boring beetles). Brightly coloured, metallic sheen; inhabit various hot, moist forests; about 15,000 species, mostly tropical; examples: *Agrilus, Sphenoptera, Chrysobothris.*

Superfamily Elateroidea

Forecoxae small; metasternum without transverse suture; larvae with no free labrum.

Family Cerophytidae. About 12 species in Europe and America; in hollow trees.

Family Cebrionidae. About 200 species; in mild regions; female often wingless.

Family Elateridae (click beetles). About 7,000 species; widely distributed; can leap when lying on back; adults, plant feeders; larvae sometimes damage plants; examples: *Pyrophorus, Agriotes, Athous.*

Family Trixagidae (*Throscidae;* throscid beetles). Small, oblong; about 200 species; widely distributed.

Family Perothopidae. Medium-sized; on trunks or branches of old beech trees.

Family Eucnemidae (false click beetles). Closely related to Elateridae; about 1,000 species, mostly in warm climates; example: *Melasis.*

Superfamily Cantharoidea

Larvae with grooved or channelled mandibles; front coxae large, projecting; 6 or 7 visible abdominal segments.

Family Brachypsectridae. A few species in Asia and California.

Family Homalisidae. A few species in Mediterranean region.

Family Karumiidae. A few species in Iran, South America.

Family Drilidae. About 80 species, mainly in Europe; larvae prey on snails.

Family Phengodidae. About 50 species in America; produce light.

Family Lampyridae (lightning bugs, fireflies). Produce light in species-characteristic flashing rhythm; wingless females and most larvae called glowworms; about 1,000 species; widely distributed; examples: *Lampyris, Photinus.*

Family Cantharidae (soldier beetles). Soft-bodied, predatory; about 3,500 species; widely distributed; examples: *Cantharis, Rhagonycha.*

Family Lycidae (net-winged beetles). About 2,800 species, mostly tropical; often bright-coloured; distasteful to birds; example: *Dulitocola.*

Superfamily Dermestoidea

Last 3 antennal segments forming a club; tarsi of legs all 5 segmented, simple.

Family Derodontidae. About 12 widely distributed species.

Family Nosodendridae. About 30 widely distributed species; found under bark.

Family Dermestidae (skin beetles, dermestid beetles). Many economically important species; mostly scavengers on plant and animal products; small to moderate-sized; hairy or with scales; examples: *Dermestes, Anthrenus;* widely distributed.

Family Thorictidae. About 80 species, commonly in Mediterranean region; associate with ants.

Superfamily Bostrychoidea

Larvae soft-bodied, lack specialized setae (hairs), maintain a C-shaped position; adult hard, head region hoodlike; members often associated with timber, destructive.

Family Anobiidae (drugstore and death-watch beetles). Live in dry vegetable materials; some species destructive pests; examples: *Xestobium, Stegobium, Lasioderma;* about 1,100 widely distributed species.

Family Ptinidae (spider beetles). Long legs; spiderlike appearance; sometimes infest stored products; about 500 widely distributed species.

Family Bostrychidae (branch and twig borers). Attack living and deadwood; about 550 species; worldwide distribution; examples: *Sinoxylon, Dinoderus.*

Family Lyctidae (powder-post beetles). Related to Bostrychidae; damage timber and furniture; about 90 widely distributed species.

Family Psoidae (psoid branch and twig beetles). Medium-sized; sometimes cause extensive orchard damage; examples: *Psoa, Polycaon.*

Superfamily Cleroidea

Tarsi of legs always 5-segmented; forecoxae projecting or transverse; abdomen with 5 or 6 visible segments.

Family Trogossitidae (bark-gnawing beetles). About 500 species, mostly tropical; vary in shape and habits; sometimes in stored products; example: *Tenebroides.*

Family Chaetosomatidae. Three genera in New Zealand.

Family Cleridae (checkered beetles). Small; many brightly coloured; downy; most adults and larvae predatory on other insects; some adults pollen feeders; about 3,000 species, mainly tropical; examples: *Corynetes, Necrobia.*

Family Melyridae (soft-winged flower beetles). About 4,000 species widely distributed; diverse; example: *Malachius.*

Family Phloiophilidae. Rare; 1 species in Britain.

Superfamily Lymexyloidea

Antennae short, more or less serrate; abdomen with 6 or 7 visible segments.

Family Lymexylidae (ship-timber beetles). About 50 species; worldwide distribution; damage wood; examples: *Lymexylon, Hylecoetus.*

Superfamily Cucujoidea

Usually 5 visible abdominal segments; antennae filiform or clubbed, rarely serrate.

Section Clavicornia.

All visible abdominal segments usually movable; sometimes less than 7 pairs of abdominal spiracles (respiratory holes)

Family Passandridae. Few species; mostly in warm climates.

Family Smicripidae. Sometimes placed in Nitidulidae; a few species in tropical America; example: *Smicrips.*

Family Nitidulidae (sap beetles). Variable size, shape, habits; usually found around fermenting plant fluids or moldy plant materials; about 2,200 species; examples: *Meligethes, Cybocephalus.*

Family Byturidae (fruitworm beetles). Small, hairy; few genera; damage raspberry blossoms and fruit; example: *Byturus.*

Family Biphyllidae (silken fungus beetle). About 200 species; mostly tropical; example: *Biphyllus.*

Family Rhizophagidae (root-eating beetles). Usually occur under bark; few species; sometimes placed in Nitidulidae.

Family Cisidae (minute tree-fungus beetles). Occur under bark, in wood, or in dry woody fungi; about 360 species; widely distributed.

Family Protocucujidae. Two species; Chile and Australia; similar to Sphindidae.

Family Sphindidae (dry-fungus beetles). Small, dark; occur in dry fungi; about 30 species; widely distributed.

Family Cucujidae (flat bark beetles). Flat, medium-sized; predatory on mites and other insects; some species hypermetamorphic; example: *Catogenus.*

Family Silvanidae (flat grain beetles). Closely related to Cucujidae; some feed on grain (*Oryzaephilus*); another genus, *Silvanus.*

Family Hypocopridae. Little known; few species.

Family Helotidae. About 80 species in warm parts of Asia.

Family Phycosecidae. Few species; examples: *Phycosecis, Alfieriella;* in Australia, Asia, Africa.

Family Propalticidae. About 20 species in warm regions of Old World.

Family Cryptophagidae (silken fungus beetles). Mostly fungus feeders; sometimes in nests of bees and wasps; about 800 species; examples: *Cryptophagus, Antherophagus.*

Family Phalacridae (shining flower beetles). Larvae develop in certain flower heads (*e.g.,* goldenrod), about 500 species; widely distributed; example: *Olibrus.*

Family Languriidae. Feed on plant leaves and stems; example: *Languria;* about 400 species mostly in Asia and North America.

Family Erotylidae (pleasing fungus beetles). Shiny; found with fungi; about 1,600 species; many in South America.

Family Cerylonidae. Often placed in Colydiidae; few species.

Family Coccinellidae (ladybird beetles, ladybugs). Many predatory on aphids and coccids, a few serious plant pests (*Epilachna*); mostly beneficial; about 5,000 species, usually bright-coloured, spotted; widely distributed; another genus, *Rodolia.*

Family Corylophidae. About 300 species; widely distributed; minute in size.

Family Endomychidae (handsome fungus beetles). Shiny, usually brightly coloured; feed on fungi (mold); about 600 species; mostly in tropical forests; examples: *Endomychus, Mycetaea.*

Family Discolomidae. About 30 tropical species; many wingless.

Family Merophysiidae. Few species, sometimes placed in Lathridiidae.

Family Lathridiidae (minute brown scavenger beetles). Found in fungi, debris, flowers; about 600 species.

Section Heteromera.

Abdomen with 7 pairs of spiracles; forecoxae usually projecting.

Family Merycidae. A few Australasian species; example: *Meryx;* affinities uncertain.

Family Mycetophagidae (hairy fungus beetles). Mostly associated with fungi; often brightly marked; about 200 species.

Family Colydiidae (cylindrical bark beetles). More than 1,400 species; especially in tropics, but many in New Zealand; diverse; example: *Bothrideres.*

Family Pterogeniidae. Two Indo-Malayan genera of uncertain affinities.

Family Nilionidae. About 40 species, mostly in tropics; associate with fungi on trees.

Family Lagriidae. Medium-sized with metallic sheen; found in foliage, on bark.

Family Tenebrionidae (darkling beetles). Varied group; mostly plant scavengers; examples: *Eleodes, Tenebrio;* about 10,000 species; widely distributed.

Family Alleculidae (*Cistelidae;* comb-clawed beetles). Found on flowers, foliage, fungi, under dead bark; 100 species; worldwide distribution; example: *Omophlus.*

Family Monommidae. Adults occur in leaf litter; larvae

in rotten wood; about 100 species in warm regions; example: *Monomma.*

Family Zopheridae. Few species, mostly in America.

Family Elacatidae (false tiger beetles). Widely distributed; umerous species; examples: *Othnius, Eurystethus.*

Family Boridae. Widely distributed small group; sometimes placed in Tenebrionidae.

Family Inopeplidae. One genus, *Inopeplus;* related to Salpingidae.

Family Salpingidae (narrow-waisted bark beetles). Superficial resemblance to Carabidae (ground beetles); adults and larvae predatory; adults occur under rocks, or bark, in leaf litter, on vegetation; few species but widely distributed; examples: *Salpingus, Lissodema.*

Family Cononotidae. Similar to Salpingidae; several genera.

Family Mycteridae. Resemble Salpingidae.

Family Hemipeplidae. Few species in warm regions; live beneath palm leaves.

Family Trictenotomidae. About 12 species in forests of Oriental region.

Family Pythidae. Few species widely distributed in Eurasia and America; example: *Pytho.*

Family Pyrochroidae (fire-coloured beetles). Adults large, found on foliage or flowers, under bark; about 100 species in north temperate region; example: *Pyrochroa.*

Family Tetratomidae. Similar to Melandryidae.

Family Melandryidae (false darkling beetles). Usually found under bark or logs; examples: *Penthe, Osphya;* about 400 species in woodlands of temperate regions.

Family Scraptiidae. About 200 species widely distributed; associated with rotten wood, fungi; example: *Scraptia.*

Family Mordellidae (tumbling flower beetles). Wedge-shaped, hump-backed; common on flowers; active; about 650 species.

Family Rhipiphoridae (wedge-shaped beetles). About 400 species, many with specialized parasitic habits on other insects; complicated life cycle; examples: *Pelecotoma, Metoecus.*

Family Cephaloidae (false longhorn beetles). About 11 species in east Asia, North America.

Family Meloidae (blister beetles, oil beetles). Body fluids contain cantharadin, sometimes used as a drug (*Lytta*); several important plant pests (*Epicauta*); many larvae beneficial, feed on grasshopper eggs; hypermetamorphic; complicated life cycle; about 2,000 species; widely distributed.

Family Oedemeridae (false tiger beetles). Adults usually on flowers or foliage; larvae in moist decaying wood; about 600 species; widely distributed, but especially abundant in temperate regions; example: *Nacerdes.*

Family Aderidae (antlike leaf beetles). About 350 species; usually found in deadwood or vegetable refuse; example: *Aderus.*

Family Anthicidae (antlike flower beetles). Many occur in vegetable refuse; about 1,000 species; sometimes placed in Pedilidae; examples: *Anthicus, Notoxus.*

Superfamily Chrysomeloidea

Mostly wood or plant feeders; body shape very variable; antennae not clubbed.

Family Cerambycidae (long-horned, wood-boring beetles). Some large in size; plant feeders; many brightly coloured; larval stage usually wood-boring, sometimes cause tree damage; about 25,000 species; worldwide distribution; examples: *Macrotoma, Titanus, Clytus, Monochamus.*

Family Chrysomelidae (leaf beetles). Closely related to Cerambycidae; larvae usually plant feeders; many serious pest species; overwinter as adults; more than 26,000 species; widely distributed.

Family Bruchidae (seed beetles, pea weevils). Most larvae live in leguminous seeds; examples: *Acanthoscelides, Bruchus;* damage stored seeds; about 900 species; widely distributed.

Superfamily Curculionoidea (snout beetles)

One of the largest and most highly evolved groups of coleopterans; head prolonged into beak or snout; mouthparts small; antennae usually clubbed and geniculate; larvae C-shaped; mostly plant feeders; of economic importance as pests.

Family Anthribidae (fungus weevils). Found in dead wood and fungi; about 2,400 species, mostly in tropics; *Brachytarsus* predatory on scale insects.

Family Nemonychidae (pine-flower snout beetles). Small group sometimes placed in Curculionidae or Attelabidae.

Family Belidae. Small group found in Australia, New Zealand, South America attached to a variety of plants.

Family Oxycorynidae. Small group; found in South America (*Oxycorynus*), East Indies (*Metrioxena*), tropical America (*Allocorynus*).

Family Aglycyderidae (*Proterrhinidae*). About 200 species; *Proterrhinus* found in Pacific region; one *Aglycyderes* species in Canary Islands.

Family Attelabidae (leaf-rolling weevils). Form leaf rolls on various trees; moderate number of species; widely distributed.

Family Brenthidae. About 1,700 species, mostly in wooded tropical countries; variable size range; males unlike females in structure.

Family Apionidae. Moderate number of species; *Apion, Cylas;* several species injure leguminous crops.

Family Curculionidae (weevils). About 30,000 species, many with scales; many injurious species; worldwide distribution; *Anthonomus* (cotton boll weevil, apple blossom weevil); *Calandra* (granary weevil, rice weevil); *Sitona* species pests of leguminous crops.

Family Scolytidae (bark beetles). Numerous species; damage trees; *Scolytus* (elm bark beetle); other genera: *Ips; Dendroctonus; Hylastinus.*

Family Platypodidae (pin-hole borers or flat-footed ambrosia beetles). Seldom attack healthy trees; example: *Platypus;* fungus food called ambrosia.

Critical appraisal. There are many different opinions among coleopterists concerning the relationships of the various groups of beetles; the groups that should be given family status; and the placement of families in superfamilies and suborders. Little information is available about many coleopteran groups, so that their taxonomic affinities are uncertain. Although Crowson includes Strepsiptera (twisted-wing parasites) as a superfamily of Coleoptera, opinion is divided concerning their status, and they have not been included here (see also *Annotated classification,* above, in section entitled *Insects*). Some classifications of Coleoptera recognize three suborders, Archostemmata, Adephaga, and Polyphaga, rather than the four used by Crowson. (J.L.G.)

Hymenoptera (ants, bees, sawflies, wasps, and allies)

The members of the order Hymenoptera comprise the third largest and perhaps the most beneficial to man of all insect groups. More than 110,000 species have been described, including ants, bees, ichneumons, chalcids, sawflies, wasps, and lesser known types. Except in the polar regions, they are abundant in most habitats, particularly in tropical and subtropical regions.

Collectively, the Hymenoptera are most important to man as pollinators of wild and cultivated flowering plants, as parasites of destructive insects, and as makers of honey. The Hymenoptera are divided into two suborders: Symphyta (mainly sawflies and horntails) and Apocrita (wasps, ants, bees, and most parasitic forms).

The order includes the best known of the social insects—ants and some species of bees and wasps. Most species, however, are solitary in habit. Hymenopterans may be parasitic or nonparasitic; carnivorous, phytophagous, or omnivorous.

GENERAL FEATURES

Hymenopterans are chiefly small to medium-sized insects, usually with four membranous wings and a narrow waist that sets off the abdomen from the the thorax, or body midregion. The mouthparts are of the biting type or of the biting–sucking type. In the higher evolutionary forms—bees, for example—mouthparts are modified into a sucking apparatus. The ovipositor, or egg-laying organ in the female, is often very long and may be modified for piercing, sawing, or stinging. Metamorphosis is complete—*i.e.,* the insect develops through four distinct stages: egg, larva, pupa, and adult. Sex is usually determined by whether or not an egg is fertilized; fertilized eggs develop into females; unfertilized ones become males.

Hymenoptera display an array of interesting behavioral characteristics, particularly in social species. The dance-like movements of honeybees communicate information from one individual to another about the location, distance, quantity, and quality of a particular food source; such movements have been carefully studied by the German entomologist Karl von Frisch. Trophallaxis, or the mutual exchange of food between larvae and adults of bees, ants, and wasps, has been of special interest to hymenopterists. Hyperparasitism—the parasitic habit of one species upon another parasitic species—has also attracted

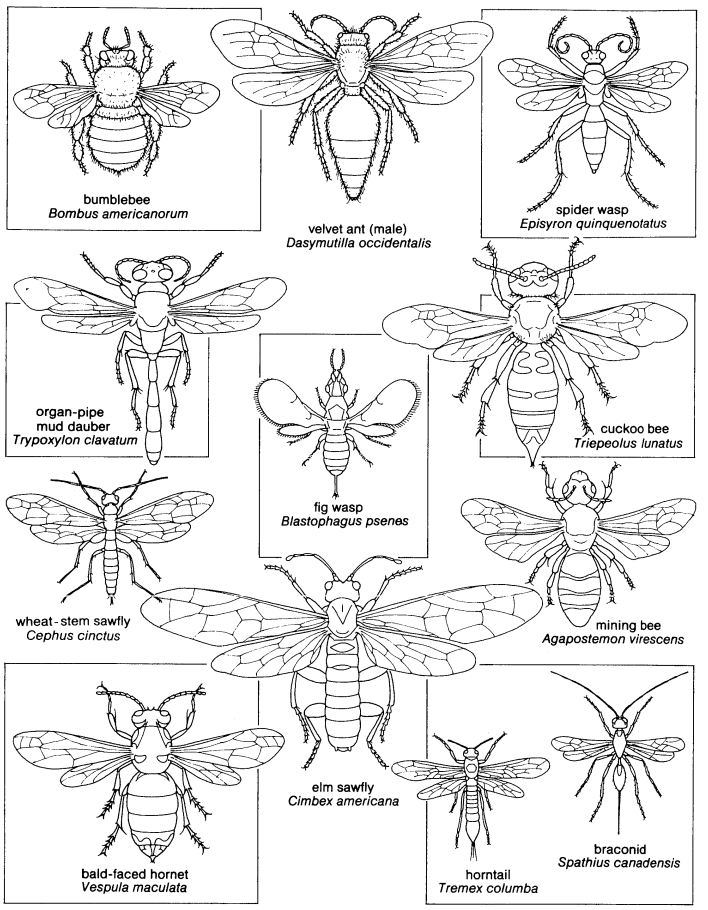


Figure 44: Representative Hymenoptera.

By courtesy of (elm sawfly, wheat stem sawfly) U.S. Department of Agriculture; (fig wasp) California Agricultural Experiment Station; (braconid) Connecticut Agricultural Experiment Station; from (horntail) *Invertebrate Identification Manual* by Richard A. Pimentel, © 1967 by Litton Educational Publishing, Inc., reprinted by permission of Van Nostrand Reinhold Company; (others) *An Introduction to the Study of Insects*, revised edition, by Donald J. Borror and Dwight M. DeLong, © 1964 by Holt, Rinehart and Winston, Inc., © 1954 by Donald J. Borror and Dwight M. DeLong, reprinted by permission of Holt, Rinehart and Winston, Inc.

attention. Polyembryony, the development of many individuals (as many as 1,000) from a single egg, is an unusual phenomenon occurring in some members of the families Chalcididae and Proctotrupidae. Parthenogenesis (production of young by females that are not fertilized by males) also occurs in some forms.

Honeydew

Certain ants are remarkable for their relationship with insects that provide so-called honeydew or other sweet fluids. Among these cooperating insects are scale insects, aphids, leafhoppers, rove beetles, and the caterpillars (larvae) of many butterflies (Lepidoptera). The so-called honey ant (*Myrmecocystus* in the United States, *Plagiolepis* in Africa) has in the nest a division of worker ants known as repletes, which are fed sugary secretions; as a consequence of high food intake, the abdomens of repletes swell into globules up to one centimetre (about 0.4 inch) in diameter. Ants that obtain sweet fluids from the caterpillars of blue butterflies (Lycaenidae) reciprocate by allowing the caterpillars to devour ant larvae.

Hymenoptera, even those equipped with a sting, are sought as food by other animals. Skunks, badgers, field mice, shrews, and other animals attack bee nests for the insects as well as for the honey. The larvae of the wax month (*Galleria mellonella*) live in bee nests, where they eat beeswax, thus damaging the nest. The cuckoo bee (Anthophorinae), a close relative of the bumblebee, lays its eggs in bumblebee nests, where the larvae are cared for and nourished by bumblebee workers.

Size range and diversity of structure. Hymenopterans range in size from the smallest fairyflies (Mymaridae), which are about 0.21 millimetre (about 0.008 inch) in length, to the largest of the Pelecinidae, which may exceed five centimetres (about two inches) in length. Most members are less than 2.5 centimetres (1 inch) long.

Principal differences in structure within the order include the presence or absence of wings (when present, there are two pairs); the presence or absence of a "waist"; modifications of the ovipositor; and adaptations of mouthparts for specific eating habits. Polymorphism—*i.e.*, the occurrence of two or more forms of a species—is highly developed in some of the social Hymenoptera. A honeybee colony, for example, contains at least one queen, plus workers and drones; each form is structurally and physiologically different from the others.

Distribution and abundance. Bees and wasps, as the most significant agents for the pollination of flowers, are found virtually everywhere that flowering plants occur. Social species may live in colonies of as many as 1,000,000 individuals. Ants are mostly heat-loving insects and are consequently most numerous in tropical and subtropical regions, both in numbers and in species. The number of ant species in arctic or Alpine regions is extremely small. The group appears to be absent over great areas north of the Arctic Circle.

IMPORTANCE

The honeybee has been valued since pre-Christian times for its honey and beeswax. Beekeeping in modern times has become a lucrative and highly developed enterprise. Royal jelly, produced by honeybee workers, has enjoyed some popularity as a cosmetic, although its beneficial properties in this respect have not been satisfactorily demonstrated.

Beneficial parasites

Certain parasitic forms are valuable control agents against insect pests. Notable among these are the parasitic wood wasps, parasitic on wood-boring beetles; braconids, parasitic on the European corn borer, other Lepidoptera, and wood-boring beetles; ichneumons, parasitic on various Lepidoptera and wood-boring beetles; eulophids, parasitic on scale insects; pteromalids, parasitic on several crop pests; chalcids and trichogrammatids, parasitic on a variety of orchard pests; and tiphiids, parasitic on the Japanese beetle. Fig insects (Agaonidae) are valuable as the only pollinators of the Smyrna fig, an important crop in the Western United States.

Relatively few Hymenoptera species are serious economic pests. Chief among these, however, are the wheatstem sawfly (Cephidae); some seed chalcids, also pests of wheat; the larch sawfly (*Pristiphora erichsonii*), which destroyed much of the larch forests in Britain and North America late in the 19th century; and the European spruce sawfly (*Gilpinia*), once a serious pest in North America. In order to control the European spruce sawfly, parasitic wasps were introduced from Europe.

The fire ant (*Solenopsis saevissima*), introduced into the United States from South America, feeds on young plants and seeds and is known to attack young mammals. The destructive habit of legionary ants, or army ants (Dorylinae), is of particular importance in South America. Armies of as many as 1,500,000 such insects destroy almost all animal life they encounter. Leaf-cutting ants (*Atta*) are serious pests, especially in Brazil, where they may ravage extensive plantings of cultivated plants overnight. These insects are eaten locally in South America.

NATURAL HISTORY

Reproduction. The ovary of the queen bee is composed of several hundred ovarioles, each of which contains about 60 eggs and so-called nutrition cells. The so-called spermatheca, a sperm reservoir that collects sperm from the male in the course of several matings, connects with the oviduct, through which eggs are carried to the outside. The sperm can remain alive and viable in the fluid medium of the spermatheca for several years.

Facultative sex determination

When an egg passes down the oviduct, it may or may not be fertilized by emerging sperm, according to the "discretion" of the female. Fertilization occurs if the female relaxes a muscular ring around the sperm duct, thus allowing the duct to open and sperm to pass through. Since unfertilized eggs result in males and fertilized eggs result in females, the queen determines the sex of the offspring by relaxing or closing the muscular ring.

In addition to this so-called facultative sex determination, hymenopteran reproduction is often remarkable in other

respects. An extreme and pronounced dimorphism, for example, often exists between the two sexes. Some male and female hymenopterans once, in fact, were thought to represent different species. The consequence of such extreme difference is that the two sexes, with markedly different ways of life, also have different tasks in the insect community. As a rule, only the female cares for the brood; the male lives only briefly, just long enough to mate. This dimorphism also involves the senses. Female bees, for instance, are sensitive to yellow, blue-green, blue, and ultraviolet light. The drones (males), however, are blind to yellow but are particularly sensitive to ultraviolet light. The yellow colour of many flowers is of no significance to the drones since they cannot feed themselves. Sunlight, however, which contains much ultraviolet light, is important in orienting the insect during the mating flight.

Parthenogenesis

Parthenogenesis, which occurs in many insect orders, is particularly common in the Hymenoptera; it occurs in three forms: arrhenotoky, thelytoky, and deuterotoky. In arrhenotoky, male honeybees are produced from unfertilized eggs laid by mated (impregnated) females or by so-called second, or supplementary, queen bees, which have not been impregnated. In thelytoky, which occurs in many species of the suborder Symphyta, unimpregnated females produce males. In deuterotoky, unimpregnated females of some Symphyta species produce females as well as males. The occurrence of these forms is not always mutually exclusive; it has been found in *Apis,* for example, that about 1 percent of the eggs laid by second queen bees may be female.

Parthenogenetic reproduction in Hymenoptera often occurs in a cycle of alternate generations. In chalcids, the first generation, consisting of males and females, produces a second generation of females, which lay unfertilized eggs; from these are produced both males and females, thus repeating the first generation. In some gall wasps (Cynipidae) no males have yet been observed. The females of these wasps invariably produce thelytokous females.

The first generation of the braconid *Microtonus brevicollis* is parthenogenetic and parasitizes the adult form of the beetle *Haltica amphelophaga.* The second generation, however, lives in the larvae of the same beetle, and the females are impregnated by males. The occurrence of parthenogenesis is determined by a nutritional or hormonal factor in the larva of the host. The mode of reproduction is also determined by the number of chromosomes (*i.e.,* cell structures that contain hereditary information) in the egg.

Polyembryony

Polyembryony, the development of more than one individual from a single egg, occurs in numerous parasitic Hymenoptera, including chalcids, encyrtids, proctotrupoids, braconids, and dryinids. In this type of reproduction, the embryo divides into several separate, identical parts at an early stage; each part then develops into an individual insect. In *Litomatix truncatellus,* which is parasitic on the larvae of the noctuid moth *Plusia gamma,* about 1,000 embryos develop from one egg. The offspring from one egg are all male or all female; as a rule, several eggs are laid at one time, so that both sexes will occur in one brood. It is believed that polyembryony is controlled by internal factors, either genetic or chemical, a view supported by the fact that in some species (Platygasteridae) only part of a brood may occur polyembryonically.

Life cycle. *Solitary forms.* The solitary wasps, of which there are about 20,000 species, usually build nests or cells, which they provision with permanently paralyzed insects or spiders. One egg is laid in each cell and the body of the paralyzed host provides all the nutriment needed by the developing larva. Some solitary wasps nest in woody or pithy stems; others dig tunnels in the soil; still others, the mason or potter wasps, construct nests of mud, which are sometimes vaselike or juglike and may be found attached to twigs or other objects.

Sawflies, which are phytophagous, lay their eggs in incisions in plants, cut by the sawlike blades of the ovipositor. Species of several genera, notably *Pontania,* induce the growth of galls on willows. In many species of sawflies in which males are unknown, parthenogenesis occurs.

The spider wasps (Pompilidae) usually construct single cells in the ground, in rotten wood, or in rock crevices and provision them with paralyzed spiders. The thread-waisted wasps (Sphecidae) have diverse nesting habits. Most of them nest in the ground and use leafhoppers, treehoppers, cicadas, stinkbugs, bees, winged ants, beetles, or caterpillars as food for the young, each species or group confining itself to one type of prey. Mud daubers (*Sceliphron, Chalybion*) build small nests of mud, often in attics, outbuildings, or eaves, and store them with the bodies of paralyzed spiders.

The females of the superfamily Ichneumonoidea deposit their eggs in or on the larvae or pupae (rarely eggs or adults) of the host species; the legless, maggot-like larvae that hatch from these eggs feed on the body fats and fluids of the host until fully grown. At this time, the larvae usually spin silken cocoons, within which they pupate and from which the adult parasites eventually emerge. Those species parasitizing exposed hosts usually develop as internal parasites; those attacking hosts concealed in wood burrows, plant stems, cocoons, or leaf mines feed externally. In the case of internal parasites, the hosts feed and behave normally until shortly before the parasitic larvae have completed their development; the hosts of the external feeders are paralyzed by the female parasites. In most cases, one parasitic larva completes its development within or upon its host; in some species, however, many larvae develop in one host.

Adults of ichneumon flies feed principally on honeydew secreted by aphids and related insects or on host juices that exude from punctures made by the ovipositor. Most cuckoo wasps (Chrysididae) lay their eggs in the nests of solitary wasps or bees. Among the exceptions is an African species that is parasitic on the tsetse fly. Tiphiidae and Scoliidae are mostly parasites of beetle grubs that live in the soil. The female wasp digs into the soil to locate the grub, stings and paralyzes it, and deposits an egg on it. The wasp larva lives on the outside of the grub.

Solitary bees

Most species of the superfamily Apoidea are solitary bees; that is, each female makes her own nest (usually a burrow in the ground) and provisions it. Among such bees there are no castes (worker, queen). The mother bee constructs cells, each of which is an enclosed space provided with a supply of pollen and honey, lays an egg in each cell, then closes it, and goes on to build and provision another.

Most solitary bees are short-lived as adults. Some species may be in flight only a few weeks, having spent the rest of the year in their cells as eggs, larvae, pupae, and young adults. Other species have several generations yearly, so that adults may be found more or less continually. In temperate climates, solitary bees usually pass the winter in their cells, either as mature larvae (prepupae) or as young adults.

Several families of bees have evolved species that are called parasitic because they lay their eggs in cells of various working bees. The young larva, which usually has a large head and large jaws, destroys the egg or young larva of the host, then eats the provisions of the host. Such bees not only lack the scopa (*i.e.,* pollen-collecting apparatus) and other structures associated with collecting and carrying pollen but also lack the various structures of the jaws and legs that other species use for making nests.

Social forms. In several families of Hymenoptera the larvae are completely dependent upon the continuous care of the adults. The colony is a family community of which every insect is an integral unit. Apart from the community he cannot properly function or survive. The essential work in the society—namely, nest building, feeding and tending the brood, and defense of the nest—is performed by the female workers. The fertile female, the queen, performs only one task: egg laying. The workers can be differentiated morphologically and physiologically as soldiers, outside workers, inside workers, and nest builders. The males play no part in everyday nest activities. They live only for a short time at a specific time of year, occur in limited numbers, and are virtual parasites of the colony that must feed them.

Conditions of social behaviour

The activities of certain solitary bees of the subfamily Halictinae are helpful in understanding certain aspects of the evolution of the highly organized hymenopteran

societies. The females of *Halictus quadricinctus* survive the hatching of their own offspring. Mother and daughter stay together in the same nest, which consists of single brood cells; thus, although each female takes care of her own cells, they build and defend the nest together. In *Augochloropsis sparsalis* there is a further development: the females of one nest stay together for one summer, and a division of labour occurs. Some of the young females return, still unmated, from the nuptial flight; they then take charge of the gathering of pollen and nectar and further building of the nest and are called worker bees. The mated females merely deposit eggs.

The social behaviour of *Halictus* (*Evylaeus*) *malachurus* has advanced another step; morphological differences are apparent between the ovipositing female and the assisting females. The latter are poorly fed as larvae; they are, as a result, smaller and their sexual organs are poorly developed. In the bumblebee colony, different castes are clearly established. Social companionship and division of labour have become "obligatory": although the mated queen can build a nest independently and raise the first worker bees herself, she needs the assistance of workers to help rear sexually mature bees.

Similar preliminary stages to a social organization are found in the Vespidae. The female *Stenogaster depressigaster* passes several generations in the communal nest, and the daughters build their own cells and care for their own offspring. In the case of *Belenogaster,* however, whose nests include about 60 cells, the females not only feed their own brood but also indiscriminately feed all larvae present. Trophallaxis, or exchange of food between workers and larvae, is a further development.

The first evolutionary step toward a division of labour occurs in *Polybia.* Some female *Polybia* only lay eggs; the gonads of the assisting females are poorly developed. These females take charge of repair and construction, larvae care, and food gathering. They thus are useful in the society, even though they produce no offspring. This society lasts at least one summer and possibly as long as several years before it dissolves, and young sexually mature insects establish a new nest in spring.

"Immortality" of colony

The life of the honeybee colony is potentially endless. Because the queen's honey-collecting apparatus and her pharyngeal glands and wax glands are degenerated, she is incapable of building a nest or feeding and tending the brood. The continued survival of the colony results from the fact that young queens replace the old. After their nuptial flight young queens return to the home nest. If, during the spring, many offspring develop, the colony population greatly increases, and the number of cells in the comb for developing young is no longer adequate. The colony then divides by swarming, during which the old queen leaves the nest with about half of the worker bees, and the old nest is relinquished to a newly hatched queen. The swarm finds a new nesting place and builds new combs. To make the task easier, all the departing workers are provided with honey.

From an illustration by James Gordon Irving from *Zoology*, a Golden Science Guide; Copyright 1958 by Western Publishing Company, Inc.

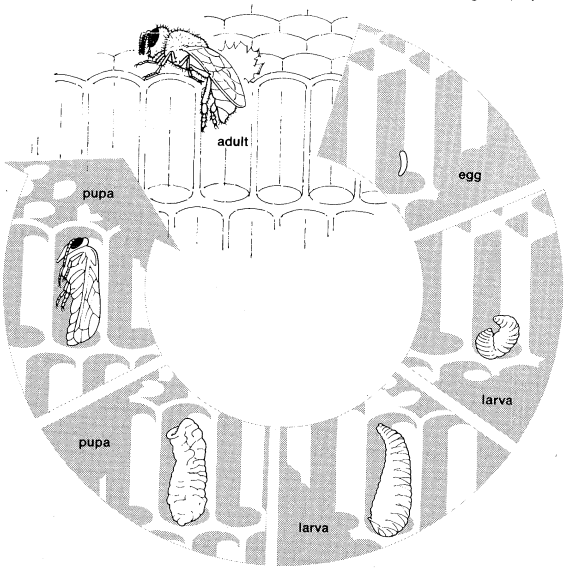


Figure 45: Life cycle of the honeybee.

For nest building the Polistinae (paper wasps) and Vespinae (*e.g.,* yellow jackets) use paperlike coverings, which they construct by gnawing wood particles from structures—such as fences, telephone poles, and barn doors—then kneading them together with saliva to form a little ball. After returning to the nest, insects roll the balls into layers of paper-thin cell walls. Vespinae also build the outer nest covering in several layers to help modulate the interior temperature, the air between the layers of paper serving as insulation. Interior warmth is produced by the body heat of the larvae and by the constant contraction of the muscles of the adults. This agitation, like human shivering, generates heat; as a result, the temperature of the hive is raised. If the nest becomes overheated, drops of water are carried in. Temperature regulation of the nest is so precise that on warm days a constant temperature of 35° C (95° F) can be maintained. An equally precise temperature regulation occurs in the hive of the honeybee.

All ant species are social in habit. The virgin queen ant, who is usually winged, mates in flight with only one male. During the flight he transfers to her seminal receptacle all the sperm she will require for the rest of her life—which may be as long as 15 years. Each fertilized queen is immediately capable of establishing or taking over a nest. After the mating flight has ended, she first seeks a place to raise her brood; then her wings drop off, and the bulky wing muscles degenerate. The substances composing the muscle tissue are processed into nutritive materials.

As soon as the wings have fallen the ovaries become functional, and egg laying begins. In primitive species the queen leaves the nest and forages for food for the larvae. In more advanced forms, the queen rarely leaves the nest. She feeds so-called nutrition eggs or other food stores within her own body to the first brood. The larvae that survive in the nest develop into dwarf workers, which forage outside the nest for food to nourish additional larvae.

A few ant species (*e.g., Atta*) have developed colonies that live for long periods, one queen succeeding another. In *Eciton* the young queens, which are wingless, mate in the nest and are dependent upon the help of the workers as the nest is being built.

Division of labour. The degree of social organization in a Hymenoptera colony is most evident in the division of labour. In honeybee colonies the division of labour is achieved in an especially interesting manner. Tasks are assigned according to age. The first day after the bee's emergence as an adult, the first tasks of the female worker are to carry out wastes, to clean the cells, and to line them with a disinfectant secretion preparatory to deposition of the egg. About three days later, this young bee advances to brood nursing—*i.e.,* she provides the older larvae with honey and pollen. On the sixth day she also nurses the young larvae with specific larval food from her pharyngeal glands. After about ten days more the bee becomes active in building, secreting wax and using it to make the comb. Soon afterward, she makes her first orientation flight outside the hive. About the twentieth day she begins serving as an entrance guard; finally she becomes a collector bee, remaining at this job until her death. This activity sequence parallels certain physiological changes in the bee. The pharyngeal glands start to secrete about the third or fourth day. The wax glands become active about the tenth day. By the time the bee moves to service outside the hive, the pharyngeal glands as well as the wax glands have degenerated.

The life of a worker bee

In the division of labour among some ant forms highly specialized types of polymorphism have been developed. The *Cryptocercus* ants, for example, make nests in hollow stems of plants, then bore a circular entrance that remains under constant surveillance by special guards whose heads are modified into pluglike structures that fit the entrance. Each guard is relieved after several hours and another guard takes its place. Nest members who wish to enter indicate that they are members by means of a special

movement of the antenna. Entrance guards are useless for other tasks. Repletes (see above *General features*) are a special worker caste of the honey ant. When food is plentiful, these workers are fed so much that the size of the abdomen is greatly increased. Unable to walk, they hang as living honey jugs from the ceiling of the nest, to be used as a food source when fresh food is scarce.

Social parasitism. The very close relationship between insects that are social parasites of the Hymenoptera and their hosts is made possible by their highly differentiated behaviour and by a special secretion of the parasite. The beetles *Lomechusa* and *Atemeles* (Staphylinidae) enter unmolested into the nests of ants, such as *Formica* and *Myrmica,* and are fed by the hosts. The beetle larvae are adopted and reared together with the ant brood. The adults and larvae of the beetle pacify their hosts by means of the secretion, which induces the ants to care for them.

Communication. The highly integrated activities of the Hymenoptera colony require sophisticated methods of passing information among its members. The so-called dance of the honeybee is perhaps the most remarkable demonstration of methods of communication in insects.

After a bee has discovered a new source of food, she returns fully loaded to the comb, delivers nectar and pollen, and notifies the other bees about the new food source before returning to the field, communicating information about the location and quality of the food source by means of various dancelike movements. Information about the plant species is conveyed by the odour of the flower, which adheres to the bee's body. The other bees sense this information through their antennae. But information about the quality and quantity of the food source is conveyed by the liveliness and duration of the dance movements of the bee. If the food source is unusually rich and of high quality, certain sounds are also made to convey this information. The location of the food source is indicated by the rhythm of the dance and by the orientation of the axis of the tail with respect to gravity. If the food source is near the hive, a "round" dance is performed. A "tail-wagging" dance indicates that the food source is more

The bee dance

From K. Von Frisch, *Bienenfibel*

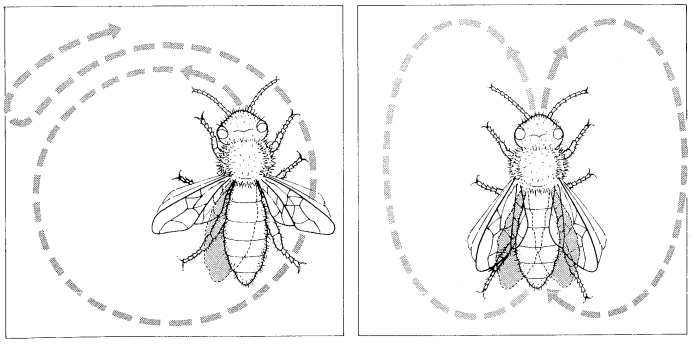


Figure 46: *Dance movements of the honeybee.*
(Left) Round dance and (Right) tail-wagging dance of the honeybee.

than 80 metres (260 feet) away. This dance transmits precise information about direction as well as distance. The number of dance cycles performed by the bee in a certain length of time is inversely related to the distance of the food source. Thus, about 10 cycles are performed every 15 seconds for a food source 100 metres (330 feet) away, but only one cycle is made in that period if the food source is 10,000 metres (33,000 feet) away. The bee measures the distance in terms of how much energy she expends in travelling to the food source. The sun and gravity are used in conveying directional information. During the flight to the food source the bee determines the angle between the line of flight and the sun. The angle to the vertical at which she then dances on the vertical face of the honeycomb describes the angle between the line of flight to the food source and a line drawn in the direction of the sun. An upward tail-wagging run means: "The flight is toward the sun." A downward run means: "The flight is away from the sun." A run 45° to the left indicates that the source is 45° to the left of the sun. Because the position of the sun changes during the day, the dance angle must also change in the course of the day; and this actually occurs. If the sun is concealed behind the clouds or behind some obstruction such as a mountain or a group of trees, the bee analyzes the pattern of polarized light coming from the sky. If only a small opening of sky appears between the clouds, it reveals to the bee's eye a typical pattern that travels along with the sun. A particular intensity of polarized light is generated toward the earth's surface from every point in the sky. The sun's position is determined by virtue of the bee's sensitivity to polarization differences.

The tail-wagging dance is also performed when the swarm is searching for a new dwelling. When a swarm accompanies its queen from the hive, it gathers first in the immediate neighbourhood of the mother hive. Scouts fly out in all directions looking for a suitable nesting place. When a suitable site is found, the scouts return to the cluster and announce the site by means of the dance, and the swarm then moves en masse.

Ecology. As the principal insect pollinators of flowering plants, the Hymenoptera have played a vital ecological role ever since the two groups evolved. The mutual dependency of many species of bees and wasps, on the one hand, and flowering plants, on the other, is firmly established. Indeed, many plants cannot reproduce without the helpful intervention of a particular insect species, most often a hymenopteran.

Parasitism, which occurs among most families of Hymenoptera, is usually apparent to none but the interested student; yet, the ecological significance of parasitism within the order might well overshadow that of pollination. Hymenoptera, the most prevalent and successful of insect parasites, exert a profound, if subtle, control over populations of other insects and certain other arthropods—groups that might otherwise overpopulate and thus upset their particular ecosystem. Man has utilized this control mechanism to his own advantage by importing, breeding, and maintaining many species of Hymenoptera parasites that prey upon insect pests.

Parasitic control of insect populations

FORM AND FUNCTION

Adult features. *External structure.* Concomitant with diversity of habit in the adult Hymenoptera is a diversity of form. This variety prevails to such a degree that only the briefest, general description applies to the order as a whole.

As in all adult insects, the segmented body consists of three primary body regions: head, thorax, and abdomen. In most forms a narrow constriction at the anterior (front) end of the abdomen separates that section markedly from the thorax. Two pairs of membranous wings are usually present. The vein pattern in the wings is usually reduced, and, in some forms, veins are entirely absent. The hindwings, noticeably smaller than the forewings, are interlocked with the latter by tiny hooklets on the anterior margin. In certain solitary wasps, particularly the wasp family called velvet ants (Mutillidae), the females are wingless, and worker ants lack wings. Winglessness also occurs in scattered genera of ichneumons and chalcids—usually only in females, sometimes in both sexes, but rarely in the males alone.

The mouthparts are usually modified for biting or for biting and sucking. The compound eyes (*i.e.,* consisting of many mosaic-like facets) are large. There are usually three ocelli, or simple eyes, arranged in a triangle on the top of the head. The antennae vary greatly in form. Rarely are they shorter than the head is wide. Usually they are moderately long, sometimes longer than the body, and composed of many segments. Often the basal segment, or scape, is greatly elongated; sometimes the segments near the tip are modified into a club. Sometimes the segments are branched. The leg is nearly always characterized by five segments, of which the fifth is the tarsus, or "foot." The abdomen of the female has an ovipositor at the tip. In the sawflies, the ovipositor is modified into a sawlike tool used for making slits in the leaves or stems of plants in which the eggs are deposited, but in all other Hymenoptera it is modified for stinging or piercing.

Internal structure. The form of the digestive system in

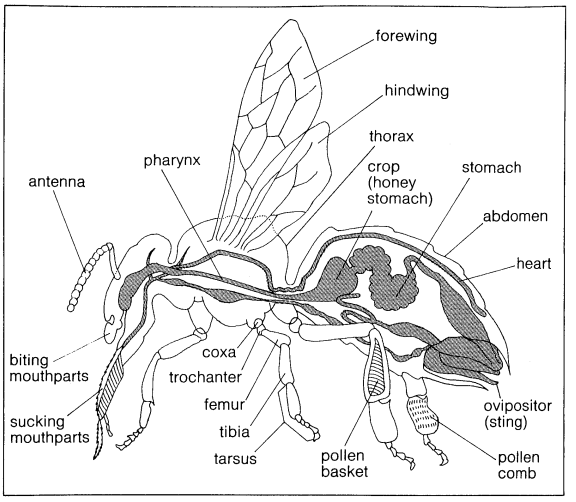


Figure 47: Body plan of a honeybee.
From H. Weber, *Grundriss der Insektenkunde*, 4th ed.; Gustav Fischer Verlag, Stuttgart

Hymenoptera is relatively uniform throughout the order. In ants, an infrabuccal chamber located under the mouth has as its apparent purpose the trapping of indigestible particles that have been ingested along with food. The solid residue is regurgitated as a pellet.

The crop, or honey stomach

In stinging forms the esophagus enlarges near the stomach into a crop, or honey stomach, which serves as a reservoir for liquids to be later regurgitated. In honey ant repletes, the crop may be greatly distended. In honeybees, it may contain as much as 75 milligrams (0.003 ounce) of nectar—about one-third the insect's total weight. In bees and wasps, the stomach, or ventriculus, is the largest part of the digestive system; in most ants, solitary wasps, and other forms, however, it is quite small.

The two pairs of salivary glands are well developed, particularly in bees. One pair is found in the head, the other in the thorax; the ducts leading from them unite to form a single canal that passes into the pharynx. Drones and queen bees also have a mass of salivary gland cells in the head near the ocelli. Worker bees have one pair of pharyngeal glands that produce food, especially royal jelly, for the young larvae. Larvae that develop into workers are fed secretion from the pharyngeal glands only during the first three days after hatching from the egg. Afterward this food is mixed with pollen and honey. The pharyngeal glands are rudimentary in drones and absent in queens.

The function of the mandibular glands, opening near the inner angle of the mandible, is not fully understood. In gall wasps, however, their secretion is known to cause an abnormal growth of cells in leaves. Among social forms, this gland in the queen produces the substance that inhibits ovary development in workers and another that attracts males during her mating flight.

The so-called Nassonow gland, opening on the dorsal side of the abdomen, produces a substance that is used to mark the entrance to the bee hive as well as food sources away from the hive. Honeybees, bumblebees, stingless bees, and many solitary bees have wax glands on the sternites (ventral body plates). The wax is used in the construction of combs—*i.e.,* of brood cells and cells for the storage of pollen and honey.

Features of immature stages. *Egg and larva.* The eggs of parasitic forms are often attached to a surface by means of a pedicel, or stalk. In some forms, the pedicel may be five or six times the length of the egg itself.

The larva typically has a distinct head region, three thoracic segments, and usually nine or ten abdominal segments. In the suborder Symphyta, the larvae are usually caterpillar-like. The head covering is especially tough and the mouthparts powerfully developed. There are usually three pairs of legs on the thorax and six or eight on the abdomen. Symphyta larvae that are wood borers or stem borers have no abdominal legs and the thoracic legs are smaller than those of nonborers.

With rare exceptions larvae of the suborder Apocrita have no legs and are maggot-like in form. The head covering is softer and thinner than in the Symphyta. In parasitic forms, the head is often greatly reduced and partially withdrawn into the prothorax, or anterior part of the thorax. Sense organs appear to be poorly developed. There are no ocelli. Antennae are absent or very small. The mandibles are toothlike, sickle-like, or spinelike. In most Apocrita larvae the stomach is a blind sac until the final larval stage, when it opens into the intestine. The larvae of stinging forms (Aculeata) generally have ten pairs of spiracles, or breathing pores; in parasitic forms nine pairs are usually present.

Pupa. In the Apocrita, the final stage, the prepupa, begins to show certain adult features—*e.g.,* adult wings and legs. The prothoracic segment has begun to distend because of the growing head. The first abdominal segment, or propodium, becomes part of the thorax. The pupa is exarate; *i.e.,* the developing adult appendages (legs and wings) are separate from the body rather than molded into its surface.

A cocoon is usually formed. It may be parchment-like in texture or made of soil particles. In stinging forms, it is usually a thin, silken lining within the larval cell. Certain ants form no cocoon whatever.

The cocoon

Special adaptations. A great variety of structural adaptations have evolved in hymenopterans. Several are of particular interest. In worker bees hairs on the tarsi of the forelegs are used to brush pollen from flowers. The tarsi of the forelegs and middle legs are used to brush pollen from the hairs of the body of the bee. During the flight from one blossom to another, the collected pollen is passed to the pollen-carrying organs, which vary among different kinds of bees.

Some primitive bees (*e.g.,* in the families Colletidae and Halictidae) have masses of long hairs on the basal segments (coxae, trochanters, femurs) of the hind legs and on the undersurface of the abdomen. These hairs constitute the scopa, or pollen-bearing structure. In many colletids and halictids, the scopa is limited to the hind legs. In two subfamilies, Panurginae and Anthophorinae, the scopa is enlarged on the fourth segment (tibiae) of the hind legs and reduced or absent on the abdomen and on the basal leg segments. In the social bees, the scopa is limited to the outer sides of the hind tibiae, where it consists of long hairs surrounding a smooth area, the entire structure being called a pollen basket, or corbicula. In leaf-cutting bees (Megachilidae), the scopa is limited to the underside of the abdomen. In some colletids (Hyaleinae), the scopa is absent, and the pollen, mixed with nectar, is carried to the nest in the crop (anterior end of the digestive tract).

Special senses. Reproduction, the search for food, and, in the case of social species, coordinated group activity require highly developed sensory and orientative capability. In these respects the Hymenoptera are the most advanced of the insects.

Optical. It has been demonstrated that the eye of the bee is sensitive to ultraviolet radiation. It is blind to red light, however. White flowers, which only partially reflect ultraviolet as a rule, appear coloured to a bee. Certain colour combinations, while invisible to the human eye, are not only visible but also may be of special interest to bees. In response to certain colour combinations, they instinctively extend the proboscis (feeding organ) into the flower for nectar.

Some flowers that appear entirely yellow to the human eye reflect ultraviolet from the outer ends of the petals. Only the inner part, then, appears yellow to the bee, directing it to the nectar source.

The facetted compound eye of the insect is apparently unable to perceive forms in sharp outline, and it seems to be the patterns of flowers that attract bees rather than their geometric shapes. In addition, striking landmarks serve for orientation from the hive to the food source.

Chemical. Recognition of a member of the same species and society and discovery of a mate and of the nest are largely determined by smell. Guards at the beehive entrance test every bee who seeks to enter to determine, by smell, if she belongs to the colony. Each colony has its specific odour.

Scent marks

Ants use scent marks, which they place on their pathways. They are thus able to find their way back to the nest and direct other colony members to a food source. When danger threatens, ants, wasps, and bees secrete an alarm substance. This marks the place of danger and notifies other colony members to be on the alert.

During the nuptial flight of the bees, the queen announces her presence to the drones by releasing a substance from her mandibular gland. If this secretion is obtained experimentally and sent aloft by means of a balloon, a swarm of drones will gather around it. A queen bee constantly identifies herself in the hive by means of the so-called queen substance, which also originates from the mandibular gland. This secretion, passed by certain workers in minute portions to all hive inmates, inhibits ovary development. It is also perceived by smelling and controls certain types of behaviour. As long as the odour is present in the hive, the bees are unable to produce new queen cells. As soon as the queen disappears or is ill, with the consequence that the substance is no longer produced, substitute queens are immediately bred from young larvae. The sense of smell is also essential to many parasitic forms in their search for a host. Some parasites can detect, chemically, whether another parasite has already laid an egg in an intended host. If such is the case, no egg will be deposited by the second visitor.

Mechanical. Nest building and mutual communication are aided in numerous ways by mechanical means of orientation. Wasp and bee nests, which have horizontal and vertical combs, as well as ant nests are designed with respect to gravity. Highly developed gravity organs have been found in bees and ants. Hair tufts on the neck and in the leg joints sense the response of the various body parts to gravity. In this way the individual parts of the body assist the organs of balance, and the brain is informed of each change in position. In ants the hair tufts on the basal antennae segment also are used for the perception of gravity.

PALEONTOLOGY AND CLASSIFICATION

Evolution. According to S.I. Malyshev, a Soviet entomologist, the first hymenopterans appeared early in the Mesozoic Era (about 225,000,000 years ago)—about the same time as the first butterflies, moths, and flies. It is his thesis that the Hymenoptera derived from the so-called Eumecoptera—ancestors of the modern scorpion fly (order Mecoptera), the first insects to undergo complete metamorphosis. Another expert fixes the appearance of the first Hymenoptera in the middle of the Jurassic Period (150,000,000 years ago). So-called protohymenopterans, found in Permian beds (250,000,000 years old) in Kansas, have been regarded by some paleoentomologists as ancestral to the modern order. These resembled modern sawflies in having forewings and hindwings of about equal size and in lacking marginal hooklets for joining the two pairs.

True sawflies (Tenthredinoidea) are also extant from the Jurassic Period. Because of the many longitudinal veins in the wings of these forms, they were believed to share a common origin with cockroaches, which have wings that exhibit a similar vein pattern.

Many fossil ants are known from the Early Tertiary Period (60,000,000 years ago), both from Europe and North America. Some of them have been assigned to living genera. Males, females, and workers were already clearly differentiated in ants at that time.

Appearance of stinging forms

The aculeates, or stinging Hymenoptera, were one of the most recent large groups of insects to evolve. By the Tertiary Period both parasitic forms and aculeates had become abundant. The first bees, according to the fossil record, appeared in the Miocene Epoch (26,000,000 years ago). These were leaf-cutting bees (*Lithargus*). Since many of the flowering plants depend upon bees for pollination, it is believed that such plants and bees evolved at about the same time.

Classification. *Annotated classification.* Some disagreement on the taxonomic structure of the order Hymenoptera exists among systematists. For many years it was customary to separate the suborder Apocrita into two subdivisions: the stinging forms (Aculeata) and the parasitic forms (Parasitica). So many exceptions to such a dichotomy were encountered, however, that it has been generally discredited. Biologically, the basis for such a separation is slim. Many forms assigned to the Parasitica are phytophagous, and a number of the Aculeata are parasites. In the generic, or nontaxonomic, sense the term aculeate still applies to the stinging forms.

The classification given below is based on that of Borror and DeLong (1964), which, in turn, is essentially that of Muesebeck *et al.* (1951) and Krombein *et al.* (1958). It covers 71 families, of which 28 are relatively scarce.

ORDER HYMENOPTERA (chalcids, ichneumons, sawflies, ants, wasps, and bees)

One of the largest insect orders; over 110,000 described species; size range from about 0.21 mm (0.008 in.) to about 5 cm (2 in.) in length; usually 4 membranous wings, hind pair smaller than front pair; wings with relatively few veins; mouthparts modified for chewing or for chewing and sucking; in some forms, especially bees, certain mouthparts (labium and maxillae) form a structure for sucking liquid food; antennae usually with 10 or more segments; in higher forms the ovipositor is modified into a sting; complete metamorphosis; larvae usually maggot-like (*i.e.,* legless); compound eyes large, usually 3 simple eyes (ocelli) present; worldwide in many types of habitat; many beneficial to man, including those that pollinate flowers, make honey and beeswax, or parasitize insect pests; many forms have a complex social organization.

Suborder Symphyta (sawflies and horntails)

The oldest hymenopterans, Jurassic to present; all fliers. Larvae mostly plant eaters, usually caterpillar-like, with variable number of legs; some serious pests of trees and shrubs; in adults, thorax joined broadly to abdomen; ovipositor well-developed.

Superfamily Megalodontoidea (primitive sawflies)

A rather scarce group.

Family Xyelidae (xyelid sawflies). Mostly less than 10 mm long; ovipositor long to very long; larvae feed on hickory, on elm, and on pine flowers.

Family Pamphiliidae (web-spinning and leaf-rolling sawflies). Stout-bodied; usually shorter than 15 mm; ovipositor short. Larvae sometimes gregarious (living in groups) in rolled leaves or webs.

Superfamily Tenthredinoidea (sawflies)

Ovipositor sawlike in shape and function.

Family Pergidae (pergid sawflies). Antennae six-segmented; larvae eat oak and hickory leaves.

Family Argidae (argid sawflies). Stout-bodied; more than 400 species, distributed worldwide; mostly black or dark; larvae feed on various trees and herbs.

Family Cimbicidae (cimbicid sawflies). Stout, often large and bumblebee-like; antennae clubbed; larvae often partly curled and covered with waxy powder.

Family Diprionidae (conifer sawflies). Includes serious pests of conifers; 13 or more antennal segments.

Family Tenthredinidae (typical sawflies). About 4,000 known species; adults wasplike, often brightly coloured; usually shorter than 20 mm; larvae feed chiefly on leaves of trees and shrubs, some highly destructive.

Superfamily Siricoidea (horntails and wood wasps)

Mostly medium to large insects.

Family Syntexidae (cedar-wood wasps). In United States represented by 1 California species; larvae eat wood of incense cedar.

Family Siricidae (horntails). Usually 2.5 cm or more in length; often black and yellow or metallic blue; female with long ovipositor; larvae wood borers, usually in trees past their prime.

Family Xiphydriidae (wood wasps). About 12–20 mm long; larvae are borers in dead and decaying deciduous trees (those that annually lose leaves).

Family Orussidae (parasitic wood wasps). A small, rare group; larvae parasitic on metallic wood-boring beetles (Buprestidae).

Superfamily Cephoidea (stem sawflies)

Contains one family, Cephidae.

Family Cephidae (stem sawflies). Slender insects; larvae live in stems of grasses (including commercial grains) and berry plants; sometimes highly destructive; wheat-stem sawfly (*Cephus cinctus*) important wheat pest in Western U.S.

Suborder Apocrita

Abdomen and thorax separated by narrow "waist"; ovipositor adapted for piercing or stinging; many species with complex social organization; many forms carnivorous; larvae usually without legs, often parasitic; some species parthenogenetic.

Superfamily Ichneumonoidea (parasitic hymenopterans)
Wasplike in appearance but seldom sting.
Family Stephanidae (stephanids). Rare insects parasitic on wood-boring beetles; about 100 species.
Family Braconidae (braconids). A large group; some forms highly beneficial to man by being parasitic on insect pests; mostly less than 15 mm long; many pupate in silk cocoons outside host's body.
Family Ichneumonidae (ichneumons). One of the largest of all insect families, about 4,000–5,000 species; hosts of these parasites include other insects, spiders, false scorpions; adults vary in size, colour, shape; many resemble slender wasps; largest about 3.7 cm (about 1.5 in.) long; ovipositor may be twice as long as body and used to penetrate host's tunnel in wood.

Superfamily Chalcoidea (chalcids)
A large, important group; chiefly found on flowers and foliage.
Family Mymaridae (fairyflies). Mostly less than 1 mm long; smallest about 0.21 mm; all are parasites of eggs.
Family Trichogrammatidae (trichogrammatids). About 0.3–1 mm long; egg parasites.
Family Eulophidae (eulophids). A large group; adults 1–3 mm long; often with brilliant metallic colours; parasites on many crop pests.
Family Elasmidae (elasmids). A small but widely distributed group; small and black; parasites on moths, butterflies, or on parasites of Lepidoptera.
Family Thysanidae (thysanids). Small, robust insects parasitic on scale insects, whiteflies (Homoptera), or other chalcids.
Family Eutrichosomatidae (eutrichosomatids). A small group parasitic on snout beetles (Curculionidae).
Family Tanaostigmatidae (tanaostigmatids). Snout-beetle parasites.
Family Encyrtidae (encyrtids). A large, widespread group, mostly 1–2 mm long; mostly parasitic on aphids, scale insects, whiteflies.
Family Eupelmidae (eupelmids). A large group of parasites with wide host range, including Lepidoptera, beetles, spiders.
Family Eucharitidae (eucharitids). Rather uncommon, medium-sized insects, usually black or metallic blue or green; mostly parasitic on ant pupae.
Family Perilampidae (perilampids). Stout-bodied insects, often green or black; mostly hyperparasitic on Diptera and Hymenoptera that parasitize caterpillars.
Family Agaonidae (fig insects). The Smyrna fig is pollinated only by fig insects; includes species with a remarkable degree of sexual dimorphism.
Family Torymidae (torymids). Rather slender, metallic green. Length 2–4 mm; ovipositor long; includes parasites as well as plant-eating species.
Family Ormyridae (ormyrids). A small group similar to torymids, but with short ovipositor.
Family Pteromalidae (pteromalids). Largest group in superfamily; mostly black, metallic green, or bronze; parasitic, with wide range of hosts including crop pests.
Family Eurytomidae (eurytomids, seed chalcids). Some parasitic, some phytophagous forms; some produce gall on grain stems; some live in bee or wasp nests; others parasitize Orthoptera eggs.
Family Chalcidectidae (chalcidectids). Mostly 2–7 mm long; parasitic on Lepidoptera, Diptera, Coleoptera, Hymenoptera.
Family Chalcididae (chalcidids). Two to 7 mm long; parasitic on various Lepidoptera, Diptera, Coleoptera.
Family Leucospidae (leucospids). Rather rare; usually black and yellow; parasites of wasps and bees.

Superfamily Cynipoidea (gall wasps and relatives)
Mostly small or minute gall-forming insects; usually black; some parasitic.
Family Ibaliidae (ibaliids). Mostly 7–16 mm long; uncommon; parasites on horntails.
Family Liopteridae (liopterids). Few species; rather large; rare.
Family Figitidae (figitids). Parasitic on pupae of Diptera and lacewings (Neuroptera).
Family Cynipidae (mostly gall wasps). The gall wasps proper comprise a single subfamily, Cynipinae, which includes most of the family; gall wasps are gall makers or inquilines (*i.e.*, living in a gall but not feeding on it).

Superfamily Proctotrupoidea (parasitic hymenopterans)
All species parasitic on immature forms of other insects.
Family Evaniidae (ensign wasps). A widely distributed group; all parasites of cockroach egg capsules; mostly 10–15 mm long; abdomen held high, thus the name ensign.
Family Gasteruptiidae (gasteruptiids). Slender insects with ovipositor about same length as body; black, sometimes with red abdomen; rather scarce; parasites of solitary wasps and bees.
Family Pelecinidae (pelecinids). Only 1 species in North America, *Pelecinus polyturator,* the female of which is about 5 cm long or more; parasitic on June-beetle larvae.
Family Vanhorniidae (vanhorniids). In North America only 1 species, parasitic on larvae of false click beetles.
Family Roproniidae (roproniids). Includes 3 North American species, all rare.
Family Heloridae (helorids). In North America only 1 species, *Helorus paradoxus,* about 4 mm long, black; parasitic on lacewing larvae.
Family Proctotrupidae (proctotrupids). Mostly 3-6 mm long; parasitic on beetle larvae.
Family Ceraphronidae (ceraphronids). A rather large group; parasitic on braconids or chalcids that are parasitic on aphids or scale insects.
Family Diapriidae (diapriids). Small black insects mostly parasitic on Diptera.
Family Scelionidae (scelionids). Small insects parasitic on insect or spider eggs; some used to combat crop pests.
Family Platygasteridae (platygasterids). Minute black insects; largest group in the superfamily; parasitic on gall midges, including the Hessian fly, a serious wheat pest.

Superfamily Bethyloidea (parasitic hymenopterans)
All rare except cuckoo wasps (Chrysididae).
Family Chrysididae (cuckoo wasps). Brilliant metallic colour, green, red, or blue; widely distributed; mostly less than 12 mm long; mostly parasites of wasp or bee larvae.
Family Bethylidae (bethylids). Rare.
Family Sclerogibbidae (sclerogibbids). Rare.
Family Dryinidae (dryinids). Rare; most species exhibit marked sexual dimorphism.
Family Trigonalidae (trigonalids). Rare.

Superfamily Scolioidea (parasitic wasps and ants)
Some families very similar to those of superfamily Vespoidea and are placed there by some authorities.
Family Tiphiidae (tiphiid wasps). Medium-sized, black, hairy wasps; many parasitic on scarab beetles, including the destructive Japanese beetle.
Family Sierolomorphidae (sierolomorphid wasps). Rare.
Family Mutillidae (velvet ants). So called because the females are wingless, antlike, and covered with short, dense hairs; mostly brightly coloured; mostly parasitic on larvae and pupae of wasps and bees.
Family Rhopalosomatidae (rhopalosomatid wasps). Rare; parasitic on crickets.
Family Scoliidae (scoliid wasps). Large, hairy, black wasps often with a yellow band or bands on abdomen; parasites of scarab-beetle larvae.
Family Sapygidae (sapygid wasps). Rare; parasitic on leaf-cutting bees.
Family Formicidae (ants). A large familiar group; worldwide distribution, but most common in tropics and subtropics; more than 5,000 species known; all social in habit; a few parasitic; some extreme polymorphism.

Superfamily Vespoidea (vespoid wasps)
Adults usually feed on nectar or sap; larvae eat spiders, other insects; antennae usually 12- or 13-segmented.
Family Vespidae (paper wasps, potter wasps, and relatives). Solitary as well as social; includes the well-known yellow jackets and hornets; a widespread group including some large species; the queen of *Vespula ducalis* of Himalaya region reaches 4 cm in length and more than 8 cm in wingspread.
Family Pompilidae (spider wasps). Distribution nearly worldwide; mostly dark-coloured, with dusky wings; usually 12–25 mm long; some 75 mm long; larvae of most species feed on spiders.

Superfamily Sphecoidea (sphecoid wasps)
All species solitary.
Family Ampulicidae (ampulicids). Small, black, rather rare.
Family Sphecidae (sphecid wasps). A large group; adults most often found in flowers; most nest in wood, earth burrows, or in cells of mud.

Superfamily Apoidea (bees)
Distribution nearly worldwide; differ from most wasps in that larvae are usually fed pollen and honey rather than animal food; maxillae and labium form a "tongue" through which nectar is sucked.
Family Colletidae (plasterer bees, yellow-faced bees). A primitive group; plasterer bees are earth burrowers; hairy, moderate in size; yellow-faced bees are small, nearly hairless, nest in crevices, plant stems, or in the earth.
Family Adrenidae (mining bees). Nest in the ground; a large, widely distributed group.
Family Halictidae (mining bees). Usually nest in the ground; often important in plant pollination; small to medium in size; often metallic in colour.

Family Melittidae (melittids). Rare; similar in habit to Andrenidae.

Family Megachilidae (leaf-cutting bees). Medium-sized, stout insects; cut pieces from leaves to line cells in nest; some parasitic; includes so-called sweat bees that are attracted by human perspiration.

Family Apidae (mining bees, digger bees, cuckoo bees, carpenter bees, bumblebees, honeybees). Social as well as solitary; important in flower pollination; the honeybee, *Apis mellifera* in particular, is one of the few domesticated insects.

Critical appraisal. Among systematists in the United States the number of families assigned to the Hymenoptera is relatively moderate. European authorities tend to assign few families to the group (less than 20 according to some), while those of South America tend to assign more. As further information is gained about the behaviour, physiology, and biochemistry of the Hymenoptera, it is probable that comparative studies will reveal unsuspected relationships within the order. As the fossil record also reveals new facts, it is inevitable that additional refinements to the present taxonomic scheme will occur. (M.Li.)

Diptera (true flies)

Although many winged insects are commonly called flies, the name is strictly applicable only to members of the order Diptera, the two-winged, or "true," flies. Diptera, one of the largest insect orders, number more than 85,000 species that are relatively small, with soft bodies. Although the mouthparts of flies are of the sucking type, individuals show considerable variation in structure. Many flies are of great economic importance. Some bloodsuckers are serious pests of man and other animals. These insects, along with many scavenging flies, are important vectors of disease; others are pests of cultivated plants. Flies are beneficial, too—as scavengers, predators or parasites of certain insect pests, pollinators of plants, and destroyers of weeds noxious to man. Dipterous larvae, often called maggots or grubs, are found in many habitats (*e.g.*, in any kind of water, in plant tissue and soil, beneath bark or stones, in decaying plant and animal matter, even in pools of crude petroleum). Adults feed on plant or animal juices or other insects. Diptera fall into three large groups: Nematocera (*e.g.*, crane flies, midges, gnats, mosquitoes), Brachycera (*e.g.*, horse flies, robber flies, bee flies), and Cyclorrhapha (*e.g.*, flies that breed in vegetable or animal material, both living and dead).

GENERAL FEATURES

Flies range in size from midges of little more than one millimetre to robber flies more than seven centimetres long. In general the more primitive flies (*e.g.*, mosquitoes, midges, fungus gnats) are fragile insects with delicate wings. The more advanced flies (*e.g.*, blow flies, houseflies) are generally squat, sturdy and bristly; they fly in a more purposeful way than midges and gnats.

Distribution and abundance

Diptera are abundant throughout the world: in the tropics, the subarctic, at sea level, and high on mountains. They colonize beaches to low-tide level, but few go into deeper water, and only one or two midges are truly marine (*e.g.*, *Pontomyia natans* in the Pacific). On the other hand, migrating flies are found far out to sea.

IMPORTANCE

The abundance, worldwide distribution, and habits of flies combine to make them a nuisance to man. Swarms of midges are a common annoyance. Sweat flies and face flies gather around the eyes, nose, and mouth and also suck blood and pus from wounds and sores. Such flies move constantly from one victim to the next; in so doing they can transfer disease-causing organisms.

The housefly (*Musca domestica*) is dangerous because it feeds on almost any food and moves from person to food, drink, or feces. By transferring infective organisms from the skin and intestine, houseflies are agents in transmitting typhoid, dysentery, cholera, summer diarrhea in children, and other intestinal virus- and bacteria-caused diseases. Eye gnats are a nuisance in warm countries. Although the larvae are plant feeders, the small active adults feed on physiological secretions, particularly those around the eyes. Other flies pierce the skin of vertebrates and feed on their blood. Mosquitoes, black flies, sand flies, biting midges, and horse flies have evolved mandibles and maxillae that are bladelike, piercing stylets.These piercing organs are developed only in females, which use blood protein in egg production. Males do not feed on blood. Other groups of flies have evolved different mechanisms for obtaining blood. Tsetse flies, stable flies or biting houseflies (*Stomoxys*), and certain parasitic flies have developed a hard drill-like labium to replace the soft spongelike one. Both males and females have evolved this labium; therefore both feed on blood. A few flies related to the housefly have a spongy proboscis equipped with small teeth for rasping skin around wounds and sores to increase the flow of blood and lymph. Other groups (*e.g.*, robber flies) have developed a piercing proboscis used only against other insects.

Disease-causing organisms in the blood can be carried by a fly inserting its proboscis into successive victims. After piercing its victim's skin, a bloodsucking fly injects saliva into the wound; without the anticoagulant properties of this saliva, bloodsucking would be impossible since the tiny hole drilled by the proboscis would clog with clotted blood. If the mouthparts are contaminated with blood that contains microorganisms, all are injected, along with the saliva, into the new victim. This is called direct, or mechanical, transmission of disease and occurs only if the fly, interrupted during a meal, finds a new victim before the microorganisms die. One contagious disease that might be spread this way is tularemia, caused by a bacterium found in wild rodents. Trappers who cut themselves while skinning animals can contract the disease. In North America the bacterium is thought to be transmitted also by the deer fly (*Chrysops discalis*), common in wooded trapper country.

Surra, a disease of horses and camels in the Middle East and the Orient, is caused by *Trypanosoma evansi* and is transmitted by horse flies. Trypanosomes, transmitted by tsetse flies, cause sleeping sickness in man and nagana in

From *Invertebrate Identification Manual* by Richard A. Pimentel. © 1967 by Litton Educational Publishing, Inc. Reprinted by permission of Van Nostrand Reinhold Company

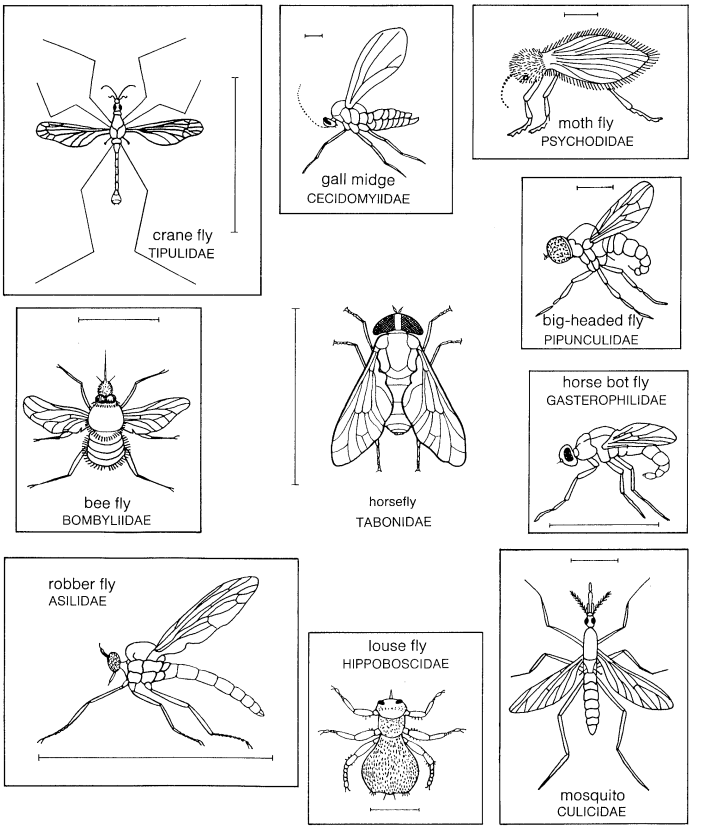


Figure 48: Diversity among Diptera. Line scales indicate the approximate size of each insect.

animals throughout tropical Africa. These trypanosomes must spend part of their life cycle in the insect before they can infect a vertebrate; this is an example of cyclic disease transmission. The relationship between the parasitic disease organism and its two hosts, vertebrate and insect, is a result of evolutionary adaptation; however, it is not known whether the trypanosome was originally a fly parasite that spread to man and other vertebrates, or whether it was a human parasite that became adapted to living in a biting fly. An important cyclically transmitted disease is malaria. *Plasmodium,* the causative agent of human malaria, is an acellular protist nourished by red blood cells in the blood of man. Its reproductive cycles cause recurrent bouts of the disease. Occasionally sexual forms occur in the victim's blood; if this form finds its way into a suitable species of bloodsucking mosquito, another stage of the *Plasmodium* begins, preparing the organism to infect another human bitten by the mosquito host. Other diseases known to be cyclically transmitted include yellow fever, filariasis, and many viral diseases. It is likely that there are others not yet recognized.

njuries to lants

Fly larvae are serious agricultural pests; they feed on young crop plants, retarding growth or killing them. Cultivated crops, because they provide pests with an almost unlimited food supply within a small area, can be devastated by uncontrolled population growth of a pest. On the other hand, wild food plants, because they are scattered and mixed with other varieties, do not usually provide so abundant a food supply and thus serve as a check on population growth. Frit flies can cause a 20 percent loss of an oat crop, and to the value of the lost oats must be added the cost of control measures necessary to save the remainder. Some crops, notably fruit trees and ornamental shrubs, are a financial loss if slightly disfigured by insect attack, though the life of the plant is not endangered. Fruit, although edible after attack by Mediterranean fruit flies, cannot be sold; a few infested fruits can result in loss of an entire consignment. Larvae of gall midges and leafminers lower the commercial value of ornamental plants.

NATURAL HISTORY

redomi- ant stage n life cycle

Life cycle. *General features.* The life cycle of a fly consists of four stages: egg, larva, pupa, and adult. Since larval forms, always morphologically distinct from adults, also occupy different habitats, flies in effect live two distinct

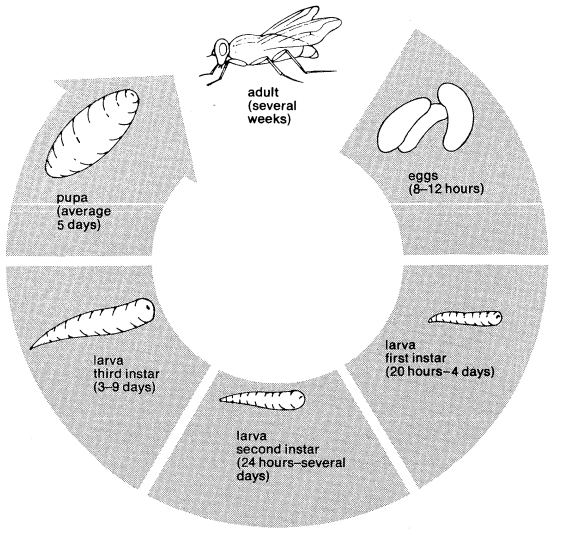


Figure 49: Life cycle of the housefly.

lives and thus are able to adapt successfully to environmental changes. In some flies (*e.g.,* robber flies) neither the larval nor the adult stage predominates; the larva feeds actively in soil, and adult flies of both sexes catch other insects in flight. Among mosquitoes, black flies, and related bloodsucking flies, the larvae have characteristic structures and live active lives under water; the complex mating process of the adults is followed (in the case of females) by bloodsucking and egg laying.

There are many flies in which one stage is predominant. Swarms of adult midges (Chironomidae), for example, are conspicuous and troublesome; but the adult midge lives just long enough—usually less than a day—to mate and lay eggs. Thus, most of the life cycle is occupied under water by the larval stage. The larvae are wormlike in appearance. Some are adapted to oxygen-poor situations; the "bloodworm," for example, which lives in the mud of stagnant waters, uses hemoglobin as a respiratory pigment. Other midge larvae live in silken tubes, either filtering minute organisms from water for food or preying upon larger creatures. Some midge larvae have evolved an elaborate symbiosis, or mutualism, with other aquatic organisms; for example *Nostoc* (a genus of blue-green algae) and certain midge larvae utilize each other's excreta. Larval life as complex as this is not mere preparation for adult life; rather, the adult stage is a revitalizing and distributory stage for the larval one. The adult stage is of relatively little importance in a few other groups, too.

At the opposite extreme are tsetse flies (*Glossina*) and three families of pupipara parasites (*e.g.,* Hippoboscidae, which feed on the blood of mammals and birds; both Nycteribiidae and Streblidae feed only on bat blood). In these families a single egg is produced at one time and hatched internally. The larva, retained and nourished in a kind of womb, is expelled when it has matured and immediately forms a pupa. Thus, these flies have no free larval stage, no independent larval life. Since the pupa is immobile, the active life of the fly is passed as an adult. Most Hippoboscidae and Streblidae, and all tsetse flies, have wings and usually migrate to new hosts, but some species of these families, and all Nycteribiidae, cannot fly and often are wingless. Wingless flies can be identified as flies only after detailed morphological examination.

Eggs. The majority of flies lay eggs, which hatch into tiny larvae after a few hours or several days. The number of eggs laid by a female varies from 1 to about 250; however, a number of successive batches may be laid. The greenbottle fly (*Lucilia sericata*) has laid nearly 2,000 eggs in captivity; however, the total is probably fewer than 1,000 in the natural state when time and energy are lost looking for suitable places to lay. Egg-laying sites, chosen instinctively by the females, are related closely to larval habitats. Since many fly larvae feed in soft organic materials, many females have developed telescopic ovipositors, formed from the last three or four abdominal segments. The female uses the ovipositor to press the eggs into a mass of decaying material. Blow flies and houseflies push their eggs between the membranes of meat or into any convenient cavity in decaying organic material. The small fruit flies (*Drosophila*), which lay in rotting fruits and fermenting materials, also have this type of ovipositor; however, the large fruit flies (*e.g.,* Mediterranean fruit fly), which lay eggs in the rind of growing fruits, have a stiffer ovipositor. Elaborate ovipositors found in the robber flies are used to push eggs into the interstices of flower heads and the axils of grasses, sometimes even into plant tissues, to conceal them and protect them from drying. When hatched, the larvae drop to the ground and burrow under the soil.

Egg-laying sites

Larvae. Fly larvae have one common characteristic: all lack true, jointed, thoracic legs. Many fly larvae have "false legs" (prolegs or pseudopods) similar to those that support the fleshy abdomen of a caterpillar. Flies, much more versatile in this respect than caterpillars, can have prolegs around any body segment. Prolegs help the larvae crawl through narrow spaces or push through soil.

The evolutionary trend among fly larvae has been toward structural simplification; thus, generally, larvae of primitive flies are more structured than are larvae of more highly evolved flies, which show greater physiological versatility. Larvae of most members of the suborder Nematocera (see below *Annotated classification*) have a well-developed head, with antennae, palpi, and complex mouthparts similar to those of many adult insects. Often they are so structurally adapted to their special way of life that they are unable to adapt to any other. This is especially true among aquatic larvae (*e.g.,* mosquitoes) and perhaps reaches the extreme in mountain midge larvae,

which live in rushing torrents and crawl on submerged rocks. Their body segments are equipped with clinging processes and suckers.

Maggots

In contrast to highly specialized larvae, about half the fly species have larvae known as maggots. The maggot has lost the complicated head capsule of primitive flies; its pointed anterior end contains one or a pair of mouth hooks. The blunt posterior end has a pair of posterior spiracles (external airholes) that appear to the naked eye as black spots. Microscopically the spiracles are seen as a complex pattern of slits or pores that are useful in distinguishing species.

Although maggots show structural uniformity, they are diverse physiologically. Most maggots feed on decaying organic matter, but there are wide differences in the food preferences of different flies. Eight "waves" of maggots have been distinguished; each wave attacks dead animals in a strict sequence as decay progresses from the newly dead corpse through rigor and putrefaction to mummification. Although some maggots appear only during a clearly defined stage of animal decomposition, the large voracious maggots of many blow flies feed on any animal matter, including living tissues. The best-known blow flies are sheep blow flies, principally species of *Lucilia.* Maggots of *Lucilia sericata,* for example, feed in small dead animals, abattoirs, and garbage cans; they oviposit in soiled wool around the anus of sheep or in the pus exuding from scratches and wounds, where they are important agents of sheep strike disease. These maggots sometimes occur in soil near buildings in cities; their food source is not known.

Other maggots feed in living or decaying vegetable matter. Larvae of the frit fly of oats and the gout fly of barley are maggots of flies that belong to the plant-feeding family Chloropidae. The hessian fly of wheat is the destructive larva of *Mayetiola* (*Phytophaga*) *destructor* of the nematoceran family Cecidomyiidae (the gall midges). Although the external structure of most nematoceran larvae is complex, the structure of the gall midges, which live completely immersed in plant tissue, has evolved in the direction of simplification; gall midges are among the most difficult fly larvae to identify. Also known as gall gnats because feeding larvae cause formation of disfiguring galls on leaves or stems, gall midges harm many kinds of plants. Thus they have evolved simplified structure and physiological diversity regarding food plants as have maggots of more advanced flies.

Molts and larval stages. Among insects in general, the evolutionary tendency has been toward decreasing the number of molts during development, and flies are no exception. The number of larval stages, or instars, is six or seven in black flies (Simuliidae) and four in most other Nematocera. Along the second line of evolution of flies, Brachycera have from five to eight instars while the maggots of the most advanced flies (Cyclorrhapha) have only three. One or two species have no molts. Sometimes molts occur before the larva hatches from the egg. Muscidae, for example, are arranged in three groups according to whether they are trimorphic (*i.e.,* have three free larval instars), dimorphic (*i.e.,* pass the first instar in the egg, have two free larval instars), or monomorphic (*i.e.,* pass the first two instars in the egg, have one free larval instar). Monomorphic larvae are always predatory; trimorphic and dimorphic larvae feed first on decaying matter (are saprophagous), but they may or may not be predatory in their final instar.

Pupa. The external features of the adult fly (*i.e.,* eyes, antennae, wings, legs) are clearly visible in the pupa. The pupa, however, is not always exposed to view; it may be enclosed either in a cocoon of extraneous matter (*e.g.,* soil, or silk, or a mixture of the two) or in a puparium, which is a case formed by the hardening of the larval skin. A puparium is formed in flies of the family Stratiomyidae and others that have maggots as larvae (all Cyclorrhapha).

Cocoons

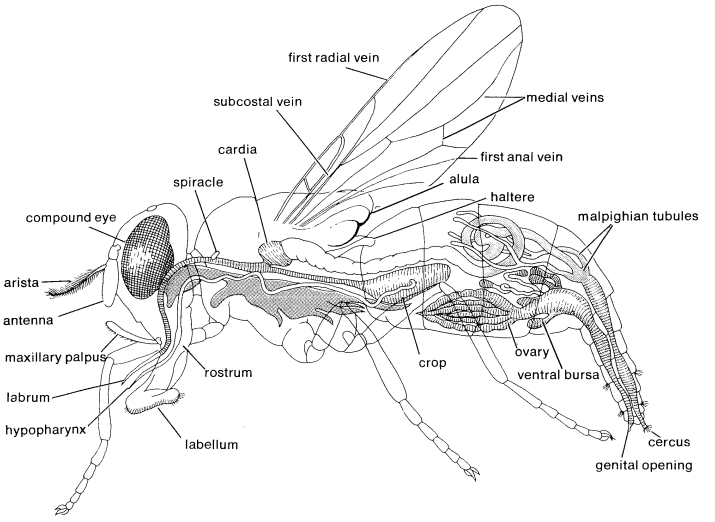
Many families of flies form cocoons sporadically; the cocoon has evolved as an adaptive device that provides extra protection to the pupa. The pupae of mosquitoes, of black flies (Simuliidae), and of a few aquatic midges swim actively. Many pupae that lie in soil or in wood have developed thorns and spines in order to help them work their way to the surface just before emergence of the adult insects.

Adult. The adult fly emerges from the pupa soft and crumpled with a colourless skin (integument) and perfectly formed (though not fully pigmented) hairs and bristles. The newly emergent adult swallows air to expand its body and wings and to force blood through its body. In the more advanced flies of the group Schizophora (see below), the ptilinum, an inflatable membranous sac in the head, is used to aid this process. The ptilinum shrivels away after it has performed its function; however, it leaves behind the ptilinal suture, a horseshoe-shaped groove that runs over and beside the antennal sockets and is only found in Schizophora.

Ecology. It has been said that there is hardly any life-supporting medium in which dipterous larvae have not been observed. It is not possible to discuss all dipteran habitats, but the annotated classification below provides many examples. Maggots, however, are the most important larvae, because they play an essential role in breaking down and redistributing organic matter. The waste products excreted by the larvae provide nutrients for molds, fungi, and plants. In addition, the bodies of larvae, pupae, and many adult flies are an important food source for higher animals. Examples are aquatic larvae of midges and mosquitoes, which are staple food for fish. The terrestrial maggots of many flies also have a role in food chains. Since a blow fly can lay one to two thousand eggs, the blow fly population would increase calamitously if more than a few of them survived. Most of the larvae die of malnutrition, desiccation, or drowning, or are consumed by birds. The adult flies are snapped up by birds, small mammals, frogs, and toads. Swallows, swifts, and martins devour large numbers of flies that have been carried up into the air by convection currents. Thus, the population is maintained at a constant level.

The most fundamental importance of flies, therefore, lies not in the few familiar families that contain mosquitoes, tsetse flies, houseflies, and other nuisance insects, but

From H. Weber, *Grundriss der Insektenkunde*



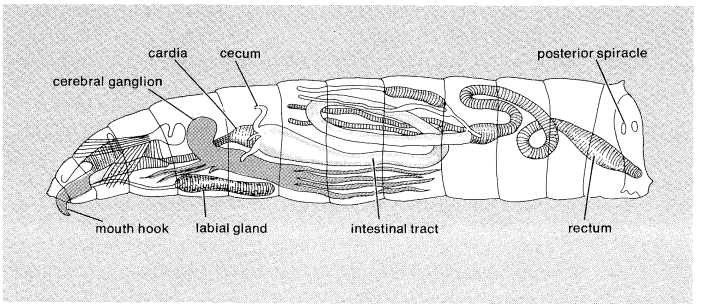


Figure 50: *Body plan of Diptera.*
(Top) Adult; (bottom) larva.

rather in the large numbers of unfamiliar species that are an essential element in the food chains upon which all life depends.

FORM AND FUNCTION

External features of adult. *General appearance.* The thorax, abdomen, and legs of adult flies vary from long to short; the appearance of the fly is functional as well as decorative. Sometimes the bright colour and pattern of many flies is metallic (*e.g.,* blow flies), but most often the fly is covered with a fine coating called tomentum or dusting. Many flies, particularly those of more highly evolved families, are bristly; and the strongest bristles have a precise location, particularly on the thorax. The arrangement of bristles and the identification method based on them is called chaetotaxy.

Wings. Adult flies have only one pair of wings, on the mesothorax or second thoracic segment. The hind wings, modified into halteres, have a stalk and a knob, or club, that may be large and heavy relative to the size of the fly. The halteres vibrate up and down in time with the wings and act as gyroscopes in flight. If the fly yaws, rolls, or pitches during flight, the halteres, maintaining their original plane of movement, twist at their bases, where special nerve cells detect the twist and cause the fly to correct its flight attitude.

Wing venation

The wings of flies have a defined pattern of veins; each has a name and characteristic location, often of taxonomic value. Few true flies have a reticulation (*i.e.,* network of small veins) such as those in many other insects that are mistakenly called flies (*e.g.,* mayflies, dragonflies, dobsonflies). Primitive flies tend to have complex wing venation, while advanced ones have reduced and simplified venation. Some of the small midges (*e.g.,* Cecidomyiidae, Sciaridae, Mycetophilidae) have reduced wing venation also. Reduction or loss of wings occurs in many families, particularly those that inhabit windy places (*e.g.,* mountains, islands) or caves, or that are external parasites among fur and feathers.

Eyes. The eyes of flies often occupy most of the surface of the head, especially in males, where the eyes may meet in the middle line (holoptic). In female flies, with few exceptions, the eyes do not meet (dichoptic). In some families, notably robber flies and small acalyptrate flies, both sexes are dichoptic. Parasitic flies, or those that live in secluded places, may have very small eyes or none at all. Typically, however, the compound eyes of flies contain many facets; for example, the housefly has 4,000 facets in each eye, about average for insects.

Mouthparts. The mouthparts of flies are adapted for sucking. Most flies have maxillae; many also have mandibles, elongate blades that overlie a groove in the labium and form a tubular channel for sucking liquids. In some females (*e.g.,* bloodsucking flies, mosquitoes) the mandibles act as piercing stylets for drawing blood. Mandibles became functionless or were lost entirely relatively early in fly evolution and therefore bloodsucking families that evolved later had to develop other piercing methods. Tsetse flies and stable flies use the hardened labium; robber flies and dance flies use the hypopharynx; and Dolichopodidae (small, metallic green flies with very long legs) envelop prey in the spongy labella of the labium and crush it with specially evolved teeth. Most flies suck their food; the few exceptions have reduced mouthparts and possibly do not feed at all as adults. Thus the food of flies must be liquid or solids that can be liquefied by saliva and stomach juices. Flies also have a pair of labial palpi equipped with sensory cells that act as organs of touch, taste, and smell. The palpi and the antennae are essential for examining possible food sources and suitable sites for egg laying.

Evolutionary specializations

Antennae. All flies have antennae. Members of the suborder Nematocera (*e.g.,* crane flies, various midges, and gnats) have whiplike antennae with two basal segments (scape and pedicel) and a flagellum of many similar segments. All other flies, properly called Brachycera, or short horns because the flagellum is contracted into a compound third segment, have remnants of the terminal flagellar segments remaining as a pencillike style or a bristle-like arista. Considerable antennal structural differences exist among related genera and species.

Larval features. Larvae of flies have no wings, show no external traces of wingbuds (endopterygote insects), and do not have segmented thoracic legs. Larvae of primitive flies (most Nematocera and Brachycera) have a well-developed head, with chewing mouthparts. Evolution has favoured reduction of the head capsule and replacement of chewing mouthparts with a pair of mouth hooks that move in a vertical plane. Larvae with adaptive external structures (*e.g.,* prolegs) generally belong to the Nematocera or Brachycera. The maggots of the Cyclorrhapha have little external structure other than black mouth hooks and the posterior spiracles. Although a few of these larvae show secondary complexities (*e.g.,* some aquatic larvae of hover flies and shore flies), most cannot be identified beyond the family level.

Nutritional requirements. *Adults.* Nutrition involves balance between feeding habits of larval and adult flies. Primary feeding occurs during the larval stage; adult feeding serves to compensate the shortcomings of larval nourishment. At one extreme are nonbiting midges, with larvae that vigorously filter microorganisms from water; the adults do not feed. Related to nonbiting midges are biting midges, mosquitoes, and black flies; adult females in these families must supplement an insufficient larval diet. Although one batch of eggs occasionally is laid without a meal of blood, blood is necessary to mature a second batch. Flies that lay one batch of eggs without blood are autogenous; those that cannot lay at all without blood are anautogenous. One species can have both types, possibly as a result of shifting populations or races arising from natural selection. For example, in the far north large populations of biting flies (*e.g.,* mosquitoes, biting midges, black flies, horse flies) occur during the short Arctic summer; obviously there are insufficient numbers of warm-blooded animals to provide food. If the flies find blood, they use it; if not, they still survive.

Blood, pollen, and nectar

Most adult flies visit flowers, which provide water, nectar, and pollen. Pollen, more difficult for a sucking insect to obtain than blood, is rich in protein and is an important source of this nutrient. Certain hover flies crush pollen grains between hardened portions of the labella before swallowing them; many flies actively probe into flowers, covering their heads and eyes with pollen grains. Nectar from flowers contains carbohydrates, and most adult flies use this syrupy liquid. Although their role in pollination is less well known than that of bees, flies are important pollinators of flowers. Some plants (*e.g.,* spurges) are often covered with small flies of different families. Small flies also feed on honeydew from aphids (see section on Homoptera). Although the name *Drosophila* means "lover of dew," this insect sucks water and any other obtainable fluid. Flies feed on dung and liquid products of either animal or vegetable decay. They obtain nutrients from farmyard manure heaps and garbage dumps. These places also harbour many larvae that feed either directly on available organic food or are carnivorous on other larvae. A familiar example is the yellow dung fly; adults prey on other insects visiting the dung.

Larvae. The adaptability of flies is evident in the wide range of foods that larvae eat. Apart from parasites, the most specialized feeders are larvae that live in plant tissues (*e.g.,* leafmining Agromyzidae, many restricted to one plant species or group). Generally agricultural and horticultural pests (*e.g.,* cabbage root fly) are versatile species, feeding on a variety of wild hosts and modifying their diets when presented with concentrated plantings of commercial crops. Many carnivorous fly larvae (*e.g.,* asilids) probably live in soil and eat vegetable or animal matter, whichever is available. Since adult asilids (robber flies), however, feed on other insects, the larval nourishment is presumed to be inadequate. Some larvae, particularly maggots, that feed on vegetable matter during the first and second instars, become carnivorous during the third instar, when most of the growth takes place.

Respiration

Larval respiration is adapted to the medium in which the larvae live. Although a few parasitic larvae (*e.g.,* Pipunculidae, parasitic in froghoppers, and Drosophilidae, internal parasites of scale insects) get oxygen through the

skin, most dipterous larvae need a tracheal system to distribute oxygen. Primitively, the tracheal system probably opened exteriorly by paired spiracles on each segment of the body. The soil dwellers, Bibionidae and Scatopsidae, retain this system, although most families have kept spiracles only on the thorax (one pair) and one at the tip of the abdomen. Even these are closed in some aquatic larvae (*e.g.,* luminous larvae of some fungus gnats and larvae of biting midges). However, mosquito larvae and those of most other water-living flies surface frequently to renew their oxygen supplies. Some larvae pierce the stems of underwater plants to obtain oxygen formed as a result of photosynthesis. Maggots of Cyclorrhapha rely heavily on complex posterior spiracles. Pupae respire through prothoracic spiracles that are sometimes equipped with long tubes extending outside the cocoon or puparium.

EVOLUTION AND PALEONTOLOGY

Diptera belong to the panorpoid complex, which includes Mecoptera (scorpionflies), Trichoptera (caddisflies), Lepidoptera (butterflies and moths), Siphonaptera (fleas), and Diptera (true flies). All are believed to have evolved from an ancestor that lived in moss; four-winged insects that resemble crane flies have been preserved as fossils in Permian deposits (see FOSSIL RECORD). Strata of the Lias Period (Lower Jurassic) contain many true midges; early Brachycera began to appear in Mesozoic; Cyclorrhapha appear in the Cretaceous. By the end of the Eocene Epoch most modern families of flies had evolved. Flies in Oligocene amber and copal are similar to living genera.

CLASSIFICATION

Distinguishing taxonomic features. The wings are the most distinctive feature of Diptera; they consist of a pair of functional forewings and reduced hind wings called halteres that serve as balancing organs. Except for male scale insects (see section on Homoptera), only Diptera have hind wings modified into halteres. The thorax consists almost entirely of mesothorax filled with muscles that operate the forewings. This feature is useful in identifying wingless flies. The single pair of wings also distinguishes Diptera from other insects called flies (*e.g.,* caddisflies, dragonflies), while the posterior halteres separate the Diptera from other insects that have one pair of wings (*e.g.,* some mayflies and beetles).

Division into suborders is based on structure of antennae and wing venation. Another major feature is chaetotaxy, the arrangement of strong bristles, many in fixed positions and given individual or group names. Separation of Diptera into families is based on habitats and habits (*e.g.,* feeding) of larvae and adults. Genera and species are distinguished by details of head structure, shape and degree of separation of eyes, profile of head, and shape and proportions of leg segments. Abdominal shape often determines characteristic appearance of a genus, but it is difficult to define; the shape varies as the insect is starved, well fed, or pregnant (viviparous flies, such as tsetse).

Annotated classification.

ORDER DIPTERA

Size range one millimetre to 7.5 cm; wings, when present, number two; hind wings reduced to halteres; sucking mouthparts; 85,000 species; worldwide distribution; diverse habitats and diets in both larvae and adults.

Suborder Nematocera

Antennae consist of scape, pedicel, and flagellum with numerous similar segments; maxillary palpi with more than three segments, often pendulous; anal cell of wing open; larvae usually with well-defined head, mandibles horizontally opposed.

Family Tipulidae (crane flies). Elongated body, wings, legs; slow-flying; larvae in soil (leatherjackets), moss, rotten wood, mud, fresh water, littoral, marine.

Family Mycetophilidae (fungus gnats). Fragile, slender; flit about in damp, shady places, among decaying vegetation.

Family Sciaridae. Similar to fungus gnats but more compact, more often indoors.

Family Bibionidae (march flies in Northern Hemisphere). Compact, well-armoured flies; strong spurs on legs; often abundant on spring blossoms; larvae in soil, sometimes found in a tangled mass near roots of plants.

Family Scatopsidae. Similar to march flies, more often indoors.

Family Cecidomyiidae (gall midges). Tiny flies seldom seen as adults; shapeless larvae burrow into plant tissues, cause formation of plant galls, and deform leaves, stems, and roots; some horticultural and agricultural pests.

Family Psychodidae (moth flies). Tiny, with hairy wings; often seen singly in kitchens, on windows above sinks; some larvae numerous in sewage sedimentation tanks; larvae mostly aquatic.

Family Phlebotomidae (sand flies). Closely related to Psychodidae; adult females suck blood, carry dermal and intestinal leishmaniasis and sand fly fever.

Family Ceratopogonidae (biting midges). Tiny, often with spotted wings (*e.g., Culicoides*); adult females with irritating bite suck blood, carry some parasitic worms; *Forcipomyia* suck blood of insects.

Family Chironomidae (nonbiting midges). Related to biting midges, but females do not suck blood; larvae aquatic; important fish food; adults swarm near water.

Family Simuliidae (black flies). Also buffalo gnats; small, humpbacked, with short antennae; females suck blood, carry parasitic worms that cause "river blindness"; forms nodules under skin; larvae aquatic, filter feeders, attached to stones, underwater vegetation, or fresh-water crustaceans.

Family Culicidae (mosquitoes). Small; elongated; proboscis prominent; palpi often long; best recognized by scaly wings; many females suck blood, carry human diseases (Anophelini carry malaria; Culicini carry yellow fever, filariasis, dengue, viral encephalitis); larvae and pupae aquatic.

Suborder Brachycera-Orthorrhapha

Name usually shortened to Brachycera; flagellum of antennae nearly always fused into a compound third segment, remaining diminutive segments form a stumpy "style" or bristle-like arista; anal cell of wing narrowed, nearly always closed on or before wing margin; palpi seldom with more than three segments, often two or one, held forward (porrect); larvae usually with well-defined head, mandibles move vertically or parallel, cannot be opposed; adult escapes from pupa by a rectangular slit ("Orthorrhapha").

Family Stratiomyidae (soldier flies). Colourful flies, found resting on vegetation with wings closed; males sometimes dance in air; larvae sometimes elongate, aquatic, active, carnivorous (*Stratiomys*); others in decaying vegetation (*Hermetia*).

Family Rhagionidae (snipe flies). Inconspicuous, usually rest on vegetation; some females (*e.g., Symphoromyia*) suck blood; most larvae in soil or in water (some *Atherix* females form egg-laying swarms); some make pits in dust, like ant lions (*Vermileo*).

Family Pantophthalmidae. Large, archaic flies, now found only in tropical forests of South America; wood boring larval grubs sometimes damage commercial timber.

Family Tabanidae (horse flies, deer flies; march flies in Australia). Squat flies with big heads, brilliantly coloured eyes; some females (*Chrysops, Tabanus, Haematopota*) suck blood, are livestock pests; many primitive genera feed only from flowers; larvae in mud or wet soil, either vegetarian (*Chrysops*) or carnivorous (*Tabanus, Haematopota*).

Family Asilidae (robber flies). Adults catch other insects in flight, suck their blood; size varies from a few millimetres to eight centimetres (longest of all flies); characteristic "moustache" of bristles probably protects eyes from damage by fly's victim; larvae in soil or wood; eat any food.

Family Bombyliidae (bee flies). Hairy, scaly; superficially resemble bees, hover over flowers in similar way; often brightly patterned, pattern destroyed by rubbing scales; larvae scavenge in bee and wasp nests, or are parasitic (*e.g.,* locust egg pods, tsetse pupae).

Family Scenopinidae (window flies). Tiny black flies, on windows indoors; develop from larvae in carpets, feed on flea and clothes moth larvae; natural habitat, bird nests or similar dry debris.

Family Therevidae (stiletto flies). Adults resemble Asilidae, but not predatory; larvae like Scenopinidae, elongated, wormlike, carnivorous but sometimes attack plant roots.

Family Nemestrinidae. Rather like Bombyliidae; larvae parasitic in grasshoppers, locusts, perhaps beetles; remarkable for beautiful hovering.

Family Acroceridae (balloon flies). Grotesque; abdomen swollen, thorax small, head tiny; larvae parasitic in spiders.

Family Empididae (dance flies). Adults suck insect blood, also feed from flowers. *Hilara* darts over water, catches small

insects; larvae in many habitats (*e.g.,* marine and freshwater mud, decaying vegetation, fungi, running sap from trees).

Family Dolichopodidae (long-legged flies). Tiny, metallic, bristly flies; large numbers sit on leaves in wet places; predatory on other insects; larvae like Empididae, elongated, with little external head structure, same habitats.

Suborder Brachycera-Cyclorrhapha

Usually shortened to Cyclorrhapha; characteristically form pupa inside last larval skin as a puparium; adult fly pushes off a circular cap, hence the name Cyclorrhapha; most families (Schizophora) with a ptilinum (membranous sac inside head), which emerges from a horseshoe-shaped ptilinal suture (identifies adult Schizophora) above antennae, is puffed in and out to help fly escape from puparium or soil or to inflate fly's body; ptilinum atrophies and only ptilinal suture remains; a small group of Aschiza, without ptilinal suture, are recognized chiefly by their wing venation.

Series Aschiza

Family Lonchopteridae. Little known; notable for parthenogenesis; few species; worldwide; sometimes abundant.

Family Phoridae (coffin flies). Tiny flies sometimes numerous indoors; larvae live in any organic debris rich in protein or nitrogenous decay products and scavenge in nests of wasps, bees, ants, termites; breed in carrion; many adults wingless or with short wings (brachypterous).

Family Pipunculidae. Tiny flies; head spherical, noted for precise hovering; larvae parasitic in Homoptera.

Family Platypezidae. Small flies; peculiar legs; rarely seen; appear to dance in smoke of wood fires; larvae live in fungi.

Family Syrphidae (hover flies). Vena spuria in wing runs between third and fourth veins; familiar everywhere; hover over flowers, settle on leaves; some larvae aquatic ("rat-tailed" maggots); larvae of many species feed on aphids on plant stems and leaves.

Family Conopidae (thick-headed flies). Wasplike flies; larvae parasitic in bees and wasps; may be a separate evolutionary line.

Series Schizophora

All flies with a ptilinal suture in head; larvae with no external head structure, mouth hooks visible through cuticle, one pair of prothoracic spiracles and one pair of posterior spiracles, each with either three slits or a mass of small pores; larvae with fore end pointed and hind end truncate are called maggots; larvae with both ends blunt and fleshy, with bulges and tracts of spines, are called grubs.

Section Acalyptrata

Thoracic squamae (*i.e.,* calypters that join base of wing to thorax) are small or evanescent; small soft-bodied flies; major families well established; placement of genera uncertain; families can be grouped according to food preferences of larvae.

Flies breeding in vegetable compost and dung.

Family Lauxaniidae. Larvae in decaying vegetable matter.

Family Helomyzidae. Like Lauxaniidae; most generalized of Acalyptrata.

Family Dryomyzidae. Like Lauxaniidae, but with wider range of food, including fungi; yellow flies often seen in winter.

Family Chyromyiidae. Yellow flies, 1 or 2 millimetres long; breed in debris of bird nests, mammal burrows, caves, cellars; seen singly on windows indoors.

Family Celyphidae (beetle flies). Scutellum enormously enlarged until it covers both abdomen and wings when at rest; tropical dung breeding.

Family Mormotomyiidae. Contains one wingless, African species; looks like a spider; known from only one locality in Kenya; breeds in bat dung.

Family Coelopidae (kelp flies, seaweed flies). Breed in wrack (*i.e.,* heaps of decaying seaweed stranded on beaches) chiefly in temperate countries; adults of some species attracted by trichloroethylene; sometimes pests.

Flies breeding in animal refuse, dung, carrion.

Family Sepsidae. Small, black, roundhead flies, sometimes with spots at wing tips; may breed to infestation level in sewage sludge.

Family Piophilidae (cheese skippers). Larvae in cheese, ham, cured meats, dried fruits, preserved skins and pelts; natural habitat in mummifying carrion; called "skippers" because larvae move both by crawling and "skipping" (*i.e.,* gripping the tip of the abdomen with mouth hooks and flipping the body through a relatively long distance).

Family Micropezidae. Large, long-legged flies; often with conspicuously patterned, blue-black wings; spectacular in tropics.

Family Sphaeroceridae. Tiny, black-brown flies; first tarsal segments of hind legs swollen; abundant throughout world in dunglike materials; some members live in seaweed on beaches; many short-winged or wingless species.

Family Sciomyzidae. Aquatic larvae eat both living and dead snails; may be valuable as controlling agents for injurious snails.

Family Milichiidae. Breed in dung; adults attach to predatory insects and spiders and feed on them; called "insect jackals"; *Madiza glabra* sometimes numerous indoors.

Family Carnidae. Scavenge in nests and burrows. Adults of *Carnus hemapterus* scavenge among bird feathers, break off wings.

Family Neottiophilidae. Nest-breeding; larvae suck blood of nestling birds.

Family Thyreophoridae. Among the rarest of flies; larvae in dead bodies of large animals.

Family Chamaemyiidae. Predatory larvae; known as controlling agents of aphids.

Family Braulidae (bee louse). *Braula caeca,* wingless fly, lives in beehives; larva feeds on wax and pollen stores; adult attaches to bee, may solicit nutritious saliva like other members of bee colony.

Flies with plant-feeding larvae.

Family Ephydridae (shore flies). Transitional; wide range of larval habitats; no substance unpalatable for larvae (*e.g.,* algae, sewage, excrement, carrion, urine, brine, hot springs, tar pools); carnivorous petroleum fly (*Psilopa petrolei*) lives in pools of crude petroleum seepage preying on trapped insects; many larvae feed in terrestrial and aquatic plants.

Family Diopsidae (stalkeyed flies). Some larvae live in decaying plant tissue, others mine in living plants.

Family Chloropidae (frit flies). Most important plant feeders; includes economic pests of cereal and other crops.

Families Opomyzidae, Geomyzidae, Psilidae. Small, usually yellow or grayish flies, plant feeders; *Psila rosae,* the carrot fly, an agricultural pest.

Family Agromyzidae (leafminers). Larvae feed in parenchymatous tissue of leaves, render epidermis transparent and produce either serpentine or "blotch" mines; rarely cause severe damage, but disfigure ornamental trees and shrubs.

Flies with fruit-feeding larvae.

Family Trypetidae (large fruit flies). Form galls in certain flowers particularly Compositae; many Trypetidae larvae feed in living fruits, and ruin them; now worldwide distribution; economic damage by several members (*e.g.,* the Mediterranean fruit fly *Ceratitis capitata*) has resulted in worldwide quarantine laws to regulate entry of fruit into countries.

Family Drosophilidae (small fruit flies). Larvae in decaying and fermenting fruit or any sweet substance; includes *Drosophila melanogaster,* used in genetic studies.

A number of smaller families have been formed to accommodate genera closely related to the two above. Otitidae (Ortalidae) and Lonchaeidae are the most clearly defined. Others such as Ulidiidae, Pallopteridae, Phytalmidae, Camillidae, and Diastatidae are debatable.

Section Calyptrata

Characterized by large squamae (calypters that join base of wing to thorax); Scatophagidae are transitional.

Family Scatophagidae (dung flies). Live around dung, other decaying materials; many also predacious as larvae and as adults.

Family Muscidae (housefly and allies). Many species include the housefly; some larvae carnivorous, especially in third instar; breed in decaying vegetable matter or dung; larvae of *Fannia,* the "lesser housefly" like materials soaked in urine; economically important muscid larvae feed on plant stems and roots; subfamily (sometimes a separate family) Anthomyiinae contains dipteran plant pests; stable fly, *Stomoxys,* (biting proboscis in both sexes) may be placed in a separate family, Stomoxyidae; tsetse fly *Glossina,* confined to Africa, peculiar structurally and biologically, sometimes placed in the family Glossinidae, occurred in North America in the Miocene.

Family Calliphoridae (blow flies). Some bristly flies with carrion-feeding maggots; common blow flies, *Calliphora* (bluebottles), feed as larvae in dead meat; *Lucilia* (greenbottles) sometimes attack living flesh; screw-worms (*e.g., Cochliomyia, Callitroga*) are dangerous feeders in living tissue.

Family Cuterebridae. Offshoot of Calliphoridae above; larvae are parasitic in rodents; one larva, *Dermatobia hominis* (hu-

man bot fly) also attacks man; eggs sometimes attached to mosquitoes and other biting flies and carried to their prospective prey.

Family Oestridae (bots and warbles). Larvae live under skin, in nose, and in other head cavities of large mammals; includes the sheep nostril fly (*Oestrus ovis*), warble flies of cattle (*Hypoderma bovis* and other species).

Family Gasterophilidae (horse bots). Larvae live in stomachs of horses, zebras, rhinos and elephants, attached to intestinal lining; relationship with other bot flies problematical; currently classified with other bot flies.

Family Sarcophagidae (flesh flies). Large, gray and black; common around refuse dumps; larval habits diverse, in living or dead animal matter; many viviparous species.

Family Tachinidae (tachinid flies). Ecologically important in balance of nature because larvae are parasites in other insects, spiders, woodlice, and centipedes; employed in biological control of pests.

Section Pupipara

Disputed group, families may merely be convergent in habit; lay living larvae, adults of both sexes feed exclusively on blood.

Family Hippoboscidae (louse flies). Feed as adults on blood of mammals and birds; many fly, some have wings reduced or lost (*e.g.,* sheep ked, *Melophagus ovinus*).

Family Streblidae (bat flies). Distinct, rounded head, wings often functional but fly little; cling closely to host.

Family Nycteribiidae (wingless bat flies). Always wingless; thorax weakened and de-sclerotized; live exclusively on bats; scarcely recognizable as flies.

Critical appraisal. Although there is general agreement concerning major groups of Diptera, disputes concerning relatively minor problems are not uncommon. After extensive study of relationships among families, probable lines of evolution within the order were traced in 1958. The order was surveyed according to the evidence of paleontology, and many fossil flies were illustrated in 1964; this resulted in subdividing the order into an unusually large number of families. Evolution of flesh-feeding maggots and classification and probable evolution of Oestridae has also been investigated. (H.O.)

Siphonaptera (fleas)

Fleas, which comprise the order Siphonaptera, are parasites that live on the exterior of the host (*i.e.,* are ectoparasitic); they are bloodsucking insects, important carriers of disease, and sometimes serious pests. As the chief agent transmitting the Black Death (bubonic plague) in the Middle Ages, they were an essential link in the chain of events that killed a quarter of the population of Europe.

GENERAL FEATURES

Fleas are small, wingless insects with a tough cuticle bearing many bristles and frequently combs (ctenidia) of broad, flattened spines. The adult flea varies from about 0.1 to 1 centimetre (0.039 to 0.39 inch) in length and feeds exclusively on the blood of mammals (including man) and birds. With more than 1,600 species and subspecies known, the order is still a small one compared with many other groups of insects; however, it is widely distributed. The rat flea and the mouse flea have been carried all over the world by man; native species of fleas are found in polar, temperate, and tropical regions.

IMPORTANCE

Infestation by fleas may cause severe inflammation of the skin and intense itching. Although many animals acquire partial immunity after constant or repeated attacks, occasionally individuals (especially man) become sensitized after exposure and develop allergies. Species that attack man and livestock include the cat flea (*Ctenocephalides felis*), the so-called human flea (*Pulex irritans*), the dog flea (*Ctenocephalides canis*), the sticktight flea (*Echidnophaga gallinacea*), and the jigger, or chigoe, flea (*Tunga penetrans*). Poultry may be parasitized by the European chicken flea (*Ceratophyllus gallinae*) and (in the U.S.) by the western chicken flea (*Ceratophyllus niger*).

Certain fleas that feed primarily on rodents or birds sometimes attack man, particularly in the absence of their usual prey. When rats are dying of bubonic plague, their hungry fleas, themselves infected with plague bacilli and seeking food elsewhere, can transmit the disease to man, especially in buildings heavily infested with rats. The Oriental rat flea (*Xenopsylla cheopis*) is the most efficient carrier of plague, but other species of fleas (*e.g., Nosopsyllus fasciatus, Xenopsylla brasiliensis, Pulex irritans*) can also transmit the disease to man. Although there are occasional cases of plague in tropical regions, the disease in man has been controlled by early diagnosis and antibiotics. Plague (sylvatic plague), a widespread disease in hundreds of species of wild rodents throughout the world, is maintained by fleas that parasitize these animals. More than 100 species of fleas are known to be infected by the plague bacillus, and an additional ten species are implicated as carriers of the classic type of urban plague. (See INFECTIOUS DISEASES.)

Fleas as carriers of disease

Fleas, particularly *Xenopsylla cheopis,* are believed to be the principal carriers of murine typhus, a rickettsial disease of man; rats and mice are sources of infection. Fleas are considered important in the maintenance and spread of many locally restricted infections among rodents and other mammals, including tularemia and Russian spring-summer encephalitis. Fleas transmit myxomatosis, a viral disease of rabbits used deliberately to control rabbits in areas where they are severe pests (*e.g.,* in Australia). Fleas are probable carriers of a filarial worm of dogs and serve as the intermediate host of a common tapeworm (*Dipylidium caninum*) of dogs and cats (and occasionally children). If heavily infested, animals can suffer severe damage or be killed by the effects of flea bites and considerable loss of blood. Fleas are subject to parasitism by external mites; internal nematode worms; and bacterial, fungal, and protozoan infections.

The female chigoe, or jigger, flea (not to be confused with "chigger," a larval mite), burrows into the skin of its host, generally on the feet, and lives within a cyst that forms around it. Intense itching accompanies the development and enlargement of the cyst as the abdomen of the pregnant flea swells to the size of a pea; secondary infections may constitute serious complications.

NATURAL HISTORY

Life cycle. Details of the life cycle are available for a few species only. The four stages are the egg, larva, pupa, and adult. Pearly white, oval eggs are deposited on the body of, or in the nest or habitat of, the host animal. The larva, a small legless caterpillar, feeds on organic debris, such as dried excrement, dried bits of skin, dead mites, or dried blood found in the host's nest. The parent fleas pass freshly imbibed blood rapidly through their gut to produce fecal matter for the nourishment of their offspring, which is essential to the successful metamorphosis of certain species of flea larvae. After three (exceptionally, two)

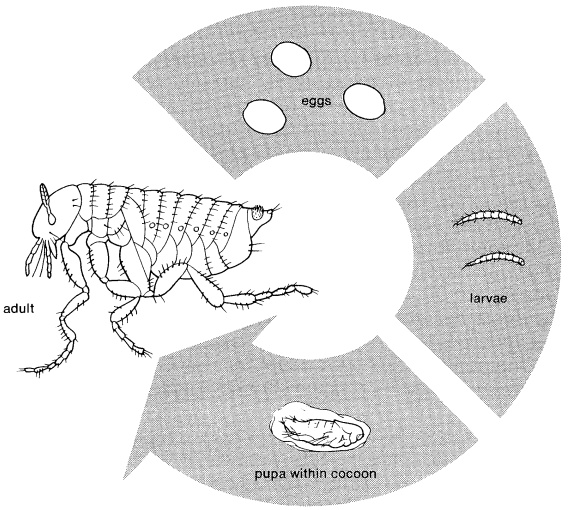


Figure 51: Stages in the life cycle of fleas.

molts, the larva spins a silk cocoon that includes debris from the nest; this is the beginning of the pupal stage. The pupa emerges as an adult some days or months later. Depending upon the species or environmental conditions, the time required for a complete life cycle varies from two weeks to several months. The life-span of the adult flea varies from a few weeks (*e.g., Echidnophaga gallinacea*) to a year or more (*Pulex irritans*). The life cycle of the European rabbit flea (*Spilopsyllus cuniculi*) and its host are perfectly synchronized. The sexual development of male and female fleas is under the direct control of the rabbit's sex hormones. Thus, the eggs of the female flea mature successfully only if she is feeding on a pregnant doe rabbit. When the young rabbits are born, both flea sexes now mature and leave the parturient doe for the nestlings and the nest, where they copulate and lay their eggs, thus assuring the flea larvae a suitable habitat for development. If the sex hormones of the female rabbit are controlled artificially by the administration of synthetic progestins (contraceptives), the sexual development of the female flea is also arrested. Although no similar case is yet known among other species of fleas, it has been recorded that rat fleas are less fertile if fed on baby mice than on their parents and that the mouse flea (*Leptopsylla segnis*) is more prolific if reared on family units rather than on individual adult mice. It seems likely, therefore, that the influence of the hosts' hormones is more widespread than hitherto suspected.

Relation to life cycle of host

Parasitism. Some fleas (*e.g.,* shrew fleas and rabbit fleas) are highly host specific; other species parasitize a variety of mammals. The cat flea infects not only the

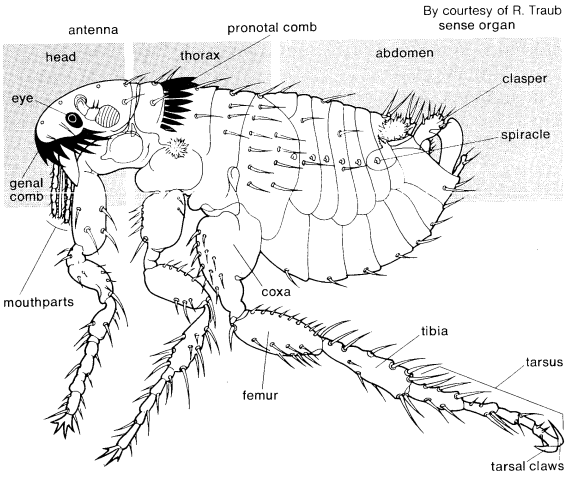


Figure 52: Body plan of the cat flea.

domestic cat but dogs, foxes, civets, mongooses, opossums, leopards, and other mammals including man if accustomed hosts are not available. Related mammals tend to be parasitized by fleas that are themselves related. Thus the rabbitlike pikas (*Ochotona*) living in the Rocky Mountains are infested with two peculiar genera of fleas that occur also on pikas in the mountains of Asia; this indicates a close phylogenetic relationship between these geographically separated hosts. Bird fleas have adapted themselves in relatively recent times to their avian hosts. These share several common adaptive features; one of the most obvious responses to the new environment is an increase in the number of comb spines on the upper surface of the thorax.

It is interesting that monkeys and apes do not harbour fleas, nor do horses or the majority of ungulates. The most heavily parasitized group of mammals are the rodents (*e.g.,* rats, mice, squirrels). Their habit of nest building in holes favours development of flea larval stages; animals with no permanent home tend to have fewer fleas.

Although both flea sexes feed avidly and repeatedly on blood (a single male exception never feeds), they survive for various periods away from the host. The rabbit flea, for example, can live for nine months at temperatures around the freezing point without feeding.

FORM AND FUNCTION

Anatomically, adult fleas are a rather uniform but distinctive group, with many interesting modifications and few obvious links with other orders. The compressed body and backwardly projecting spines or combs enable them to move swiftly through hairs or feathers of the host. The mouthparts, modified for sucking blood, include barbed stylets that aid both in penetration of the flea into the host skin and in attachment of species that spend long periods fixed to the host (*e.g.,* the sticktight fleas). Generally fleas that live on diurnal hosts have well-developed eyes; species that parasitize subterranean hosts (*e.g.,* moles or nocturnal animals like bats) have poorly developed eyes or none at all. The most impressive adaptations are highly developed jumping legs and the powerful muscles that control them. During their evolution, fleas, like the majority of parasitic insects, lost their wings. Certain parts of the flight mechanism have been retained, however, and incorporated in the jumping mechanism. Because fleas are able to leap horizontal or vertical distances 200 times their body length and to develop an acceleration of 200 gravities, they have been described as insects that fly with their legs. Certain species that live in nests high above the ground or in other unusual habitats crawl rather than jump. An incidental use of the unusual strength of fleas is in "flea circuses" in which they pull miniature carts and perform other feats.

Locomotion

CONTROL

In controlling fleas it is best to treat simultaneously both the breeding sites of fleas and the infested host, since, as mentioned earlier, the larval and pupal stages usually develop away from the host's body. For infested animals a commercial dust, spray, dip, or aerosol containing an insecticide such as pyrethrum, rotenone, malathion, or methoxychlor is used. Methoxychlor and malathion may continue to kill fleas for a few weeks after application. The long-acting chlorinated hydrocarbons, such as lindane and chlordane, can be applied to the fur of certain infested animals (but not to cats or dogs). In some regions, however, fleas have become resistant to these hydrocarbon insecticides, and new toxicants are required. Organic phosphorus compounds offer promise as systemic insecticides. When animals to be protected are fed these compounds, the fleas feeding on them die. For the control of larval and adult fleas away from the host, dieldrin, lindane, or malathion may be applied to the pens and haunts of the affected animals. Repellents containing benzyl benzoate, phthalate compounds, or diethyltoluamide are effective in preventing attack by fleas.

EVOLUTION, PALEONTOLOGY, AND CLASSIFICATION

The Siphonaptera form a small group of insects probably descended from an ancestor of the Mecoptera (scorpionflies), with which they share certain characteristics. Both groups have a spined gizzard (proventriculus), sexual differences in the number of ganglia in the ventral nerve cord, six rectal glands, and a simple type of ovary. The males have a similar type of spermatozoon, unique in the phylum Arthropoda, in which a motile flagellum, or tail, lacking the outer ring of nine tubules, is deployed around the mitochondria (cell organelles). A fossil flea, discovered in Australia, is claimed to be 200,000,000 years old. Two other known fossil fleas come from the Baltic amber (Oligocene) and closely resemble "modern" fleas.

Distinguishing taxonomic features. Although taxonomic separation of these groups rests upon a combination of superficially trivial morphological characteristics, they do reflect the fundamental differences between the groups. At the family or generic levels, classification is based principally upon shape of head and thorax, arrangement of combs, modifications of the male copulatory organ and female reproductive organs, general chaetotaxy (arrangement of bristles), and other characteristics.

Annotated classification. Fleas of today can be divided into three superfamilies: the Pulicoidea, Malacopsylloidea, and Ceratophylloidea. The Pulicoidea are generally adapted more to life on the host's body than are the latter two, which spend more time in the nest.

Order Siphonaptera (fleas)

Small, wingless, laterally compressed insects; adults ectoparasitic on warm-blooded vertebrates; eyes present or lacking; antennae short and stout, reposing in grooves; mouthparts adapted to piercing and sucking, maxillary and labial palps present; thoracic segments free; legs with large coxae, tarsi five-segmented; larvae elongated, legless, caterpillar-like; pupae with appendages free, enclosed in cocoons.

Superfamily Pulicoidea. Outer internal ridge of midcoxa of leg absent; mesonotum (dorsal sclerite of mesothorax) without pseudosetae under collar; metepimeron extending far upward, with first abdominal spiracle placed above the metepisternum; metanotum (dorsal sclerite of metathorax) and abdominal terga (dorsal surface) without apical spines or spinelets; spiracles circular; abdominal terga (segments II-VII) with, at most, one row of bristles; no bristles above spiracle of terga VIII; sensillium with either 8 or 14 pits each side; hind tibia of leg without an apical tooth outside; one family Pulicidae with genera *Pulex, Xenopsylla, Tunga,* and others.

Superfamily Malacopsylloidea. Rod of the lateral sclerites of the mesothorax (mesopleural rod) not forked; anterior tentorial arm present in head; no genal combs (spines on the anteroventral border of head) or pronotal combs (spines on the dorsal sclerite of the prothorax); ventral margin of pronotum never bilobed; fifth tarsal segment of leg always with four pairs of bristles (Neotropical region and extreme south Nearctic); two families, Malacopsyllidae and Rhopalopsyllidae.

Superfamily Ceratophylloidea. All fleas that do not possess combinations of characters listed above are included with the Ceratophylloidea, 12 families are included.

Critical appraisal. There is general agreement among flea specialists regarding the basic classification of the order, but it is probable that the family Hystrichopsyllidae will be accorded the rank of a superfamily.

(M.L.R./R.T.)

BIBLIOGRAPHY

Insects. *General and classification:* A.D. IMMS, *A General Text-Book of Entomology,* 9th ed. entirely rev. by O.W. RICHARDS and R.G. DAVIES (1957); V.B. WIGGLESWORTH, *The Life of Insects* (1964); C.T. BRUES, A.L. MELANDER, and F.M. CARPENTER, *Classification of Insects,* 2nd rev. ed. (1954).

Insect form: R.E. SNODGRASS, *Principles of Insect Morphology* (1935); J.S. KENNEDY (ed.), *Insect Polymorphism* (1963).

Insect physiology and biochemistry: M. ROCKSTEIN (ed.), *The Physiology of Insecta,* 3 vol. (1964–65); V.B. WIGGLESWORTH, *The Physiology of Insect Metamorphosis* (1954), *The Principles of Insect Physiology,* 6th ed. rev. (1965), *Insect Hormones* (1970); V.G. DETHIER, *The Physiology of Insect Senses* (1963); D. GILMOUR, *The Biochemistry of Insects* (1961); V.J.A. NOVAK, *Insect Hormones: The Physiology, Morphology, and Phylogeny of the Insect Endocrines* (1966); W. MORDUE *et al., Insect Physiology* (1981).

Insect behaviour and communications: C.G. JOHNSON, *Migration and Dispersal of Insects by Flight* (1969); C.T. BRUES, *Insect Dietary: An Account of the Food Habits of Insects* (1946); J.W.S. PRINGLE, *Insect Flight* (1957); P.T. HASKELL, *Insect Sounds* (1961); LORUS J. MILNE and MARGERY MILNE, *Insect Worlds* (1980), a popularly written description of insect adaptability and use of the environment; MICHAEL D. ATKINS, *Introduction to Insect Behavior* (1980).

Economic and ecological importance: C.L. METCALF and W.P. FLINT, *Destructive and Useful Insects: Their Habits and Control,* 4th ed. (1962); J.R. BUSVINE, *Insects and Hygiene: The Biology and Control of Insect Pests of Medical and Domestic Importance,* 2nd ed. (1966); R.M. GORDON and M.M.J. LAVOIPIERRE, *Entomology for Students of Medicine* (1961); T.R.E. SOUTHWOOD, *Ecological Methods: With Particular Reference to the Study of Insect Populations* (1966).

Apterygote. K. CHRISTIANSEN, "Bionomics of Collembola," *A. Rev. Ent.,* 9:147–178 (1964), a concise review of the early literature; A.D. IMMS, *A General Textbook of Entomology,* 9th ed. rev. by O.W. RICHARDS and R.G. DAVIES (1957), the standard textbook; S.M. MANTON, "Introduction to Classification of Arthropoda," in R.C. MOORE (ed.), *Treatise on Invertebrate Paleontology,* pt. R, *Arthropoda 4* (1969), a recent brief coverage of the subject; Z. MASSOUD, "Monographie des Neanuridae, Collemboles Poduromorphes à pièces buccales modifiées," *Biologie de l'Amérique Australe,* 3:7–399 (1967), an important revision and classification of Collembola with sucking mouthparts; J. PACLT, *Biologie der primär flügellosen Insekten* (1956), an extensive review of earlier literature and research; J.T. SALMON, *An Index to the Collembola,* 3 vol. (1964–65); A.G. SHAROV, *Basic Arthropodan Stock: With Special Reference to Insects* (1966), an excellent work that is based on personal palaeontological research and emphasizes phylogenetic relationships from worms to winged insects; S.L. TUXEN, *The Protura* (1964), a taxonomic coverage of this group.

Ephemeroptera. L. BERNER, *A Tabular Summary of the Biology of North American Mayfly Nymphs (Ephemeroptera)* (1959); B.D. BURKS, *The Mayflies, or Ephemeroptera, of Illinois* (1953), chiefly taxonomic, with good illustrations; A.E. EATON, *A Revisional Monograph of the Ephemeridae, or Mayflies,* 6 pt. (1883–88), taxonomy of world animal life with historical value; G.F. EDMUNDS, JR., "The Principles Applied in Determining the Hierarchic Level of the Higher Categories of Ephemeroptera," *Syst. Zool.,* 11:22–31 (1962), and "Biogeography and Evolution of Ephemeroptera," *A. Rev. Ent.* (1972); J.W. and F.A. LEONARD, *Mayflies of Michigan Trout Streams* (1962), biology and ecology with colour plates; J.G. NEEDHAM, J.R. TRAVER, and Y.C. HSU, *The Biology of Mayflies* (1935, reprinted 1969), comprehensive handbook.

Odonata. *General:* R.J. TILLYARD, *The Biology of Dragonflies—Odonata or Paraneuroptera* (1917), a classic, college-level textbook emphasizing systematics and functional morphology; P.S. CORBET, *A Biology of Dragonflies* (1962), a college-level textbook emphasizing ecology and behaviour.

Regional: C. LONGFIELD, *The Dragonflies of the British Isles* (1949); P.S. CORBET, C. LONGFIELD, and N.W. MOORE, *Dragonflies* (1960); F.C. FRASER, *Odonata,* vol. 1–3, *The Dragonflies of India,* in "The Fauna of British India Series" (1933–36); E.M. WALKER, *The Odonata of Canada and Alaska,* vol. 1–2 (1953–58); J.G. NEEDHAM and M.J. WESTFALL, *A Manual of the Dragonflies of North America (Anisoptera)* (1954); E.C.G. PINHEY, *The Dragonflies of Southern Africa* (1951); P.-A. ROBERT, *Les Libellules (Odonates)* (1958), a Swiss work on western European Odonata. Works listed in the above section include identification manuals incorporating biological observations.

Technical: F.C. FRASER, *A Reclassification of the Order Odonata* (1957), the definitive work on classification; S. ASAHINA, *A Morphological Study of a Relic Dragonfly Epiophlebia superstes Selys (Odonata, Anisozygoptera)* (1954), a comprehensive study of the only living genus of Anisozygoptera; P.S. CORBET, "The Life History of the Emperor Dragonfly, *Anax imperator* Leach (Odonata: Aeshnidae)," *J. Anim. Ecol.,* 26:1–69 (1957), with emphasis on larval growth, metamorphosis, and emergence; M.E. JACOBS, "Studies on Territorialism and Sexual Selection in Dragonflies," *Ecology,* 36:566–586 (1955), a detailed treatment of sexual behaviour and egg-laying; C.H. KENNEDY, "The Relation of American Dragonfly-Eating Birds to Their Prey," *Ecol. Monogr.,* 20: 103–142 (1950), an exhaustive study of bird predation; F.X. WILLIAMS, "Biological Studies in Hawaiian Water-Loving Insects, Pt. II, Odonata or Dragonflies," *Proc. Hawaii. Ent. Soc.,* 9:273–349 (1936), a description of the ecology of the only known species with a terrestrial larva.

Orthopteran. R.D. ALEXANDER, "Acoustical Communication in Arthropods," *A. Rev. Ent.,* 12:495–526 (1967), with many orthopteran examples; "Life Cycle Origins, Speciation, and Related Phenomena in Crickets," *Qt. Rev. Biol.,* 43:1–41 (1968), a valuable summary of recent advances in cricket biology; and with D. OTTE, "The Evolution of Genitalia and Mating Behavior in Crickets (Gryllidae) and Other Orthoptera," *Misc. Publs. Mus. Zool. Univ. Mich.,* 133:5–62 (1967), well illustrated; E.D. BALL *et al.,* "The Grasshoppers and Other Orthoptera of Arizona," *Tech. Bull. Univ. Ariz.,* 93:257–373 (1942), well illustrated; W.S. BLATCHLEY, *Orthoptera of Northeastern America* (1920), an important guide, some parts out of date; I.J. CANTRALL, "The Ecology of the Orthoptera and Dermaptera of the George Reserve, Michigan," *Misc. Publs. Mus. Zool. Univ. Mich.,* 54:1–182 (1943), a survey of 78 species of orthopterans; R.F. CHAPMAN, *The Insects: Structure and Function* (1969), with many orthopteran examples; L. CHOPARD, "Dictyoptères," "Notoptères," "Chéleutoptères," and "Orthoptères," in *Traité de zoologie,* vol. 9 (1949), a general treatise on Orthoptera; V.M. DIRSH, "A Preliminary Revision of the Families and Subfamilies of Acridoidea," *Bull. Br. Mus. Nat. Hist. (Ent.),* 10:351–419 (1961), the current classification of grasshoppers; "The Post-Embryonic Ontogeny of Acridimorpha (Orthoptera)," *EOS, Madr.,* 43:413–514 (1968), basic information on growth changes in grasshoppers; S.K. GANGWERE, "A Monograph on Food Selection in Orthoptera," *Trans. Am. Ent. Soc.,* 87:67–230 (1961), a summary of techniques for studying food preferences of orthoptera and a review of information on the food eaten by various orthopteran groups, chiefly U.S. species; A.B. GURNEY, "Praying Mantids of the United States: Native and Introduced," *A. Rep. Smithson. Instn. for 1950,* pp. 339–362 (1951), a semipopular review with photos of several species; DANIEL OTTE, *The North American Grasshoppers,* 3 vol. (1981–), a revision of the taxonomic order. D.M. GUTHRIE and A.R. TINDALL, *The Biology of the Cockroach* (1968), emphasis on the physiology of cockroach species commonly used as laboratory animals; P.T. HASKELL, *Insect Sounds* (1961), many references to orthoptera; M. HEBARD, "The Dermaptera and Orthoptera of Illinois," *Bull.*

Ill. St. Nat. Hist. Surv., 20:125–279 (1934), keys to species, illustrated; J.R. HELFER, *How to Know the Grasshoppers, Cockroaches and Their Allies* (1963), a semipopular handbook; A.D. IMMS, *A General Textbook of Entomology*, 9th ed. rev. by O.W. RICHARDS and R.G DAVIES (1957), an important but detailed text; F.A. MCKITTRICK, "Evolutionary Studies of Cockroaches," *Mem. Cornell Univ. Agric. Exp. Stn. 389* (1964), classification of cockroaches; K. PRINCIS, "Blattariae," in *Orthopterorum Catalogus*, 8 pt. (1962–71), a catalog of living cockroaches, of use to researchers in systematics and distribution; D.R. RAGGE, *Grasshoppers, Crickets and Cockroaches of the British Isles* (1965), a well-illustrated standard work; J.A.G. REHN, "Man's Uninvited Fellow Traveler, the Cockroach," *Sci. Mon.*, 61; 265–276 (1945), popular and factual; with H.J. GRANT, *A Monograph of the Orthoptera of North America (North of Mexico)*, vol. 1 (1961), detailed and authoritative; L.M. ROTH, "Oothecae of the Blattaria," *A. Ent. Soc. Am.*, 61:83–111 (1968), an illustrated review; "The Evolution of Male Tergal Glands in the Blattaria," *A. Ent. Soc. Am.*, 62:176–208 (1969), a review.

"Evolution and Taxonomic Significance of Reproduction in Blattaria," *A. Rev. Ent.*, 15:75–96 (1970), a summary of reproductive habits; with T. EISNER, "Chemical Defenses of Arthropods," *A. Rev. Ent.*, 7:107–136 (1962), concerned with secretory glands that produce repugnant substances; and with E.R. WILLIS, *The Biotic Associations of Cockroaches* (1960), a documented summary; A.G. SHAROV, in *Trans. Paleont. Inst., Moscow*, 118:1–208 (1968; Eng. trans., "Phylogeny of the Orthopteroidea," pp. 1–251, 1971), evolutionary development; B.P. UVAROV, *Grasshoppers and Locusts*, vol. 1 (1966), the first of a 2-volume book on this subject; T.J. WALKER, "Cryptic Species Among Sound-Producing Ensiferan Orthoptera (Gryllidae and Tettigoniidae)," *Q. Rev. Biol.*, 39:345–355 (1964), emphasis on the large number of orthoptera species reproductively isolated by inconspicuous calling songs; F.E. ZEUNER, *Fossil Orthoptera Ensifera*, 2 vol. (1939), a review with excellent illustrations.

Isoptera. K. KRISHNA and F.M. WEESNER (eds.), *Biology of Termites*, 2 vol. (1969–70), an advanced treatise, dealing with the fundamental aspects of termite biology; S.H. SKAIFE, *Dwellers in Darkness* (1961), a popular account of termites for the informed layman; W.V. HARRIS, *Termites: Their Recognition and Control* (1961); and T.E. SNYDER, *Our Enemy the Termite*, rev. ed. (1948), two general works, with emphases on termite control.

Phthiraptera. G.F. FERRIS, "The Sucking Lice," *Mem. Pacif. Cst. Ent. Soc.*, vol. 1 (1951), an outline of the morphology, growth, and identification of the Anoplura, with a host-parasite list; G.H.E. HOPKINS, "The Host-associations of the Lice of Mammals," *Proc. Zool. Soc. Lond.*, 119:387–604 (1949), an account of the biology and methods of preservation, examination, and distribution of the mammalian lice, with a list of mammals and their lice; and with T. CLAY and G. TIMMERMANN, *First Symposium on Host Specificity among Parasites of Vertebrates* (1957), papers on the distribution of the Anoplura and Mallophaga of birds and mammals, with a discussion on host-parasite relationships; S. VON KELER, "Bibliographie der Mallophagen," *Mitt. Zool. Mus.* 36:147–403 (1960), a full bibliography of the literature on the Mallophaga arranged chronologically, together with an author and subject index; M. ROTHSCHILD and T. CLAY, *Fleas, Flukes and Cuckoos: A Study of Bird Parasites*, 3rd ed. (1957), a popular account of the Mallophaga of birds; H. ZINSSER, *Rats, Lice and History* (1935, reprinted 1965), a readable account of the scientific and historical aspects of *Typhus*.

Thysanoptera. The literature on Thysanoptera is widely scattered throughout numerous entomological journals with occasional summaries given in monographs or lists. The monographs are predominantly faunal studies, though they often include regional revisions of some of the taxa. Only a few complete generic revisions have been published. Some of the latest, though not necessarily recent, faunal monographs are: S.F. BAILEY, "The Thrips of California, Part 1: Suborder Terebrantia," *Bull. Calif. Insect Surv.*, vol. 4, no. 5, pp. 143–220 (1957); H.E. COTT, *Systematics of the Suborder Tubulifera (Thysanoptera) in California* (1956); L.J. STANNARD, "The Thrips, or Thysanoptera, of Illinois," *Bull. Ill. Nat. Hist. Surv.*, vol. 29, no. 4, pp. 215–552 (1968); G.D. MORISON, "Thysanoptera of the London Area," *Lond. Nat.*, reprint no. 59 (1947–49). Two comprehensive catalogues and annotated synopses are C.F. JACOT-GUILLARMOD, "Catalogue of the Thysanoptera of the World (Part 1)," *Ann. Cape Provincial Museums (Nat. Hist.)* (1970); and L.J. STANNARD, *The Phylogeny and Classification of the North American Genera of the Suborder Tubulifera (Thysanoptera)* (1957). Proposals for the classification of the thrips that have received general acceptance are those by H. PRIESNER, "Zur Vergleichender Morphologie des Endothorax der Thysanoptera," *Zool. Anz.*, vol. 159, no. 7–8, pp. 159–167 (1957), and "Das System der Tubulifera (Thysanoptera)," *Anz. öst. Akad. Wiss.*, 13:283–296 (1960). Morphological treatments of the Thysanoptera are found in R.G. DAVIES, "Observations on the Morphology of the Head and Mouthparts in the Thysanoptera," *Proc. R. Ent. Soc. Lond.* (A), vol. 33, no. 7–9, pp. 97–106 (1958); and P. PESSON, "Ordre des Thysanoptera," in P. GRASSE, *Traité de zoologie*, vol. 10, no. 2, pp. 1805–1869 (1951). A source of ecological information on thrips is that by L. CEDERHOLM, *Ecological Studies on Thysanoptera* (1963).

Homoptera. *A Bibliography of the Homoptera (Auchenorrhyncha)*, 2 vol. (1943), *General Catalogue of the Homoptera* (1954–68), *Fulgoridae of Eastern North America* (1923), taxonomic treatment with keys to families and excellent illustrations; K.C. DOERING, "Synopsis of the Family Cercopidae (Homoptera) in North America," *J. Kans. Ent. Soc.*, 3:53–64, 81–108 (1930), a synoptic treatment of all known North American cercopids; P.B. LAWSON, "The Cicadidae of Kansas," *Kans. Univ. Sci. Bull.*, 12: 309–376 (1920), taxonomic and distributional discussion of Kansas cicadas; P.W. OMAN, *The Nearctic Leafhoppers (Homoptera: Cicadellidae): A Generic Classification and Check List* (1949); H. OSBORN, *The Membracidae of Ohio* (1940), taxonomy of treehoppers of Ohio and proximity; D.L. CRAWFORD, *Monograph of the Jumping Plant Lice or "Psyllidae" of the New World* (1914), a taxonomy of jumping plant lice of North America; G.F. FERRIS, *Atlas of the Scale Insects of North America*, 4 vol. (1937–42), a comprehensive study of genera and species, excellent illustrations; A.D. MACGILLIVRAY, *The "Coccidae"* (1921), synoptic treatment of scale insects; F.C. HOTTES and T.H. FRISON, *The Plant Lice or Aphididae of Illinois* (1931), aphid taxonomy; J.W. EVANS, *A Natural Classification of Leafhoppers (Jassoidea, Homoptera)*, 3 pt. (1946–47), a taxonomy of leafhoppers with emphasis on higher categories.

Heteroptera. W.S. BLATCHLEY, *Heteroptera of Eastern North America* (1926), introduction to Heteroptera, especially those of eastern North America; E.A. BUTLER, *A Biology of the British Hemiptera-Heteroptera* (1923), classic work with biological data; W.E. CHINA and N.C.E. MILLER, "Checklist and Keys to the Families and Subfamilies of the Hemiptera-Heteroptera," *Bull. Br. Mus. (Nat. Hist.) Ent.*, 8:1–45 (1959), phylogenetic dendrogram and assignment of names, includes a bibliography of source studies; R.H. COBBEN, *Evolutionary Trends in Heteroptera, part 1, Eggs: Architecture of the Shell, Gross Embryology and Eclosion* (1968), structural and developmental details with interpretive comments; W.L. DISTANT, *Rhynchota*, in *Fauna of British India* (1902–18), five-volume study of Asian Heteroptera, *Biologia Centrali Americana Insecta*, vol. 1, *Rhynchota Heteroptera* (1881–1909), classic study of tropical American fauna; P.T. HASKELL, *Insect Sounds* (1961), readable, though in places technical, introduction to a fascinating topic; H.B. HUNGERFORD, "The Biology and Ecology of Aquatic and Semiaquatic Hemiptera," *Kans. Univ. Sci. Bull.*, 11:1–265 (1919), classic and basic study; N.C.E. MILLER, *The Biology of the Heteroptera* (1956); H.M. PARSHLEY, *A Bibliography of North American Hemiptera-Heteroptera* (1927); T.R.E. SOUTHWOOD and D. LESTON, *Land and Water Bugs of the British Isles* (1959), keys and notes through species; W. STICHEL, *Illustrierte Bestimmungstabellen der Wanzen, 2, Europa* (1955–62), four volumes with keys to European species and a checklist for Palearctic Region; R.L. USINGER, *Monograph of Cimicidae (Hemiptera-Heteroptera)*, vol. 7 (1966), incorporates modern methodology in monographic study of important bedbugs; E.P. VAN DUZEE, *Catalogue of the Hemiptera of America North of Mexico* (1917), important list and literature summary of North American Heteroptera; H. WEBER, *Biologie der Hemipteren* (1930), important source for original and compiled information.

Neuropteran. F.J. KILLINGTON, *Monograph of the British Neuroptera*, 2 vol. (1936–37), an excellent monograph; C.L. WITHYCOMBE, "Some Aspects of the Biology and Morphology of the Neuroptera, with Special Reference to the Immature Stages and their Possible Phylogenetic Significance," *Trans. Ent. Soc. Lond.*, pt. 3–4, pp. 303–411 (1924–25), a good source for information.

Trichoptera. F.C.J. FISCHER, *Trichopterorum Catalogus* (1960–), a continuing checklist of world Trichoptera; N.E. HICKIN, *Caddis Larvae: Larvae of the British "Trichoptera"* (1967), an excellent, comprehensive work on caddis; M.E. MOSELY and D.E. KIMMINS, *The Trichoptera (Caddis-Flies) of Australia and New Zealand* (1953), a basic study of fauna that includes species of importance in world classification; H.H. ROSS, "The Evolution and Past Dispersal of the Trichoptera," *A. Rev. Ent.*, 12:169–206 (1967), a brief work including a revolutionary reclassification of the group; and GARY LAFONTAINE, *Caddisflies* (1981), for the general reader.

Lepidoptera. The literature on Lepidoptera is enormous and includes books and articles published in nearly all countries. The following gives a representative sample, including major works with extensive bibliographies.

General works: "Lepidoptera," in A.D. IMMS, *A General Textbook of Entomology*, 9th ed. rev. by O.W. RICHARDS and R.G. DAVIES (1957); and "Lepidoptera," in D.J. BORROR and D.M. DE

LONG, *An Introduction to the Study of Insects,* rev. ed. (1964), give a key to families of Lepidoptera in North America; E.B. FORD, *Butterflies* (1945) and *Moths* (1955), are replete with details of natural history, ecology, physiology, and evolution, chiefly of British species; M. HERING, *Biologie der Schmetterlinge* (1926), covers life histories and ecology, particularly of European groups; A.B. KLOTS, *The World of Butterflies and Moths* (1958), covers natural history and ecology, with striking illustrations. See also P.J. PORTIER, *La biologie des Lépidoptères* (1949), for many details of the biology; C.B. WILLIAMS, *The Migration of Butterflies* (1930); and F.A. URQUHART, *The Monarch Butterfly* (1960). *Lepidopterorum Catalogus,* published in many parts by various authors, is the definitive scientific listing of all genera and species, with bibliographies. PAUL E.S. WHALLEY, *Butterfly Watching* (1980), discusses behaviour for the general reader.

Regional works: (*North America*): P.R. EHRLICH *et al., How to Know the Butterflies* (1961), is an excellent handbook of the North American species; W.T.M. FORBES, *Lepidoptera of New York and Neighboring States,* 4 pt. (1923–1960), is the definitive work for eastern North America for technical classification, especially larvae, life histories, and foodplants; W.J. HOLLAND, *The Moth Book* (1914, reprinted 1968), and *The Butterfly Book,* rev. ed. (1931), are the long-time standards, with many colour illustrations; A.B. KLOTS, *Field Guide to the Butterflies of North America, East of the Great Plains* (1951), covers all the species of butterflies and skippers of eastern North America, with much on life histories, biology, and ecology, with many illustrations. (*Europe:*) The following are more or less local in scope, covering the butterflies and moths of various countries or regions: DAVID CARTER, *Butterflies and Moths in Britain and Europe* (1982); L.T. FORD, *A Guide to the Smaller British Lepidoptera* (1949); W. FORSTER and T.A. WOHLFAHRT, *Die Schmetterlinge Mitteleuropas,* 3 vol. (1954–60); S. HOFMEYER, *Die Danske Ugler* (1962); T.W. LANGER, *Nordens dagsommerfugle* (1958); F. NORDSTROM and E. WAHLGREN, *Svenska fjärilar* (1941); R. SOUTH, *The Moths of the British Isles* rev. ed., 2 vol. (1939); A. SPULER, *Die Schmetterlinge Europas und ihre Raupen,* 4 vol. (1908–10); and M. STURANI, *Vita delle Farfalle* (1947). (*Other regions:*) A. SEITZ *et al.* (eds.), *Macrolepidoptera of the World* (1906–), English, French, and German editions, 16 vol. (1–4 Palaearctic, 5–8 American, 9–12 Indo-Australian, 13–16 Ethiopian), was issued in parts, by many authors, with supplements; although out of date in much nomenclature and taxonomy, it is still the standard for most of the world for identification. The following are books covering specific regions giving many details about life histories and biologies and mostly illustrated very fully: D.A. SWANEPOEL, *The Butterflies of South Africa: Where, When, and How They Fly* (1953); G. VAN SON, *The Butterflies of Southern Africa* (1949–63); G.A. WATERHOUSE and G. LYELL, *The Butterflies of Australia* (1914); and M.A. WYNTER-BLYTH, *Butterflies of the Indian Region* (1957).

Coleoptera. *General:* In addition to information given in textbooks of entomology, the following works should be consulted: J.R. DIBB, *Field Book of Beetles* (1948); G. TAYLOR, *Some British Beetles* (1948); UNITED STATES DEPARTMENT OF AGRICULTURE, "Insects," *1952 Yearbook of Agriculture* (1952), many chapters of this work are devoted to pestiferous beetles—their ravages and control; A.G. BØVING and F.C. CRAIGHEAD, "An Illustrated Synopsis of the Principal Larval Forms of the Order Coleoptera," *Ent. Amer.,* vol. 11 (1930–31); J.H. FABRE, *Souvenirs entomologiques,* 10 vol. (1879–1907). Many of these detailed popular accounts of natural history are available in English. A standard anatomical and biological work, based upon a study of the water beetle *Dytiscus,* is E. KORSCHELT, *Der Gelbrand Dytiscus marginalis,* 2 vol. (1923–24). See also ROSS H. ARNETT, N.M. DOWNIE and H.E. JACQUES, *How to Know the Beetles,* 2nd ed. (1980); RICHARD E. WHITE, *A Field Guide to the Beetles of North America* (1983).

Advanced: R.A. CROWSON, *The Natural Classification of the Families of Coleoptera* (1955, reprinted 1967), a descriptive synopsis of all families of beetles, with illustrated key, and many new arrangements; W.W. FOWLER, *Coleoptera: General Introduction and Cicindelidae and Paussidae* (1912), a general treatment of morphology, including the internal anatomy of beetles, with a general discussion of organs, biology, and classification; J.T. LACORDAIRE, *Histoire naturelle des insectes: Genera des coléoptères, ou exposé méthodique et critique de tous les genres proposés jusqui'ici dans cet ordre d'insectes,* 12 vol. (1854–76, reprinted 1954–72), the standard work on all the families and genera of beetles known at the time, still extremely useful, with keys to all categories (in French).

Catalogues of the species of Coleoptera include the *Catalogus Coleopterorum Europeae,* 2nd ed. (1906) by L. HEYDEN, E. REITTER, and J. WEISE; C.W. LENG, *Catalogue of the Coleoptera of America, North of Mexico* (1920), with five supplements (1927–48); W. JUNK, *Coleopterorum Catalogus* (1910–39); R.E. BLACKWELDER (comp.), *Checklist of the Coleopterous Insects of Mexico, Central America, the West Indies, and South America,* 6 pt. (1944–57).

Hymenoptera. *Hymenoptera, general:* K.V. KROMBEIN *et al., Hymenoptera of America North of Mexico* (1958); *Catalog of Hymenoptera in America North of Mexico,* 2 vol. (1979).

Hymenoptera, evolution: S.I. MALYSHEV, *Genesis of the Hymenoptera and the Phases of Their Evolution* (1968; orig. pub. in Russian, 1966).

Social insects, general: C.D. and M.H. MICHENER, *American Social Insects* (1951); D.W. MORLEY, *The Evolution of an Insect Society* (1954); W.M. WHEELER, *The Social Insects: Their Origin and Evolution* (1928) E.D. WILSON, *The Insect Societies* (1971).

Hymenoptera, specific taxonomic groups: C.H. ANDREWES, *The Lives of Wasps and Bees* (1969); W.S. CREIGHTON, *The Ants of North America* (1950); H.E. EVANS, *The Comparative Ethology and Evolution of the Sand Wasps* (1966) and *Wasp Farm* (1963); M. LINDAUER, *Communication Among Social Bees,* 3rd ed. (1971); T.B. MITCHELL, *Bees of the Eastern United States,* 2 vol. (1960–62); C.R. RIBBANDS, *The Behaviour and Social Life of Honeybees* (1953); K. VON FRISCH, *The Dance Language and Orientation of Bees* (1967); W.M. WHEELER, *Ants: Their Structure, Development and Behaviour* (1910); PAUL D. HURD *et al., Principal Sunflower Bees of North America with Emphasis on the Southwestern United States* (1980).

Diptera. J.R. BUSVINE, *Insects and Hygiene,* 2nd ed. (1966), comprehensive survey of insects, hygiene, insecticides; C.N. COLYER and C.O. HAMMOND, *Flies of the British Isles,* 2nd ed. (1968), limited to British flies but has general application; C.H. CURRAN, *Families and Genera of North American Diptera,* 2nd rev. ed. (1965), useful keys for primary identification; M. DEMEREC (ed.), *Biology of Drosophila* (1950, reprinted 1965), accounts "normal" development so that experimental results can be assessed; W. HENNIG, *Die Larvenformen der Dipteren,* 3 vol. (1948–52), the only comprehensive account of fly larvae, with references to published descriptions; M.T. JAMES, "The Flies that Cause Myiasis in Man," *U.S. Dept. Agric. Misc. Publ. 631* (1947), useful handbook on larvae that infest the human body; E. LINDNER, *Die Fliegen der palaearktischen Region* (1924, parts still appearing), comprehensive account of flies with much information applicable throughout the world; I.M. MACKERRAS, "The Zoogeography of the Diptera," *Aust. J. Sci.,* 12:157–161 (1950); H. OLDROYD, *Diptera: Introduction and Key to Families,* in the "Handbooks for the Identification of British Insects Series" (1970); *Collecting, Preserving and Studying Insects,* rev. ed. (1973); *The Natural History of Flies* (1964), the only comprehensive account of flies in English; *Insects and Their World,* 2nd ed. (1966), simple account of the basic principles of structure and function in insects; and *Elements of Entomology: An Introduction to the Study of Insects* (1968), general reading about basic relationships of insects; L.S. WEST, *The Housefly: Its Natural History, Medical Importance and Control* (1951), comprehensive treatise on every aspect of housefly biology. More specialized references on classification may be found in W. HENNIG, "Die Familien der Diptera Schizophora und ihre phylogenetischen Verwandschaftsbeziehungen," *Beitr. Ent.,* 8:505–688 (1958); and in F. ZUMPT, "Some Remarks on the Classification of the Oestridae," *J. Ent. Soc. S. Africa,* 20:154–161 (1957).

Siphonaptera. G.P. HOLLAND, "The Siphonaptera of Canada," *Tech. Bull. Dep. Agric. Can. 70* (1949), the best work on identification, affinities and host-relations of relevant North American fauna; "Evolution, Classification and Host-Relations of Siphonaptera," *A. Rev. Ent.,* 9:123–146 (1964), a critical review of the literature; G.H.E. HOPKINS and M. ROTHSCHILD, *An Illustrated Catalogue of the Rothschild Collection of Fleas (Siphonaptera) in the British Museum (Natural History),* 5 vol. (1953–71), indispensable for global identification and general morphology of families treated to date; W.L. JELLISON, "Fleas and Disease," *A. Rev. Ent.,* 4:389–414 (1959), a summary of the medical importance of fleas; P.T. JOHNSON, "A Classification of the Siphonaptera of South America with Descriptions of New Species," *Mem. Ent. Soc. Wash.,* 5:1–299 (1957), sole opus on South American fleas, well-prepared and well-illustrated; M. ROTHSCHILD, "Fleas," *Sci. Am.* 213:44–53 (Dec. 1965), an account of interesting aspects of the biology and behaviour of fleas; "The Rabbit Flea and Hormones," in *The Biology of Sex,* ed. by A. ALLISON, pp. 189–199 (1967), reviews findings on hormonal interrelationships of the rabbit flea and its host; and with T. CLAY, *Fleas, Flukes and Cuckoos,* 3rd ed. (1957), information on bionomics and habits of bird fleas and other bird parasites; R. TRAUB, *Siphonaptera from Central America and Mexico* (1950), includes generic revisions and the first study on comparative anatomy of the aedeagus of fleas; *United States Department of Agriculture,* "Fleas—How to Control Them," Leaflet no. 392 (n.d.), one of a series of pamphlets on methods of control; ALLAN H. BENTON, *An Atlas of the Fleas of the Eastern United States* (1980).

Insurance

Insurance is a method of coping with risk. Its primary function is to substitute certainty for uncertainty as regards the economic cost of loss-producing events. Insurance may be defined more formally as a system under which the insurer, for a consideration usually agreed upon in advance, promises to reimburse the insured or to render services to the insured in the event that certain accidental occurrences result in losses during a given period.

This article is divided into the following sections:

Kinds of insurance 678
- Fire insurance 678
  - Standard provisions
  - Indirect losses
- Marine insurance 679
  - Ocean marine insurance
  - Inland marine insurance
- Property and liability insurance 680
  - Liability insurance
  - Theft insurance
  - Aviation insurance
  - Workers' compensation insurance
  - Credit insurance
  - Title insurance
  - Miscellaneous insurance
- Suretyship 683
  - Major types of fidelity bonds
  - Major types of surety bonds
- Life and health insurance 684
  - Life insurance
  - Private health insurance
- Group insurance 685
  - Group life insurance
  - Group health insurance
  - Group annuities

Insurance practice 687
- Underwriting and rate making 687
  - Underwriting principles
  - Rate making
- Legal aspects of insurance 688
  - Government regulation
  - Contract law
  - Liability law

Historical development of insurance 690
- England
- United States
- Soviet Union
- Japan
- Worldwide operations
- Mergers

Bibliography 691

Insurance relies heavily upon the "law of large numbers." In large homogeneous populations it is possible to estimate the normal frequency of common events such as deaths and accidents. Losses can be predicted with reasonable accuracy, and this accuracy increases as the size of the group expands. From a theoretical standpoint, it is possible to eliminate all pure risk if an infinitely large group is selected.

nsurable isk

From the standpoint of the insurer, an insurable risk must meet the following requirements:

1. The objects that are to be insured must be numerous enough and homogeneous enough to allow a reasonably close calculation of the probable frequency and severity of losses.

2. The insured objects must not be subject to simultaneous destruction. For example, if all the buildings insured by one insurer are in an area subject to flood, and a flood occurs, the loss to the insurance underwriter may be catastrophic.

3. The possible loss must be accidental in nature, and beyond the control of the insured. If the insured could cause the loss, the element of randomness and predictability would be destroyed.

4. There must be some way to determine whether a loss has occurred and how great that loss has been. This is why insurance contracts specify very definitely what events must take place, what constitutes loss, and how it is to be measured.

From the viewpoint of the insured person, an insurable risk is one for which the probability of loss is not so high as to require excessive premiums. What is "excessive" depends on individual circumstances, including the insured's attitude toward risk. At the same time, the potential loss must be severe enough to cause financial hardship if it is not insured against. Insurable risks include losses to property resulting from fire, explosion, windstorm, etc.; losses of life or health; and the legal liability arising out of use of automobiles, occupancy of buildings, employment, or manufacture. Uninsurable risks include losses resulting from price changes and competitive conditions in the market. Political risks such as war, punitive taxation, or currency debasement are usually not insurable by private parties but may be insurable by governmental institutions. Very often contracts can be drawn in such a way that an "uninsurable risk" can be turned into an "insurable" one through restrictions on losses, redefinitions of perils, or other methods.

Kinds of insurance

FIRE INSURANCE

The standard fire insurance policy is made up of two parts—the insuring agreement and the conditions and stipulations. The policy coverage is limited to the actual cash value, defined in practice as the replacement cost of the property less an allowance for depreciation. (The subtraction of depreciation from loss settlements may be waived by special provision under an endorsement known as replacement cost insurance.) Settlement must be made within a reasonable time following the loss but without any allowance for increased cost of repair by reason of building codes or other laws.

The basic fire insurance policy does not provide any coverage for indirect loss from fire such as might result from the interruption of a business. This peril can be covered under a separate contract, described below. The insuring agreement covers not only loss from fire and lightning but also losses to property that had to be removed from endangered premises.

Recovery under the basic standard policy is limited to the interest of the insured person. For example, if two individuals own a certain property but only one of them is insured, and the other is not named in the contract as an insured interest, the limit of recovery is that portion of the property owned by the insured person. The policy refers to all direct losses by fire. This implies that if other causes are instrumental in causing the loss, there will be no recovery except for losses due directly from the fire. Thus, if an explosion knocks a building down and it becomes a pile of rubble that later burns, the fire loss is limited to the loss of rubble. Only losses that are proximately caused by fire are paid.

Direct losses

Standard provisions. Fire insurance policies contain various specific provisions and stipulations. Among the more important provisions are the following:

1. Certain property is declared uninsurable or excluded from the basic coverage. This includes accounts, bills, currency, deeds, evidence of debt, money, or securities. Such property may be covered under a separate form.

2. The basic fire policy customarily excludes loss from certain perils such as war, invasion, insurrection, revolution, usurped power, neglect by the insured, and theft.

Thus if the insured sees the covered property threatened by a fire in a neighbouring building and fails to use preventive measures, recovery may be denied.

3. Insurance coverage is suspended at any time the insured is increasing the hazard or when the property is vacant or unoccupied beyond a period of 60 days, or during other periods as stated in the policy.

4. The policy may be cancelled by either party for any reason. If the insurer cancels it, however, the insured must be given prior notice.

5. A mortgagee's interest can be protected under the basic fire insurance policy, generally without extra charge.

6. If the insured has other policies of insurance on the same property, each policy contributes to the loss in the proportion that its limits bear to the total limits of all insurance on the property. This prevents the insured from collecting more than the actual cash loss by taking out a number of policies in separate companies.

7. The insured is required to give immediate written notice of the loss and to prove the amount of the loss within a specified period. The policy contains arbitration provisions in the event of disagreement over the amount.

8. The insurer is allowed to rebuild or replace the damaged property with other property of like kind and quality, rather than make a cash settlement. Abandonment of destroyed property to the insurer is not permitted.

9. Legal action to recover loss is not permitted unless all of the requirements for arbitration have been complied with and unless it is commenced within a limited period after the loss.

10. If a party other than the insured burns down a property and is held legally liable for it, the insurer reserves the right to bring a legal action against this third party in the name of the insured. This prevents the insured from collecting twice—once from the insurance company and once from a liable third party.

Co-insurance requirement

A widely used clause is the coinsurance clause, used on insurance covering commercial property. Its purpose is to prevent underinsurance. It requires the insured to carry enough coverage to equal a specified percentage, usually 80 to 90 percent, of the true replacement value of the property. If the insured fails to carry this minimum amount, recovery will be limited to only that portion of any fire loss equal to the ratio of the amount of insurance actually carried to the amount required under the coinsurance clause.

Fire insurance rates are influenced by four basic risk factors: the fire protection class of the city of location; the type of construction; the class of occupancy of the building; and the degree to which the building is exposed to losses originating outside it. For example, if a city has very little fire protection its insurance rates will be higher than those in a city that is superior in its water supply, fire department, protection and alarm systems, building codes, police protection, and so forth. A building made of fireproof material obviously requires a lower fire rate than one of frame construction. A manufacturer of explosive chemicals would receive a higher fire rate than the owner of a hardware store. If the building is located near another in which explosives are stored, it will carry a higher rate than a building not so exposed.

The standard fire policy insures only against the basic perils of fire, lightning, and removal from the premises. If the insured wishes additional perils to be covered, the form must either list additional perils or be drawn up on an "all risk" basis. One of the most common extensions of perils is known as the extended coverage endorsement, which provides coverage against windstorm, hail, explosion, riot, riot attending strike, civil commotion, aircraft, vehicles, and smoke. Each peril is specifically defined and delimited. For example, smoke from a fireplace may not be defined as "smoke."

Indirect losses. An entirely different branch of the fire insurance business has been developed to insure losses that are indirectly the result of one of the specified perils. A prominent example of this type of insurance is business interruption insurance. The insurer undertakes to reimburse the insured for lost profits or for fixed charges incurred as a result of direct damage. For example, a retail store might have a fire and be completely shut down for one month and partially shut down for another month. If the fire had not occurred sales would have been much higher, and therefore substantial revenues have been lost. In addition, fixed costs such as salaries, taxes, maintenance, and so forth must continue to be paid. A business interruption policy would respond to these losses.

Other forms of indirect insurance include the following: (1) contingent business interruption insurance, designed to cover the consequential losses if the plant of a supplier or a major customer is burned and results in either reduced orders or reduced deliveries that force a shutdown of the insured firm; (2) extra expense insurance, which pays the additional cost occasioned by having extra expenses to pay, such as rent on substitute facilities after a fire; (3) rent and rental value insurance, covering losses in rents that the owner of an apartment house may incur if the building is destroyed.

MARINE INSURANCE

Marine insurance is actually transportation insurance. After insurance coverage on ocean voyages had been developed, it was a natural step to offer insurance on inland trips. This branch of insurance became known as inland marine. In many policy forms the distinction between inland and ocean marine has disappeared; it is common to cover goods from the time they leave the warehouse of the shipper, even if this warehouse is situated many miles from the nearest seaport, until they reach the warehouse of the buyer, which likewise may be located far inland.

Insurance of goods in transit

Ocean marine insurance. Ocean marine contracts are written to cover four major types of property interest: (1) the vessel or hull; (2) the cargo; (3) the freight revenue to be received by the ship owner; and (4) legal liability for negligence of the shipper or the carrier. Hull insurance covers losses to the vessel itself from specified perils. Usually there is a provision that the marine hull should be covered only within specified geographic limits. Cargo insurance is usually written on an open contract basis under which shipments, both incoming and outgoing, are automatically covered for the interests of the shipper, who reports periodically the values exposed and pays a premium based upon these values. By means of a negotiable open cargo certificate, which is attached to the bill of lading, insurance coverage is automatically transferred to whoever has legal title to the goods in the course of their movement from seller to buyer.

Freight revenue may be insured in several different ways. If there is an obligation by the shipper to pay the carrier's freight bill regardless of whether or not the goods are delivered, the value of the freight is declared a part of the value of the cargo and is insured as part of this value. If the freight revenue is contingent upon safe delivery of the goods, the carrier insures the freight as a part of the regular hull coverage.

Major clauses or provisions that are fairly standardized are (1) the perils clause; (2) the "running down" clause, or R.D.C.; (3) the "free of particular average," or F.P.A., clause; (4) the general average clause; (5) the sue and labour clause; (6) the abandonment clause; (7) coinsurance; and (8) express and implied warranties. Each of these will be discussed in turn.

Perils clause. Until 1978 the main insuring clause of modern ocean marine policies was preserved almost unchanged from the original 1779 Lloyd's of London form. The clause is as follows:

> Touching the adventures and perils which we the assurers are contented to bear and do take upon us in this voyage: they are of the seas, men-of-war, fire, enemies, pirates, rovers, thieves, jettisons, letters of mart and countermart, surprisals, takings at sea, arrests, restraints, and detainments of all kings, princes, and people, of what nation, condition, or quality soever, barratry of the master and mariners, and of all other perils, losses, and misfortunes, that have or shall come to the hurt, detriment, or damage of the said goods and merchandises, and ship, etc., or any part thereof.

Although the clause reads as if it were an all-risk agreement, courts have interpreted it to cover only the perils mentioned. Essentially, the clause insures the voyage from

perils *of* the sea. Perils *on* the sea, such as fire, are not covered unless specifically mentioned. Furthermore, although the perils clause indicates coverage from "enemies, pirates, rovers, thieves," the policy does not cover losses from war. (War risk insurance is offered in some nations through governmental agencies.)

In 1978, at the request of the UN Conference on Trade and Development, the 1779 language was modernized and a revised insuring clause was proposed. The new form restricts coverage on losses from poor packing, places the burden of proof of seaworthiness on the shipper rather than on the carrier, and excludes losses resulting from insolvency of the common carrier, with the burden of proof placed on the shipper that the carrier is financially sound. The revised form has not been adopted by all insurers.

R.D.C. clause. The R.D.C., or "running down" clause, provides coverage for legal liability of either the shipper or the common carrier for claims arising out of collisions. (Collision loss to the vessel itself is part of the hull coverage.) The R.D.C. clause covers negligence of the carrier or shipper that results in damage to the property of others. A companion clause, the protection and indemnity clause (P and I), covers the carrier or shipper for negligence that causes bodily injury to others.

F.P.A. clause. The F.P.A., or "free of particular average," clause excludes from coverage partial losses to the cargo or to the hull except those resulting from stranding, sinking, burning, or collision. Under its provisions, losses below a given percentage of value, say 10 percent, are excluded. In this way the insurer does not pay for relatively small losses to cargo. The percentage deductible varies according to the type of cargo and its susceptibility to loss.

General average clause. The general average clause in ocean marine insurance obligates the insurers of various interests to share the cost of losses incurred voluntarily to save the voyage from complete destruction. Such sacrifices must be made voluntarily, must be necessary, and must be successful. For example, if a shipper's cargo is voluntarily jettisoned in a storm in order to save the vessel from total loss, the general average clause requires the insurers of the hull and of all other cargo interests to make a contribution to the loss of the shipper whose goods were sacrificed. Other types of losses may also be covered. It has been held, for example, that losses suffered from efforts to put out a fire on shipboard, which result in damage to specific goods, can be included in a general average claim. Similarly, losses from salvage efforts to free a stranded vessel may qualify under a general average claim to which all interests must contribute.

'oluntary
)sses

Sue and labour clause. The sue and labour clause requires the ship owner to make every attempt to reduce or save the exposed interests from loss. Under the terms of the clause, the insurer pays for any necessary costs incurred in carrying out the requirements of the sue and labour clause. Thus, if a ship is stranded, under the sue and labour clause the hull owner would be required to hire salvors to attempt to save the ship. Such expenses are paid even if the salvage attempts fail.

Abandonment clause. If salvaging or rehabilitating a ship or cargo following a marine loss costs more than the goods are worth, the loss is said to be constructively total. Under such conditions, the ocean marine policy permits the insured to abandon the damaged ship or cargo to the insurer and make a claim for the entire value. In this case, the salvage belongs to the insurer, who may dispose of it in any way. Abandonment is not permitted in other forms of property insurance.

Coinsurance. Although there is no coinsurance clause as such in the ocean marine policy, losses are settled as though a 100 percent coinsurance clause existed. Thus, if an insured takes out coverage equal to 50 percent of the true replacement cost of the goods, only 50 percent of any partial loss may be recovered.

Warranties. In the field of ocean marine insurance there are two general types of warranties that must be considered: express and implied. Express warranties are promises written into the contract. There are also three implied warranties, which do not appear in written form but bind the parties nevertheless.

Examples of expressed warranties are the F.C.&S. warranty and the strike, riot, and civil commotion warranty. The F.C.&S., or "free of capture and seizure," warranty excludes war as a cause of loss. The strike, riot, and civil commotion warranty states that the insurer will pay no losses resulting from strikes, walkouts, riots, or other labour disturbances. The three implied warranties relate to the following conditions: seaworthiness, deviation, and legality. Under the first, the shipper and the common carrier warrant that the ship will be seaworthy when it leaves port, in the sense that the hull will be sound, the captain and crew will be qualified, and supplies and other necessary equipment for the voyage will be on hand. Any losses stemming from lack of seaworthiness will be excluded from coverage. Under the deviation warranty, the ship may not deviate from its intended course except to save lives. Clauses may be attached to the ocean marine policy to eliminate the implied warranties of seaworthiness or deviation. The implied warranty of legality, however, may not be waived. Under this warranty, if the voyage itself is illegal under the laws of the country under whose flag the ship sails, the insurance is void.

Implied warranties

Inland marine insurance. Although there are no standard forms in inland marine insurance, most contracts follow a typical pattern. They are usually written on a named-peril basis covering such perils of transportation as collision, derailment, rising water, tornado, fire, lightning, and windstorm. The policies generally exclude losses resulting from pilferage, strike, riot, civil commotion, war, delay of shipments, loss of markets, illegal trade, or leakage and breakage.

Rates are generally negotiated on an individual basis and vary according to the volume of shipments during the year. The final premium is often determined by audit when the volume of the year's shipments are known. The inland transit form also contains a clause covering loss caused by the negligence of third parties. Usually the insurer maintains its right of subrogation against the railroad or other common carrier for all loss for which the carrier is legally liable.

The scope of inland marine is greatly extended by means of "floater" policies. These are used to insure certain types of movable property whether or not the property is actually in transit. Business floater policies are purchased by jewelers, launderers, dry cleaners, tailors, upholsterers, and other persons who hold the property of others while performing services. Personal property floaters are used to cover, on a comprehensive basis, any item of personal property owned by a private individual. They may also cover the property of visitors, or the property of servants while on the premises of the insured. They exclude certain types of property for which other contracts have been designed, such as automobiles, aircraft, motorcycles, animals, or business and professional equipment.

Floater policies

PROPERTY AND LIABILITY INSURANCE

Property and liability insurance, also known as casualty insurance, covers loss associated with hazards to persons and property, including legal hazards as well as those arising from accident and sickness. It does not include life insurance. Major classes are liability (including automobile), theft, aviation, workers' compensation, credit and title, and a variety of miscellaneous types covering machinery and equipment, plate glass windows, and so forth.

Liability insurance. Liability insurance arises mainly from the operation of the law of negligence. Individuals who, in the eyes of the law, fail to act reasonably or to exercise due care may find themselves subject to large liability claims. Court judgments have been issued for sums so large as to require a lifetime to pay.

There are at least four major types of liability insurance contracts: (1) liability arising out of the use of automobiles, (2) liability arising out of conduct of a business, (3) liability arising from professional negligence (applicable to doctors, lawyers, etc.), and (4) personal liability, including the liability of a private individual operating a home, carrying on sporting activities, and so on.

Practically all liability contracts falling in these four categories have some common elements. One is the insuring

Elements of liability contracts

clause, in which the insurer agrees to pay on behalf of the insured all sums that the insured shall become legally obligated to pay as damages because of bodily injury, sickness or disease, or wrongful death or injury to another person's property. The liability policy covers only claims that an insured becomes legally obligated to pay; voluntary payments are not covered. It is often necessary to resort to legal or court action to determine the amount of these damages, although in a vast majority of cases the damages are settled out of court by negotiation between the parties.

All liability insurance contracts contain clauses that obligate the insurer to conduct a court defense and to pay any settlement, including premiums on bonds, interest on judgments pending appeal, medical and surgical expenses that are necessary at the time of the accident, and other costs. Liability insurance has sometimes been termed defense insurance because of this provision. The insurer agrees to defend a suit even though it is false or fraudulent, so long as it is a suit stemming from a peril insured against. The insured is required to cooperate with the insurer in all court actions by appearing in court, if necessary, to give testimony.

Practically all liability insurance policies contain limitations on the maximum amount of a judgment payable under the contract. The cost of defense and supplementary payments are paid in addition to any judgments.

Definition of "the insured"

Liability insurance contracts have in common the fact that the definition of "the insured" is broad. Automobile liability policy, for example, includes not only the owner and driver but anyone else operating the car with permission. In business liability insurance, all partners, officers, directors, or proprietors are covered by the policy regardless of their direct responsibility for any act of negligence. Other parties may be included for an extra premium.

Another element common to all liability insurance policies is certain exclusions. Policies covering business activities almost invariably exclude liability arising out of the personal activities of the insured. Each kind of liability contract tends to exclude the liability for which another contract has been devised: a personal liability policy, for example, excludes automobile liability since a special contract has been created for this particular type of liability. Another common exclusion is damage to property under the care, custody, or control of the insured, on the theory that one cannot be liable to oneself for one's own negligence; thus if damage to property results from the owner's negligence, the owner may not look to a liability policy for recovery but is expected to have some type of physical damage insurance.

Another common element in liability policies is subrogation: the insurer retains the right to bring an action against a liable third party for any loss this third party has caused.

Automobile insurance. Automobile insurance in the United States and the United Kingdom is commonly written as a comprehensive contract that includes not only liability but also medical payments, theft, and physical damage. Usually the policy defines the insured broadly enough to include anyone who is driving the car with permission of the insured or the insured's spouse. Practices in other countries vary. Compulsory insurance laws obtain in most European countries, and it is not unusual for the language of the contracts to be specified by law.

Business liability insurance. Business liability contracts commonly written include the following: liability of a building owner, landlord, or tenant; liability of an employer for acts of negligence involving employees; liability of contractors or manufacturers; liability to members of the public resulting from faulty products or services; liability as a result of contractual agreements under which liability of others is assumed; and comprehensive liability. The latter contract is designed to be broad enough to encompass almost any type of business liability, including automobiles. There has been increasing use of coverage for liability stemming from defective products, some court judgments for which have been enormous.

Most business liability contracts are written on what has been known as the occurrence basis, under which the business is covered for loss even if the act that produced the claim was not accidental. The only requirement is that the result of the act be accidental or unintended. Thus if a contractor is making an excavation that produces large amounts of dust and this dust causes loss to neighbouring property, the contractor's liability policy would respond to claims for loss, if it were written on the occurrence basis, even though the act that produced the dust was a deliberate act.

Malpractice insurance

Professional liability. Known as malpractice insurance, professional liability contracts are distinguished from general business liability policies because of the specialized nature of the liability. Professional persons requiring liability contracts include physicians and surgeons, lawyers, accountants, engineers, architects, and insurance agents. Important differences between professional and other liability contracts are the following:

1. No distinction is made between bodily injury or property damage liability, and there is no limit per accident but rather a limit of liability per claim. This recognizes the fact that one negligent act on the part of a professional person may involve more than one party, each of whom could bring a legal action against the professional person. Thus a doctor might administer the wrong medicine to a number of patients, each of whom could bring a legal action.
2. Claims against a professional person may have an adverse effect upon his or her reputation. The policy therefore permits the insured to carry any action to court, since an out-of-court settlement might conceivably imply guilt in the eyes of the professional's public or clientele.
3. In professional liability insurance there is an exclusion for any agreement guaranteeing the result of any treatment. Suits stemming from clients' dissatisfaction with the service performed are thus not covered.

Personal liability insurance. The most common form of personal liability insurance is issued as a comprehensive personal liability insurance policy. It is an all-risk agreement and contains relatively few exclusions. The policy covers any act of negligence of the insured or residents of the home that results in legal liability. It may also include medical payments insurance covering accidental injury to guests and other nonresidents without regard to the question of negligence.

Theft insurance. Theft generally covers all acts of stealing. There are three major types of insurance contracts for burglary, robbery, and other theft. Burglary is defined somewhat narrowly to mean the unlawful taking of property within a premises that has been closed and in which there are visible marks evidencing forcible entry. This narrow definition is necessary to restrict burglary coverage to a particular class of criminal act. Robbery is defined as that type of unlawful taking of property in which another person is threatened by either force or violence. In the robbery peril, therefore, the element of personal contact is necessary. Theft policies further categorize robbery according to whether it takes place inside or outside designated premises; different premium rates and loss limits apply depending on where the robbery occurs. The insured peril is generally restricted to acts of persons not in a position of trust—"outsiders." Employees, officers, directors, and other "insiders" must be covered under fidelity bonds.

Underwriting is difficult, and rates tend to be high because those businesses most subject to theft are most likely to apply for coverage—liquor stores, pawnshops, jewelry stores, and so forth. Another reason for high rates is that owners tend to underinsure on the assumption that a burglar or robber can steal only so much at any one time.

Major types of theft coverage

Perhaps the most common of all burglary coverages is on safes. Often the loss in the form of damage to the safe itself from the use of explosives and other devices is as great as the loss of the money, jewelry, or securities it contains. Accordingly, the policy covers both types of claims. Another common burglary policy applies to mercantile open stock. In this policy there is usually a limit applicable on any article of jewelry or any article contained in a showcase where susceptibility to loss is high. In order to prevent underinsurance the mercantile open stock policy is usually written with a coinsurance requirement or with some minimum amount of coverage.

Another common theft policy for business firms is a comprehensive crime contract covering employee dishon-

esty as well as losses on money and securities both inside and outside the premises, loss from counterfeit money or money orders, and loss from forgery. This policy is designed to cover in one package most of the crime perils to which an average business is subject.

A broad form of crime protection for individuals is the personal theft policy. This is offered both as a separate contract and as part of a "homeowner's policy." It covers all losses of personal property from theft and mysterious disappearance, including money and securities.

Aviation insurance. Aviation insurance normally covers physical damage to the aircraft and legal liability arising out of its ownership and operation. Specific policies are also available to cover the legal liability of airport owners arising out of the operation of hangars or from the sale of various aviation products. These latter policies are similar to other types of liability contracts.

Perhaps the major underwriting problem is the "catastrophic" exposure to loss. The largest passenger aircraft may incur losses of $100,000,000 or more, counting both liability and physical damage exposures. The number of aircraft of any particular type is not large enough for the accurate prediction of losses, and each type of aircraft has its special characteristics and equipment. Thus a great deal of independent individual underwriting is necessary. Rate making is complex and specialized. It is further complicated by rapid technological change and by the constant appearance of new hazards.

Hull insurance for aircraft is generally written on an all-risk basis. Rates are usually expressed as a percentage, such as 2 percent, of the aircraft's replacement value. Policies generally exclude losses from damage that is incurred while the aircraft is being used for unlawful purposes, or outside the geographical limits specified in the policy, or by someone other than an authorized pilot, or for purposes other than those that are authorized in the policy, or in violation of an airworthiness condition or of civil air regulations. Also excluded are losses resulting from war, strike, riot, civil commotion, wear and tear, depreciation, and loss of revenue from failure to use the aircraft.

Policies are written to cover liability of the owner or operator for bodily injury to passengers or to persons other than passengers, and for property damage. Medical costs, including loss of income, are usually paid to passengers suffering permanent total disability without the requirement of proving negligence. This type of coverage has been called admitted liability insurance.

ndustrial njury nsurance

Workers' compensation insurance. Workers' compensation insurance, sometimes called industrial injury insurance, compensates workers for losses suffered as a result of work-related injuries. Payments are made regardless of negligence. The schedule of benefits making up the compensation is determined by statute.

The scope of employment injury laws, originally limited to persons in forms of employment recognized as hazardous, has, as the result of associating the right to compensation with the existence of a contract of service, been gradually extended to clerical employment, which is safer than being at home or on the street. Nevertheless, the large exception of agricultural employees remains in some underdeveloped countries (*e.g.,* former British colonies), Canada, much of the United States, and the U.S.S.R. (collective farmers). Other classes of exception are employees in very small undertakings and domestic servants. The exclusion of employees with middle-class salaries persists in parts of the former British Empire. In a few countries, working employers are enabled to insure themselves as well as their employees.

The notion of employment injury was at first confined to injuries of accidental origin, but during the 20th century it was extended to include occupational diseases in increasing number. To entitle the worker to benefit, the accident must occur during employment, and many laws also require the accident to have been caused by the employment in some way; but the trend seems to be toward accepting the former condition as sufficient. Following the German law of 1925, some 30 countries included accidents occurring on the way to and from work. Injuries due to the employee's willful misconduct are generally excluded. Occupational diseases are covered to some extent by virtually all national laws.

Classes of benefits. Four classes of benefits are provided by compulsory insurance, and, except for certain diseases, a right to them is acquired without any qualifying period of previous employment. First is a medical benefit, which includes all necessary treatment and the supply of artificial limbs. If its duration is limited, the maximum is likely to be one year. Second is a temporary incapacity benefit, which lasts as long as the medical benefit except that a waiting period of a few days is frequently prescribed. Its rate varies in different countries, from 50 percent of the employee's wage (*e.g.,* in former British colonies and protectorates and in a few Latin-American countries) to 100 percent (in the U.S.S.R., Yugoslavia, Argentina, Brazil, Colombia, and Mexico), with a maximum, and often a minimum, limit ($66^2/_3$ percent and 75 percent are the commonest rates). Third is a permanent incapacity benefit, which, unless the degree is very small, in which case a lump sum is paid, takes the form of a pension. If the incapacity is total, the pension is usually equal to the temporary incapacity benefit. If the incapacity is partial, the pension is proportionately smaller. In some 60 countries an additional pension is granted if the victim needs constant attendance. In cases of death the pensions are distributed to the widow (or invalid widower) and minor children, and, if the maximum total has not then been attained, other dependents may receive small pensions. The maximum is the same as for total incapacity.

Permanent incapacity benefit

In a growing number of industrialized countries (Austria, France, West Germany, Ireland, Israel, The Netherlands, and Switzerland) the fourth type of benefit—systematic arrangements for retraining seriously injured persons—is provided, and employers may even be required to provide employment to such persons.

Financing and administering employment injury insurance. Almost all systems of employment injury insurance are financed by employers' contributions exclusively, and in almost all of these systems the contribution is proportional to the risk represented by the class of activity in which the employer is engaged. Usually the insurance institution adapts the contribution to the accident experience of the undertaking individually or to any special preventive measures it may have taken. On the other hand, mainly for simplicity, but partly perhaps in order to subsidize basic but dangerous industries, a uniform contribution rate for all classes of activity has been established in Austria, several people's democracies, and many underdeveloped countries.

Social insurance against employment injury, as against other risks, is in most countries administered by institutions under the joint management of employers and employees and often of government representatives as well; in eastern Europe, however, the administration is entrusted to trade unions. Disputes are settled by arbitral organs without resort to the courts.

In the United States, an employer may comply with the provisions of most workers' compensation laws in three ways: by purchasing a private workers' compensation and employer liability policy from a commercial insurer; by purchasing coverage through a state fund set up for this purpose; or by setting aside reserves sufficient to cover the risks involved. Most workers' compensation benefits are financed by the first two methods.

State laws in the United States are not uniform with respect to the size or length of time for which income payments are made. For example, only about half of the states give lifetime income benefits for occupational injuries. In others there is a statutory limitation of between 400 and 500 weeks of payments. Again, most states provide liquidating damages for an injury that is permanent but does not totally incapacitate the worker, such as the loss of an arm or leg. The size of these liquidating damages varies greatly. Most state laws also provide complete medical benefits, including rehabilitation expenses, and survivors' benefits in the event of the worker's death.

Coverage of credit risks

Credit insurance. The use of credit in modern societies is so various and widespread that many types of insurance have grown up to cover some of the risks involved.

Examples of these risks are: the risk of bad debts from insolvency, death, and disability; the risk of loss of savings from bank failure; the risk attaching to home-loan debts when installments are not paid for various reasons, resulting in foreclosure with subsequent loss to the creditor; and the risk of loss from export credit because of war, currency restrictions, cancellation of import licenses, or other political causes.

Merchandise credit insurance. Credit insurance for domestic buyers and sellers is available in the United States, Canada, Mexico, and most European countries. It is sold only to manufacturers, wholesalers, and certain service agencies, not to retailers. The insurance is designed to enable the seller to recover a certain percentage of losses from insolvency of the debtor, but the contracts list a number of conditions under which the creditor may initiate a claim regardless of the question of insolvency. The policy is designed primarily to meet the needs of those sellers whose business is concentrated among a few buyers, insolvency of any one of which would seriously jeopardize the financial stability of the seller.

Export credit insurance. A special form of credit insurance is available to exporters against losses from both commercial and political risks. In the United States, for example, export credit insurance is written through a consortium of insurance companies organized by the Foreign Credit Insurance Association (FCIA). The Export-Import Bank of the United States assumes the ultimate liability for loss, while the FCIA serves as the underwriting agency. Coverage is usually limited to 90 or 95 percent of the account. Prior approval from the FCIA is usually required before export credit insurance is granted. In some cases, the exporter is required to purchase coverage on all credit sales in a given country as a device to reduce adverse selection.

Export credit insurance is used more widely in some countries than in others. In the United Kingdom, approximately a quarter of all export sales are covered, compared with about 6 percent in the United States. Export sales are not eligible for insurance if they are made for cash or financed directly or indirectly through government-guaranteed loans.

Title insurance. Title insurance is a contract guaranteeing the purchaser of real estate against loss from undiscovered defects in the title to property that has been purchased. Such loss may stem from unmarketability of the property because of defective title or from costs incurred to cure defects of the title.

The need for title insurance arises from the fact that real estate transactions are complex and technical. Any legal error, no matter how detailed or minute, may cause a defect in the title that impairs its marketability. Examples of such defects are forgeries, invalid or undiscovered wills, defective probate proceedings, or transfers of property by persons lacking full legal capacity to contract.

Public records contain a description of each piece of real property, and presumably they mention any factor that may cloud the title, such as a legal encumbrance, lien, easement, or other matters. When the property is transferred, a search of these records is made, and any defects are recorded on the title. The title insurance policy excludes losses stemming from such recorded defects, since it covers losses only from undiscovered defects. In contrast to most insurance policies, which cover losses occurring during the policy term, title insurance covers only losses caused by title defects that have already occurred but are not known when the policy is issued.

Without title insurance the potential buyer of real estate is protected only by a so-called abstract. The abstract is really a legal history of the title. An attorney may examine an abstract and render an opinion as to the validity of the title, but there is no guarantee that the attorney's opinion is correct. Title insurance, which offers such a guarantee, is normally issued by title companies that maintain elaborate records to keep current information on all the property under their surveyance. The premium for title insurance is intended mainly to cover the expenses of maintaining the records. When the property is transferred again, a new policy must be issued and a new premium paid.

Miscellaneous insurance. Special casualty forms are issued to cover the hazards of sudden explosions from equipment such as steam boilers, compressors, electric motors, flywheels, air tanks, furnaces, and engines. Boiler and machinery insurance has several distinctive features. A substantial portion of the premium collected is used for inspection services rather than loss protection. Second, the boiler policy provides that its coverage will be in excess of any other applicable insurance. In this sense, it may be looked upon as an "umbrella policy" to fill in gaps in the insured's program. Third, the policy lists the specific losses that will be paid, such as the loss of the boiler or machinery itself due to accident, expediting expenses, property damage liability, bodily injury liability, defense settlement and supplementary payments, business interruption, outage (interruption of service), power interruption, consequential loss due to spoilage of goods, and furnace explosion. The policy will satisfy each of these claims in the order in which they appear, up to the limit of the coverage.

Boiler and machinery insurance

The extensive use of plate glass in modern architecture has produced a special comprehensive insurance that covers not only plate glass but glass signs, motion-picture screens, halftone screens and lenses, glass bricks, glass doors, and so forth. It may be written to cover loss from any source except fire or nuclear radiation.

Plate glass insurance

SURETYSHIP

Surety contracts are designed to protect businesses against the possible dishonesty of their employees. Surety and fidelity bonds fill the gap left by theft insurance, which always excludes losses from persons in a position of trust. A bond involves three contracting parties instead of two. The three parties are: the principal, who is the person bonded; the obligee, the person who is protected; and the surety, the person or corporation agreeing to reimburse the obligee for any losses stemming from failures or dishonesty of the principal. The bond covers events within the control of the person bonded, whereas insurance in the strict sense covers loss from random events generally outside the direct control of the insured. In bonding, the surety always has the right to try to collect its losses from the person bonded, whereas in insurance the insurer may not attempt to recover losses from the insured. Of course, under property and liability policies the insurer may attempt to recover from liable third parties under the right of subrogation, but subrogation rights are often not possible to enforce in practice. Bonds are not usually cancellable by the insurer, whereas most insurance contracts, except life, are cancellable by the insurer upon due notice.

Fidelity bonds and surety bonds

Fidelity bonds are written to cover the obligee, usually an employer, against loss from dishonest acts of employees; surety bonds cover not only dishonesty but also incapacity to perform the work agreed upon. Surety bonds are normally written on principals who are acting in an independent or semi-independent capacity, such as building contractors or public officials, whereas fidelity bonds are written on employees acting under the guidance and supervision of their employer. Finally, surety bonds are often issued with the requirement of collateral, whereas fidelity bonds are not. The surety bond is an instrument through which the superior credit of the surety is substituted for the uncertain credit of the principal; hence if the surety is asked to bond a principal of somewhat doubtful credit, the requirement of cash collateral is frequently imposed.

Major types of fidelity bonds. Fidelity bonds differ according to whether specific persons are named as principals or whether all employees or persons are covered as a group. The latter are most frequently used by large employers because they offer automatic coverage on given classes of workers, including new employees, and greater ease of administration, including simpler claims procedures. Fidelity bonds are usually written on a continuous basis—that is, they are effective until cancelled and have no expiration date. The penalty of the bond (the maximum amount payable for any one loss) is unchanged from year to year and is not cumulative. The bonds specify a discovery period (usually two years) limiting the time for discovering losses after a bond is discontinued. When a

new bond is put into effect, it can be written to cover losses that have occurred but are undiscovered before the effective issue date of the bond. A salvage clause also is included, stating the way in which any salvage recovered by the surety from the principal is to be divided between the surety and the obligee. This clause is significant, because the obligee may have losses in excess of the penalty of the bond. Some salvage clauses require that any salvage be paid to the obligee up to the full amount of all losses, and others provide that any salvage be divided between the surety and the obligee on a pro-rata basis, in the proportion that each party has suffered loss.

Major types of surety bonds. There are various classes of surety bonds. Contract construction bonds are written to guarantee the performance of contractors on building projects. Bonds are particularly important in this field because of the general practice of awarding commercial building contracts to the lowest bidder, who may promise more than can actually be performed. The surety who is experienced in this field is in a position to make sounder judgment about the liability of the various bidders than anyone else and backs up its judgment with a financial guarantee.

Court bonds include several different types of surety bonds. Fiduciary bonds are required for court-appointed officials entrusted with managing the property of others; executors of estates and receivers in bankruptcy are frequently required to post fiduciary bonds. Litigation bonds are court bonds required of plaintiffs and defendants in court actions; examples are bail bonds, guaranteeing that a defendant will appear in court, and appeal bonds, guaranteeing that a defendant can respond to the extra costs incurred in appealing an adverse decision.

Other types of surety bonds include official bonds, lost instrument bonds, and license and permit bonds. Public official bonds guarantee that public officials will faithfully and honestly discharge their obligations to the state or to other public agencies. Lost instrument bonds guarantee that if a lost stock certificate, money order, warehouse receipt, or other financial instrument falls into unauthorized hands and causes a loss to the issuer of a substitute instrument, this loss will be reimbursed. License and permit bonds are issued on persons such as owners of small businesses to guarantee reimbursement for violations of the licenses or permits under which they operate.

LIFE AND HEALTH INSURANCE

Life insurance. Life insurance may be defined as a plan under which large groups of individuals can equalize the burden of loss from death by distributing funds to the beneficiaries of those who die. From the individual standpoint life insurance is a means by which an estate may be created immediately for one's heirs and dependents. It has achieved its greatest acceptance in Canada, the United States, Sweden, Australia, Ireland, New Zealand, The Netherlands, and Japan, in which countries the face value of life insurance policies in force generally exceeds the national income.

ife in-
ırance in
s various
ɔrms

In the United States, in 1982, nearly, $4,500,000,000,000 of life insurance was in force. The assets of the more than 2,000 U.S. life insurance companies totalled nearly $600,000,000,000, making life insurance one of the largest savings institutions in the United States. Much the same is true of other wealthy countries, in which life insurance has become a major channel of saving and investment, with important consequences for the national economy.

Life insurance is relatively little used in poor countries, although its acceptance has been increasing.

Types of contracts. The major types of life insurance contracts are term, whole life, and endowment, but innumerable combinations of these basic types are sold. Term insurance contracts, issued for specified periods of years, are the simplest. Protection under these contracts expires at the end of the stated period, with no cash value remaining. Whole life contracts, on the other hand, run for the whole of the insured's life and gradually accumulate a cash value. The cash value, which is less than the face value of the policy, is paid to the policyholder when the contract matures or is surrendered. Endowment contracts run for a stated period of years and pay their full face value at the end of the period.

Life insurance may also be classified, according to type of customer, as ordinary, group, industrial, and credit. The ordinary insurance market includes customers of whole life, term, and endowment contracts and is made up primarily of individual purchasers of annual-premium insurance. The group insurance market consists mainly of employers who arrange group contracts to cover their employees. The industrial insurance market consists of individual contracts sold in small amounts with premiums collected weekly or monthly at the policyholder's home. Credit life insurance is sold to individuals, usually as part of an installment purchase contract; under these contracts, if the insured dies before the installment payments are completed, the seller is protected for the balance of the unpaid debt.

Insurance may be issued with a premium that remains the same throughout the premium-paying period, or it may be issued with a premium that increases periodically according to the age of the insured. Practically all ordinary life insurance policies are issued on a level-premium basis, which makes it necessary to charge more than the true cost of the insurance in the earlier years of the contract in order to make up for much higher costs in the later years; the so-called overcharges in the earlier years are not really overcharges but are a necessary part of the total insurance plan, reflecting the fact that mortality rates increase with age. The insured is not overpaying for protection because of the claim on the cash values that accumulate in the early years; the policyholder may borrow on this value or may recapture it completely by lapsing the policy. The insured does not, however, have a claim on all of the earnings that accrue to the insurance company from investing the funds of its policyholders.

By combining term, whole life, and endowment insurance, an insurer can provide many different kinds of policies. Two examples of such "package" contracts are the family income policy and the mortgage protection policy. In each of these, a base policy, usually whole life insurance, is combined with term insurance calculated so that the amount of protection declines as the policy runs its course. In the case of the mortgage protection contract, for example, the amount of the decreasing term insurance is designed roughly to approximate the amount of the mortgage on a property. As the mortgage is paid off, the amount of insurance declines correspondingly. At the end of the mortgage period the decreasing term insurance expires, leaving the base policy still in force. Similarly, in a family income policy, the decreasing term insurance is arranged to provide a given income to the beneficiary over a period of years roughly corresponding to the period during which the children are young and dependent.

Package contracts

Some whole life policies permit the insured to limit the period during which premiums are to be paid. Common examples of these are 20-year life, 30-year life, and life paid up at age 65. On these contracts the insured pays a higher premium to compensate for the limited premium-paying period. At the end of the stated period the policy is said to be "paid up," but it remains effective until death or surrender.

Term insurance is most appropriate when the need for protection runs only for a limited period; whole life insurance is most appropriate when the protection need is permanent. Endowment insurance is primarily designed to ensure that a savings plan will be completed—for example, to build a retirement fund or to provide an educational fund for children.

A more recent form of package contract, known generally as "universal life," is a combination of term insurance and a flexible savings plan in which savings are invested at money-market interest rates (those rates paid on short-term notes or obligations of leading corporations). The policyholder may vary the premiums and may adjust the death protection of the plan within stated limits. The interest paid on these policies varies with market conditions. Cash values reflect both interest earnings and charges for expenses and mortality.

Settlement options. The death proceeds or cash values

of insurance may be settled in various ways. The insured may take the cash value and lapse the policy. A beneficiary may take a lump sum settlement of the face amount upon the death of the insured. The beneficiary may, instead, elect to receive the proceeds over a given number of years or in some fixed amount, such as $100 a month, for as long as the proceeds last. The money may be left with the insurer temporarily to draw interest. Or the proceeds may be used to purchase a life annuity, which in effect is another insurance policy guaranteeing regular payments for the life of the insured.

Other provisions. Life insurance policies contain various clauses that protect the rights of beneficiaries and the insured. Perhaps the best known is the incontestable clause, which provides that if a policy has been in force for two years the insurer may not afterward refuse to pay the proceeds or cancel the contract for any reason except nonpayment of premiums. Thus, if the insured made a material misrepresentation when the policy was originally obtained, and this misrepresentation is not discovered until after the incontestable period, beneficiaries may still receive the value of the policy so long as the premiums are maintained. Another protective clause is the suicide clause, which states that after a given period, usually two years, the insurer may not deny liability for subsequent suicide of the insured. If suicide occurs within the period, the insurer tenders to the beneficiary only the premiums that have been paid. If the age of the insured was misstated when the policy was taken out, the misstatement-of-age clause provides that the amount payable is the amount of insurance that would have been purchased for the premium had the correct age been stated. Many life insurance policies, known as participating policies, return dividends to the insured. The dividends, which may amount to 20 percent of the premiums, may be accumulated in cash left with the insurer at interest, used to buy additional life insurance, used to reduce premium payments, or used to pay up the contract sooner than would otherwise have been possible.

Special riders. The insured may, at a nominal charge, attach to the contract a waiver-of-premium rider under which premium payments will be waived in the event of total and permanent disability before the age of 60. Under the disability income rider, should the insured become totally and permanently disabled, a monthly income will be paid. Under the double indemnity rider, if death occurs through accident the insurance payable is double the face amount.

Private health insurance. In many countries health insurance has become a governmental institution. In some, doctors and other professional staff are employed, directly or indirectly, by a government agency on a full-time or part-time salaried basis, and health facilities are owned or operated by the government. This is the practice in Australia, Brazil, Canada, Chile, Greece, Ireland, Mexico, New Zealand, Sweden, Turkey, and the Communist countries of Europe. In other countries the government pays for medical care provided by private physicians; these countries include Austria, Denmark, West Germany, The Netherlands, Norway, and Spain. In some countries private health insurance programs exist along with, or as part of, the government program. Various combinations of programs are possible, and it is difficult to summarize all of the arrangements that actually exist. The United States provides government-run medical services in veterans' hospitals and mental hospitals, and it also has a governmental health insurance program under the Social Security Act amendments of 1965; but most health insurance in the United States still consists of private programs. Much private health insurance in the United States is operated on a group basis, generally through groups of employees whose payments may be subsidized by their employer. The following is a description of the principles of private health insurance. (Government medical services are discussed in the article SOCIAL WELFARE.)

Government-supported health insurance

Types of policies. There are five major types of health insurance coverage: hospitalization, surgical, regular medical, major medical, and disability income. Health insurance contracts are not highly standardized. The policy provisions discussed below should be considered as typical, not universal or invariant.

Hospitalization insurance indemnifies for room and board in the hospital, laboratory fees, use of special facilities, nursing care, and certain medicines and supplies. The contracts contain specific limitations on coverage, such as a maximum number of days in the hospital and maximum allowances for room and board. Surgical expense insurance covers the surgeon's charge for given operations or medical procedures, usually up to a maximum for each type of operation. Regular medical insurance contracts indemnify the insured for expenses such as physicians' home or office visits, medicines, and other medical expenses. Major medical contracts are distinguished from other health insurance policies by offering coverage without many specific limitations; usually there is only a maximum per person, a deductible amount, and a percentage deductible, called coinsurance, under which the insured usually pays 20 percent of each medical bill above the deductible amount. Disability income coverage provides periodical payments when the insured is unable to work as a result of accident or illness. There is normally a waiting period before the payments begin. Definitions of disability vary considerably. A strict definition of disability requires that one be unable to perform each and every duty of one's regular occupation for a given period, say two years, and thereafter be unable to perform the duties of any occupation for which one is reasonably fitted by training or experience. More liberal definitions of disability require only the inability to perform the duties of one's usual occupation.

An important condition of health insurance is that of renewability. Some contracts are cancellable at any time upon short notice. Others are not cancellable during the year's term of coverage, but the insurer may refuse to renew coverage for a subsequent year or may renew only at higher rates or under restrictive conditions. Thus the insured may become ill with a chronic disease and discover that upon renewal the policy excludes all future coverage for this disease. Only policies that are both noncancellable and guaranteed renewable assure continuous coverage, but these are much more expensive.

Problems. Private health insurance contracts are in general quite restricted in coverage, to the point that many consider them to be inadequate for modern conditions. They also lend themselves to abuses such as overutilization of coverage, multiple policies, and insuring for more than 100 percent of the expected loss. Health insurance, by its very existence, helps to escalate rising medical care costs; for example, insured medical losses tend to run higher than noninsured losses because physicians often charge according to "ability to pay," and insurance increases this ability. Through insurance it is also easier to pass on rising hospital costs to the patient. Finally, since there is a tendency for those most likely to have losses to take out health insurance, an element of adverse selection exists. Careful underwriting to screen out those who are trying to take advantage of the insurance mechanism to pay for known bills is considered essential, but this undoubtedly denies coverage to many who need protection.

GROUP INSURANCE

Insurance written for specific groups

Groups have always been important in the insurance field, from the burial societies of the Romans and the insurance funds of the medieval guilds to the fraternal and religious insurance plans of modern times. In the 20th century private insurance companies have written increasingly large amounts of group insurance, particularly in life insurance, health insurance, and annuities. By 1968 more than two-thirds of the industrial labour force in the United States was covered by group life and health insurance plans established by employers. Much of the impetus for these employee benefit plans came from the labour unions, which pressed for such "fringe benefits" in bargaining with employers.

Group insurance is widely used throughout the world, both in the form of private plans and as social insurance plans. Social security plans with group coverage exist in 132 nations. Private group plans are generally offered

wherever private life and health insurance companies operate. Group life insurance is the most commonly offered plan; group health plans are government-operated in many nations. In many countries, group pension plans are common as a supplement to social insurance pension schemes.

Group insurance has been especially popular in Japan, where many employees serve a company for life. All Japanese life insurance companies offer group life insurance. Health insurance is provided by the government. Funded group pensions became popular after a 1962 tax law made contributions tax deductible for Japanese employers. In addition, virtually all Japanese employers provide lump-sum retirement allowances to their workers.

Group life insurance. Under group life insurance an employer signs a master contract with the insurance company outlining the provisions of the plan. Each employee receives a certificate that gives evidence of participation in the plan. The amount of insurance depends upon the employee's salary or job classification; usually the employer pays a portion of the premium and the employee pays the rest, but sometimes the employer pays the entire cost of the plan.

A major advantage of group life insurance to an employee is that usually coverage may be obtained regardless of health. An employee who leaves the group may, without a medical examination, convert the group coverage to an individual policy. The premiums on group life insurance are considerably less than on comparable individual policies, mainly because the selling and administrative costs are minimal.

Group health insurance. Major types of health insurance written on a group basis include insurance against the losses occasioned by hospitalization, surgical expense, and disability. Hospitalization insurance is designed to cover daily room and board and other expenses. Surgical expense insurance usually provides specified allowances for physicians' charges for various operations. Regular medical expense coverage is aimed generally at covering part of the costs of medicines and doctor calls. Major medical insurance offers the insured a large monetary coverage, designed to meet catastrophic costs of illness or accident with few restrictions as to the type of medical expense for which reimbursement is allowed. The insured must bear a percentage of any loss, usually 20 percent. Temporary disability income offers the insured a weekly indemnity for a period of up to six months if the insured is temporarily disabled and unable to work. Long-term disability extends the income for periods longer than six months. Accidental death and dismemberment insurance offers an insured or a beneficiary a lump sum; it is used widely as a form of travel accident insurance.

pecial
rotection
n group
nsurance

Under the typical group health insurance contract, the insured person enjoys several elements of protection not obtainable in individual contracts. Cancellation of coverage is not permitted unless coverage for the entire group is cancelled. The insured enjoys protection against rate increases unless the rate for all members of the class is increased. Typically the group protection may be converted to some kind of individual policy, or the insured may transfer to another group plan. The insurer tends to be liberal on claims settlement because the typical premium under a group plan is large enough that the insurer is unwilling to jeopardize the good will of the clientele through miserly claims treatment.

Most group insurance plans require that certain conditions be met. Sometimes there must be a minimum number of persons covered, such as 10 or 25. The group must also have some reason for existence other than to obtain insurance. The most usual types of groups are employees of a common employer, members of a labour organization, debtors of a common creditor, or members of a professional or trade association.

Mention must also be made of nonprofit prepayment plans (such as the Blue Cross–Blue Shield plans and health maintenance organizations [HMO's] in the United States), which resemble the above plans in most respects but are not operated by insurance companies. These plans often indemnify the hospital or the physician, on the basis of services performed, rather than the patient. Health insurance plans may also be established independently by large employers, labour unions, communities, or cooperatives. Outside the United States this kind of health insurance has been taken over by government programs. In Sweden, before the enactment of the compulsory insurance program in 1955, 70 percent of the population was covered by private plans. In Great Britain, before the National Health Service was instituted in 1948, about half the population was privately covered. In The Netherlands about half the population was so covered before the government program began, and there were still many private funds run by various groups.

Group annuities. An annuity in the literal sense is a series of annual payments. More broadly it may be defined as a series of equal payments over equal intervals of time. A life annuity, a subclass of annuities in general, is one in which the payments are guaranteed for the lifetime of one or more individuals. A group annuity differs from an individual annuity in that the annuity payments are based upon the assumed length of lives of members of a given group. The size of the payments depends upon several factors: the assumed interest rate, the life expectancy of the individual or of the individuals making up the group, the length of the period during which payments are guaranteed, the length of time elapsing before the payments begin, and the number of lives on which the payments are continued. For example, if payments to an annuitant aged 65 are to be guaranteed for 20 years, they will be substantially smaller than if they are guaranteed only for the remainder of the person's life.

The typical group life annuity is written by an employer, who may pay all or part of the cost. Under the usual arrangement, every employee receives each year a credit with the life insurance company for an annuity purchased to begin at age 65. The final pension received is made up of the sum of the individual annuities purchased throughout the worker's life. As a rule an irrevocable claim to these annuity rights is gained only after the person has worked with the employer for a given number of years or has reached a given age.

The basic advantage of an annuity is that it provides an income for life that is larger than the amount that the holder would receive by putting money out at simple interest. It is the reverse of life insurance, in that the insurer pays premiums to the insured; it resembles insurance in that the payment is based on life expectancy.

The problem of inflation has led to experimentation with variable annuities in order to protect annuitants against decreases in purchasing power. The major distinguishing characteristic of a variable annuity is that the payments vary according to underlying trends in the stock market. Funds paid in for the variable annuity are invested in common stock rather than in bonds, mortgages, or other fixed-interest investments as is true of regular annuities. In simplified terms, if the stock market rises 10 percent in one year the annuitant may expect payments to go up by approximately 10 percent in the following year. Conversely, if the stock market drops 10 percent, the annuitant will suffer a 10 percent reduction in income. To the extent that the stock market reflects changes in the cost of living, the annuitant's income is automatically adjusted for these changes each year; and if the stock market also reflects increases in productivity in the economy, then the annuitant may expect to receive a share in such increases in the productivity as the economy may gain.

Variable annuities

Some variable annuity plans are tied directly to a cost-of-living index. In order to finance the increased benefits, the employer invests a portion of the funds in equities such as common stock and real estate. An assumption is made that there will be a sufficient gain from this source to enable the employer to pay the increased cost of living, but the employee is not expected to suffer reductions in annuity payments.

The problem of adjusting retirement benefits to changes in the economy has been of concern in many countries. In Finland and Israel, the governments have pegged the price of government bonds to the cost-of-living index. Retired individuals purchasing government bonds may then receive automatic increases in interest payments if the cost

of living rises. Their interest will not fall below a specified amount. Social security legislation in most parts of the world is geared in various ways to changes in the cost of living. In some cases benefits are directly tied to a price index. In other cases the legislature from time to time must be asked to make adjustments in social security benefits.

Insurance practice

UNDERWRITING AND RATE MAKING

The two basic functions in insurance are underwriting and rating, which are closely related to each other. Underwriting deals with the selection of risks, and rating deals with the pricing system applicable to the risks accepted.

The selection of risks

Underwriting principles. Underwriting has to do with the selection of subjects for insurance in such a manner that general company objectives are met. The main objective of underwriting is to see that the risk accepted by the insurer corresponds to that assumed in the rating structure. There is often a tendency toward adverse selection, which the underwriter must try to prevent. Adverse selection occurs when those most likely to suffer loss are covered in greater proportion than others. The insurer must decide upon certain standards, terms, and conditions for applicants, project estimated losses and expenses through the anticipated period of coverage, and calculate reasonably accurate rates to cover these losses and expenses. Since many factors affect losses and expenses, the underwriting task is complex and uncertain. Bad underwriting has resulted in the failure of many insurers.

In some types of insurance major underwriting decisions are made in the field, and in other types they are made at the home office. In the field of life insurance the agent's judgment is not accepted as final until the home-office underwriter can make a decision, for the life insurance contract is usually noncancellable, once written. In the field of property and liability insurance, on the other hand, the contract is cancellable if the home-office underwriter later finds the risk to be unacceptable. It is not uncommon for a property and liability insurer to accept large risks only to cancel them at a later time after the full facts are analyzed. The insurance underwriter must tread a thin line between undue strictness and undue laxity in the acceptance of risk. The underwriter's position is not unlike that of the credit manager in a business corporation, in which unreasonably strict credit standards discourage sales but overly weak credit standards invite losses.

An important initial task of the underwriter is to try to prevent adverse selection by analyzing the hazards that surround the risk. Three basic types of hazards have been identified as moral, psychological, and physical. A moral hazard exists when the applicant may either want an outright loss to occur or may have a tendency to be less than careful with property. A psychological hazard exists when an individual unconsciously behaves in such a way as to engender losses. Physical hazards are conditions surrounding property or persons that increase the danger of loss.

Moral, psychological, and physical hazards

An underwriter may suspect the existence of a moral hazard on applications submitted by persons with known records of dishonesty, or when excessive coverage is sought or the replacement value of the property exceeds its value as a profit-making enterprise. Underwriters are aware that fire losses are more likely to occur during business depressions. The underwriter can detect moral hazard in various ways: An applicant's credit may be checked; courthouse and police records may reveal a criminal history or a history of bankruptcy; and other insurance companies can be queried for information when it is suspected that an individual is trying to obtain an excessive amount of coverage or has been turned down by other insurers.

The psychological type of hazard is more interesting. Some persons are said to be "accident prone" because they have far more than their share of accidents, suggesting that unconsciously they want them. It is well known that persons applying for annuities tend to have longer than average lives, and consequently a special mortality table is used for annuitants. Certain types of insanity have to be watched for—notably the impulse to set fires.

Physical hazards include such things as wood-frame construction in buildings, particularly in areas where such properties are densely concentrated. Earthquake insurance rates tend to be high where geological faults exist (as in San Francisco, which is built almost directly over such a fault).

Each kind of insurance has its characteristic hazards. In fire insurance the physical hazards are analyzed according to four major factors: type of construction, the protection rating of the city in which the property is located, exposure to other structures that may spread a conflagration, and type of occupancy.

In underwriting automobile insurance, the underwriter considers the following factors: the age, sex, and marital status of the driver and members of the driver's household; length of driving experience; occupation; stability of employment and residence; physical impairments; accident and conviction record; extent of use of alcohol and drugs; customary use of the vehicle; age, condition, and maintenance of the vehicle; records of insurance cancellation or refusal. In some cases tests of emotional maturity are administered. Some underwriters even consider such factors as the school records of student drivers and whether or not driving courses have been taken.

The hazards considered in the underwriting of general liability insurance depend upon the type of business and the record of the person applying for coverage. In the field of contracting, for example, the underwriter is interested in the type of equipment owned or rented by the applicant; the applicant's losses in the past, attitude toward safe practice, cooperation with building inspectors, and financial position and credit standing; the stability of supervisory employees; and the degree to which the applicant has been a successful contractor in the past.

How insurance rates are established

Rate making. Closely associated with underwriting is the rate-making function. If, for example, the underwriter decides that the most important factor in discriminating between different risk characteristics is age, the rates will be differentiated according to age.

The rate is the price per unit of exposure. In fire insurance, for example, the rate may be expressed as $1 per $100 of exposed property; if an insured has $1,000 of exposed property, the premium will thus be $10. The rate reflects three major elements: the loss cost per unit of exposure, the administrative expenses, or "loading," and the profit. In property insurance, approximately one-third of the premium covers expenses and profit, and two-thirds covers the expected cost of loss payments. These percentages vary somewhat according to the particular type of insurance.

Rates are calculated in the following way. A policy, for instance, may be written covering a class of automobiles with an expected loss frequency of 10 percent and an average collision loss of $400. The expenses of the insurer are to average 35 percent of the premium, and there must be a profit of 5 percent. The pure loss cost per unit is 10 percent of $400, or $40. The gross premium is calculated by the formula $L/[1-(E+P)]$, in which L equals the loss cost per unit, E equals the expense ratio, and P equals the profit ratio. In this case the gross premium would be $40/[1 − (.35 + .05)], or $66.67.

Credibility. A rate is said to be credible when it is based on a number of exposure units large enough to satisfy statistical estimates of accuracy. Thus 10,000 automobiles may be a large enough number to produce a rate that is 99 percent credible, whereas a rate based on 500 may be only 30 percent credible. If an underwriter discovers that unit costs are averaging 25 percent higher than expected, say $50 instead of $40, and the insured group consists of only 500 autos, the rate will be adjusted upward by only 30 percent of the amount that would be required if the higher loss costs were based on an experience of 10,000 autos. The new rate, based on 30 percent credibility, is calculated according to the formula $L_1 + (L_2 - L_1)\ C/[1-(E+P)]$, in which L_1 equals loss cost based on existing loss assumptions, L_2 equals loss cost based on new information, C equals credibility of new information, E equals expense ratio, and P equals profit ratio. The new rate in the above case would be $40 + ($50 − $40) .30/[1 − (.35 + .05)], or $71.66. The new gross rate of $71.66 is 7.5 percent above the old rate of $66.67, rather than

25 percent. This increase is 30 percent of the rate increase (25 percent) that would have been made had the new experience been fully credible ($.30 \times .25 = .075$).

Credibility factors in insurance rate making are developed from statistical theory and the "law of large numbers." Although a full exposition is beyond the scope of this article, it is intuitively easy to see that as the number of exposures increases, the variation of actual losses from probable or expected losses becomes progressively smaller. This assumes that losses occur in a random manner and with frequencies that closely follow some underlying mathematical distribution. Loss predictions become more and more accurate as the number of exposures increases, and credibility factors rise correspondingly.

Standards of rate making

Four basic standards are used in rate making: (1) the structure of rates should allocate the burden of expenses and costs in a way that reflects as accurately as possible the differences in risk—in other words, rates should be fair; (2) a rate should produce a premium adequate to meet total losses but should not bring unreasonably large profits; (3) the rate should be revised often enough to reflect current costs; and (4) the rate structure should tend to encourage loss prevention among those who are insured.

Some examples will illustrate the nature and application of the criteria outlined above. In life insurance, the rate is generally more than adequate to meet all reasonably anticipated losses and expenses; in other words, the insured is charged an excessive premium, part of which is then returned as a dividend according to actual losses and expenses. The requirement that the rate reflect fairly the risk involved is much more difficult to achieve. In workers' compensation insurance, the rate is expressed as a percentage of the employer's payroll for each occupational class. This may seem fair enough, but an employer with relatively high-paid workers has fewer employees for a given amount of payroll than one whose workers are paid a lower wage. If the two employers fall into the same occupational class and have the same total payroll, they are charged the same premium even though one may have a larger number of workers than the other, and hence greater exposure to loss. Fairness may be an elusive goal.

Insurance rates are revised only slowly, and since they are based upon past experience, they tend to remain out of date. In life insurance, for example, the mortality tables used are changed only every several years, and rate adjustments are reflected in dividends. In automobile insurance, rates are revised annually or even oftener, but they still tend to be out of date.

Systems. Two basic rate-making systems are in use: the manual, or class-rating, method, and the individual, or merit-rating, method. Sometimes a combination of the two methods is used.

A manual rate is one that applies uniformly to each exposure unit falling in some predetermined class or group, such as people of the same age, workers of one employer, drivers meeting certain characteristics, or all residences in a given area. Presumably the members of each class are so homogeneous as to be indistinguishable so far as risk characteristics are concerned.

Merit rating is used to give recognition to individual characteristics. In commercial buildings, for example, fire insurance rates depend on such individual characteristics as the type of occupancy, the number and type of safety features, and the quality of housecleaning. In an attempt to reflect the true quality of the risk, a percentage charge or credit may be applied to the base rate for each of these features. Another example is found in employer group health insurance plans where the premium or the rate may be adjusted annually depending on the loss experience or on the amount of claims service provided.

In order to obtain broader and statistically sounder rates, insurers often pool loss and claims experience by setting up rating bureaus to calculate rates based on industry-wide experience. They may have an agreement that all member companies must use the rates thus developed. The rationale for such agreements is that they help insurers meet the criteria of adequacy and fairness. Rating bureaus are used extensively in fire, marine, workers' compensation, automobile, and crime insurance.

LEGAL ASPECTS OF INSURANCE

Government regulation. The insurance business is subject to extensive government regulation in all countries. In European countries insurance regulation is a mixture of central and local controls. In West Germany, for example, the federal law applies generally to some 9,000 insurers operating in the country, most of them small local companies, but only about 10 percent of these insurers are supervised directly by the Federal Insurance Department; the others are all supervised by local agencies.

Forms of government regulation

In the United Kingdom, regulation generally allows the management fairly complete liberty of action and is concerned only with final business results. In this the United Kingdom differs from most other European countries, in which the purpose of insurance supervision is to regulate more closely the conditions in which insurers operate.

In the countries of the European Economic Community (under articles 59–60 of the Treaty of Rome) an attempt is being made to obtain greater uniformity among national insurance statutes. This is intended to facilitate the operations of insurers across national borders. Rate regulation, however, remains within the jurisdiction of individual countries, being quite strict in France and Italy and relatively lax in Germany.

An important legal force influencing insurance regulation in such countries as France, Belgium, Egypt, Greece, Italy, Lebanon, Spain, Turkey, and the former French African colonies is the Code Napoleon. The influence of the Code may be seen, for example, in the matter of third-party liability, in which the burden of proof may be upon the defendant rather than upon the plaintiff.

In some countries not all classes of insurance are regulated. In The Netherlands only life insurance is regulated, and in Belgium only life, industrial injury, and third-party motor vehicle liability insurance. In some countries the scope of supervision may embrace many aspects of the insurance business, but in the United Kingdom and The Netherlands only financial matters are subject to regulation.

In several European countries insurers may not write both life insurance and general insurance (property and liability insurance). Minimum capital requirements vary, depending upon the type of business written, usually being highest for life insurance.

In most European countries policies are submitted to supervisory authorities for approval or for information. In some countries standard clauses or forms of contracts must be used; for example, in Sweden insurers must use a standard compulsory motor vehicle third-party liability policy, and in Switzerland a standard contract for war risks and life insurance is required.

Insurance is often compulsory. In general, laws often require individuals to carry third-party liability insurance and industrial injury insurance. Fire insurance is required on immovable property in Germany. A number of countries require aviation insurance (for accident and sickness) on airline passengers and crews.

Although individuals generally have the freedom to select whichever insurer they wish, there are restrictions on freedom in buying insurance from foreign or alien insurers. In some countries, buyers must use domestic insurers for compulsory coverages but are free to take out insurance from foreign or alien insurers when coverage is not available from domestic insurers. In other countries, certain types of insurance may not be placed in foreign countries. About half of the countries of the world prohibit "nonadmitted" insurance, defined as insurance written in contravention of local laws and regulations.

In the United States, where there is relatively close control over rate making by the states, insurance rates must meet three basic requirements: the rates must be adequate to meet losses, must not be excessive, and must not be unfairly discriminatory among different classes of risk. Other aspects of insurance that are subject to regulation by state governments in the United States include: minimum standards of financial solvency; minimum amounts of initial capital; restrictions on the types of assets in which insurance companies may invest; procedures for orderly liquidation of insolvent insurers; minimum levels of

security deposits with government insurance commissioners; maximum expense allowances for the acquisition of new business; standard selling practices, including agency licensing, policy forms, and provisions; taxation; and service and process procedures governing ways in which suits may be brought to recover losses.

Insurance as a contract

Contract law. In general, an insurance contract must meet four conditions in order to be legally valid: it must be for a legal purpose; the parties must have a legal capacity to contract; there must be evidence of a meeting of minds between the insurer and the insured; and there must be a payment or consideration.

To meet the requirement of legal purpose, the insurance contract must be supported by an insurable interest (see further discussion below); it may not be issued in such a way as to encourage illegal ventures (as with marine insurance placed on a ship used to carry contraband).

The requirement of capacity to contract usually means that the individual obtaining insurance must be of a minimum age and must be legally competent; the contract will not hold if the insured is found to be insane or intoxicated, or if the insured is a corporation operating outside the scope of its authority as defined in its charter, bylaws, or articles of incorporation.

The requirement of meeting of minds is met when a valid offer is made by one party and accepted by another. The offer is generally made on a written application for insurance. In the field of property and liability insurance, the agent generally has the right to accept the insured's offer for coverage and bind the contract immediately. In the field of life insurance, the agent generally does not have this power, and the contract is not valid until the home office of the insurer has examined the application and has returned it to the insured through the agent.

The payment or consideration is generally made up of two parts—the premiums and the promise to adhere to all conditions stated in the contract. These may include, for example, a warranty that the insured will take certain loss-prevention measures in the care and preservation of the covered property.

Warranties. In applying for insurance the applicant makes certain representations or warranties. If the applicant makes a false representation, the insurer has the option of voiding the contract. Concealment of vital information may be considered misrepresentation. In general, the misrepresentation or concealment must concern a material fact—defined as a fact that would, if it were known, cause the insurer to change the terms of the contract or be unwilling to issue it in the first place. If the agent of the insurer asks the applicant a question the answer to which is a matter of opinion, and the answer turns out to be wrong, the insurer must demonstrate bad faith or fraudulent intent in order to void the contract. If, for example, in answer to an agent's question, the applicant reports no history of serious illness, in the mistaken belief that a past illness was minor, the court may find the statement to be an honest opinion and not a misrepresented fact.

Subrogation and indemnity

A basic principle of property liability insurance contracts is the principle of subrogation, under which the insurer may be entitled to recovery from liable third parties. In fire insurance, for example, if a neighbour carelessly sets fire to the insured's house and the insurance company indemnifies the insured for the loss, the company may then bring a legal action in the name of the insured to recover the loss from the negligent neighbour. The principle of subrogation is complemented by another basic principle of insurance contract law, the principle of indemnity. Under the principle of indemnity a person may recover no more than the actual cash loss; one may not, for example, recover in full from two separate policies if the total amount exceeds the true value of the property insured.

Insurable interest. Closely associated with the above legal principles is that of insurable interest. This requires that the insured be exposed to a personal loss if the peril insured against should occur. Otherwise it would be possible for a person to take out a fire insurance policy on the property of others, and collect if the property burned. Any financial interest in property, or reasonable expectation of having a financial interest, is sufficient to establish insurable interest. A secured creditor such as a mortgagee has an insurable interest in the property on which money has been lent. In the field of personal insurance one is held to have an unlimited interest in one's own life. A corporation may take life insurance on the life of a key executive. A wife may insure the life of her husband, and a father may insure the life of a minor child, because there is a sufficient pecuniary relationship between them to establish an insurable interest.

In life insurance, the insurable interest must exist at the time of the contract. Continued insurable interest, however, need not be demonstrated. A divorced woman may continue life insurance on the life of her former husband and legitimately collect the proceeds upon his death even though she is no longer his wife. In the field of property insurance, on the other hand, the insurable interest must be demonstrated at the time of the loss. If an individual insures a home but later sells it, no recovery may be made if the house burns after the sale, because the insured has no loss at the time of the fire.

Liability law. In most countries, an individual may be held legally liable to another for acts or omissions and be required to pay damages. Liability insurance may be purchased to cover these contingencies. In some countries (such as the U.S.S.R.) automobile liability insurance is not permitted, on the ground that this would allow the insured to escape the consequences of negligence; some of the costs, however, are paid by social insurance (in the case of injuries) and by individual accident insurance.

The concept of negligence

Legal liability exists when an individual commits a legal injury that wrongly invades another person's rights. Such injuries include slander, assault, and negligent acts. A negligent act involves failure to behave in a manner expected when the results of this failure cause a financial loss to others. An act may be classed as negligent even if it is unintentional. Negligence may be imputed from one person to another. For example, a master is liable not only for his own act but also for the negligent acts of servants or others legally representing him. It is not uncommon for a municipality to require that businesses using city property assume what would otherwise have been the city's negligence for the use of its property. Statutes may impute liability on individuals when no liability would exist otherwise; thus a parent may be legally liable for the acts of a minor child who is driving the family automobile.

In common-law countries such as the United States and the United Kingdom, three defenses may be used in a negligence action. These are assumed risk, contributory negligence, and the fellow servant doctrine. Under the assumed risk rule, the defendant may argue that the plaintiff has assumed the risk of loss in entering into a given venture and understands the risks. Employers formerly used the assumed risk doctrine in suits by injured employees, arguing that the employee understood and assumed the risks of employment in accepting the job.

The contributory negligence defense is frequently used to defeat negligence actions. If it can be shown that one party was partly to blame, then that party may not collect from any negligence of the other party. Some courts have applied a substitute doctrine known as comparative negligence. Under this, each party is held responsible for a portion of the loss corresponding to the degree of blame attached to that party; a person who is judged to be 20 percent to blame for an accident may be required to pay 20 percent of the other person's injuries.

The fellow servant defense has been used at times by employers; an employer would argue in some cases that the injury to an employee was caused not by the employer's negligence but by the negligence of another employee.

Workers' compensation statutes in some countries have nullified such common law defenses in industrial injury cases.

In many countries the courts have tended to apply increasingly strict standards in adjudicating negligence. This has been termed the trend toward strict liability, under which the plaintiff may recover for almost any accidental injury, even if it can be shown that the defendant has used "due care" and thus is not negligent in the traditional sense. In the United States, manufacturers of polio

vaccine that was found to have caused polio were required to pay large damage claims although it was demonstrated that they had taken all normal precautions and safeguards in the manufacture of the vaccine.

Historical development of insurance

Ancient origins of insurance

Insurance in some form is as old as historical society. So-called bottomry contracts were known to merchants of Babylon as early as 4000–3000 BC. Bottomry was also practiced by the Hindus in 600 BC, and was well understood in ancient Greece as early as the 4th century BC. Under a bottomry contract, loans were granted to merchants with the provision that if the shipment was lost at sea the loan did not have to be repaid. The interest on the loan covered the insurance risk. Ancient Roman law recognized the bottomry contract in which an article of agreement was drawn up and funds were deposited with a money changer. Marine insurance became highly developed in the 15th century.

In Rome there were also burial societies that paid funeral costs of their members out of monthly dues.

The insurance contract also developed very early. It was known in ancient Greece and among other maritime nations in commercial contact with Greece.

England. Fire insurance arose much later, obtaining impetus from the Great Fire of London in 1666. A number of insurance companies were started during the so-called bubble era in England after 1711. Many of them were fraudulent, get-rich-quick schemes, concerned mainly to sell their securities to the public. Nevertheless, two important and successful English insurance companies were formed during this period—the London Assurance Corporation and the Royal Exchange Assurance Corporation. Their operation marked the beginning of modern property and liability insurance.

Lloyd's of London

No discussion of the early development of insurance in Europe would be complete without reference to Lloyd's of London, the international insurance market. It began in the 17th century as a coffeehouse patronized by merchants, bankers, and insurance underwriters, gradually becoming recognized as the most likely place to find underwriters for marine insurance. Edward Lloyd supplied his customers with shipping information gathered from the docks and other sources; this eventually grew into the publication *Lloyd's List,* still in existence. Lloyd's was reorganized in 1769 as a formal group of underwriters accepting marine risks. (The word underwriter is said to have derived from the practice of having each risk taker write his name under the total amount of risk that he was willing to accept at a specified premium.) With the growth of British sea power Lloyd's became the dominant insurer of marine risks, to which were later added fire and other property risks. Today Lloyd's is a major reinsurer as well as primary insurer, but it does not itself transact insurance business; this is done by the member underwriters, who accept insurance on their own account and bear the full risk in competition with each other.

United States. The first American insurance company was organized by Benjamin Franklin in 1752 as the Philadelphia Contributionship. The first life insurance company in the American colonies was the Presbyterian Ministers' Fund, organized in 1759. By 1820 there were 17 stock life insurance companies in the state of New York alone. Many of the early property insurance companies failed from speculative investments, poor management, and inadequate distribution systems. Others failed after the Chicago fire of 1871 and the San Francisco fire and earthquake of 1906. There was little effective regulation, and rate making was difficult in the absence of cooperative development of sound statistics. Many problems also beset the life insurance business. In the era following the U.S. Civil War, bad practices developed: dividends were declared that had not been earned, reserves were inadequate, advertising claims were exaggerated, and office buildings were erected that sometimes cost more than the total assets of the companies. Thirty-three life insurance companies failed between 1870 and 1872, and another 48 between 1873 and 1877.

After 1910 life insurance enjoyed a steady growth in the United States. The annual growth rate of insurance in force over the period 1910–82 was approximately 8.2 percent—amounting to a 300-fold increase for the 72-year period. Over the period 1917–76, property and liability insurance assets grew some 65 times, an increase of about 7.3 percent annually.

Soviet Union. Following the revolution in 1917, insurance in the U.S.S.R. was nationalized. It is now administered by the organization known as Gosstrakh, which writes two basic forms of coverage: personal and property. Property insurance is the more important and is further divided into two general types: voluntary and required. Property insurance is required on all government-owned property and on certain property of collective farms. Voluntary property insurance is available for non-state-owned livestock, crops, household goods, automotive equipment, and freight.

Personal insurance coverages such as term and whole life insurance, accident insurance, and certain annuity contracts are available on a voluntary basis in the U.S.S.R. Insurance against tort liability is not sold in the U.S.S.R., on the ground that to allow such coverage would permit negligent persons to escape the financial consequences of their behaviour. Reinsurance also does not exist, although Ingosstrakh, which handles all foreign coverages, accepts reinsurance from other world insurance organizations. Most state-owned enterprises in the U.S.S.R. do not carry insurance as such but use self-insurance in the sense that losses are absorbed in the state budget.

Japan. Insurance in Japan is mainly in the hands of private enterprise, although government insurance agencies write crop, livestock, forest fire, fishery, export credit, accident and health, and installment sales credit insurance as well as social security. Private insurance companies are regulated under various statutes. Major classes of property insurance written include automobile and workers' compensation (which are compulsory), fire, and marine. Rates are controlled by voluntary rating bureaus under government supervision, and Japanese law requires rates to be "reasonable and non-discriminatory." Policy forms generally resemble those of Western nations. Personal insurance lines are also well developed in Japan and include ordinary life, group life, and group pensions. Health insurance, however, is incorporated into Japanese social security.

Worldwide operations. Because of the great expansion in world trade and in the extent to which business firms make investments outside the boundaries of their home countries, the market for insurance on a worldwide scale has expanded rapidly in the 20th century. This development has required a worldwide network of offices to provide brokerage services, underwriting assistance, claims service, and so forth. The majority of the world's insurance businesses are concentrated in Europe and North America. These companies must service a large part of the insurance needs of the rest of the world. The legal and regulatory hurdles that must be overcome in order to do so are formidable.

Mergers. In the mid-20th century there was a definite trend toward mergers and consolidations of insurance companies. In the United States during the period 1955–65 most such consolidations took place between property-liability companies and life insurance companies. After 1965, a new type of merger began to appear: holding companies known as congenerics, with subsidiaries consisting not only of insurance companies but also of companies performing other services such as banking, data processing, loss-prevention engineering, mutual fund management, real estate appraisal, and credit and collection. The advantages claimed for such conglomerates are various. A life insurance company may find that its total tax burden is less if it is part of a holding company rather than the owner of a number of subsidiary companies. Moreover, a holding company is able to expand by issuing preferred stock, a course not permitted to a life insurance company in many countries. This allows greater "leverage" in financial operations—that is, a company can control larger resources with a smaller investment than would otherwise be necessary. There are also merchandis-

Congeneric holding companies

ing advantages in wide diversification. The sales force can sell not only insurance but mutual funds, loss-prevention engineering, real estate appraisal services, and other related financial services as well.

BIBLIOGRAPHY. Basic surveys of the insurance field are given in MARK R. GREENE and JAMES S. TRIESCHMANN, *Risk and Insurance,* 6th ed. (1983); DAVID L. BICKELHAUPT, *General Insurance,* 11th ed. (1983); and ROBERT I. MEHR and EMERSON CAMMACK, *Principles of Insurance,* 7th ed. (1980). Group insurance is thoroughly represented in R.D. EILERS and ROBERT M. CROWE (eds.), *Group Insurance Handbook* (1965); more specialized treatments are ROBERT I. MEHR and SANDRA GUSTAVSON, *Life Insurance,* 3rd ed. (1984); and JERRY S. ROSENBLOOM (ed.), *The Handbook of Employee Benefits* (1984). Treatments of international developments in insurance are ROBERT M. CROWE (ed.), *Insurance in the World's Economies* (1982); and MICHAEL E. HOGUE and DOUGLAS G. OLSON (eds.), *World Insurance Outlook* (1982). Regulation of insurance is covered in SPENCER L. KIMBALL and HERBERT S. DENENBERG (eds.), *Insurance, Government and Social Policy* (1969), which stresses American regulation; and in a study by the ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, *Supervision of Private Insurance in Europe,* 2 vol. (1963), which compares regulation in major European countries. Current world developments in insurance are reported in *Sigma* and *Experiodica,* monthly publications of the Swiss Reinsurance Co., Zürich.

(M.R.G.)

Integumentary Systems

Every living organism is separated from its environment by a covering, or integument, that delimits its body. This covering serves as protection from foreign matter and gives communication with the environment. The integument enables an organism to live in a particular environment. Many animals are provided with smooth skins that are bathed in slime, which serves as protection (often with antibiotic qualities) and lubrication.

Among unicellular organisms, such as bacteria and protozoans, the integument corresponds to the cell membrane (see CELLS: *The plasma membrane*) and any secreted coating that the organism produces. In most invertebrate animals a layer (or layers) of surface (epithelial) cells—often with additional secreted coatings—constitutes the integument (Figure 1). Among the vertebrates the boundary covering—with a variety of derived elements such as scales, feathers, and hair—has assumed the complexity of an organ system, the integumentary system.

The integument is composed of layers that may be of single cell thickness, as in many invertebrates, or multiple cell thickness, as in some invertebrates and all vertebrates. In every case the cells that give rise to the integuments belong to that class of tissue called epithelium, which in most animals is called epidermis. Underlying the epidermis and supplying it with nourishment is the dermis. In addition to the cellular layers, the integument often includes a noncellular coating, or cuticle, that is secreted by the epidermis. Such coatings are found in most invertebrates. The vertebrate skin has generated many kinds of glands and a variety of horny structures, but it lacks coatings.

The wide diversity of integuments among vertebrates further exemplifies the adaptive character of the body covering: from the almost impenetrable shield of an armadillo and the dense furry coat of an Arctic bear to the slimy, scaled covering of a cod and the exceptionally smooth skin of a porpoise. Amphibians and fishes often have mucous glands that lubricate their skins and prevent waterlogging and deterioration. Reptiles have thick, leathery skins that help reduce water loss and serve as an armour against enemies. Birds use their feathers—skin derivatives—to fly and to insulate their bodies. The hairy or furry coats of many terrestrial mammals insulate them, shed water, and provide a dense guard against injury.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 334, 421, and 423.

This article is divided into the following sections:

Invertebrate integuments 691
  Components 691
  Differences among invertebrate groups 691
The vertebrate integumentary system 693
  Skin layers 693
  Skin derivatives and appendages 693
  Variations among vertebrates 695
  Embryology and evolution 698
  Biodynamics 698
The human skin 699
  Distinctive features 699
  Dermis 699
  Blood and lymph vessels 699
  The skin surface 700
  The epidermis 700
  Hair 701
  Sebaceous glands 702
  Sweat glands 703
  Nails 703
  Cutaneous sense organs 704
Skin diseases 704
  General features 704
  Hereditary skin diseases 705
  Generalized skin diseases 705
  Localized skin diseases 708
  Metabolic disorders 709
  Aging and the skin 709
Bibliography 710

Invertebrate integuments

COMPONENTS

The integuments of unicellular organisms are coincident with the cell membrane together with any coating it secretes. Multicellular invertebrates, however, have a single outer layer of epithelial cells, and these may secrete a variety of surface coatings, ranging from the mucous coat of cnidarians (polyps, sea anemones, jellyfish) to the tough cuticle of insects. The epithelial layer may include cells of several types, such as sensory cells, gland cells, and stinging cells, and the outer surfaces may bear microvilli, cilia, or other fine projections. In addition, the cells may produce excrescences such as bristles, spines, knobs, and ridges.

The firmness of some invertebrate animals, such as annelids (earthworms, marine worms, and leeches) and certain mollusks, depends on the distension by water of the individual cells that form the body wall. In many other forms rigid skeletal materials are deposited either within the cells or on the outer surface. The noncellular coatings of invertebrate integuments are exceedingly varied in composition and extent, and they cut across taxonomic categories. These rigid structures may provide no more than a protective armour, but in the arthropods, including crustaceans, insects, and spiders, a multilayered and hardened integument forms an exoskeleton to which muscles are attached. In the echinoderms, the exoskeleton lies below the epidermis.

Secreted coatings

DIFFERENCES AMONG INVERTEBRATE GROUPS

Protozoans. The secreted coatings of protozoans exhibit all grades between soft forms (as in *Amoeba*) and forms

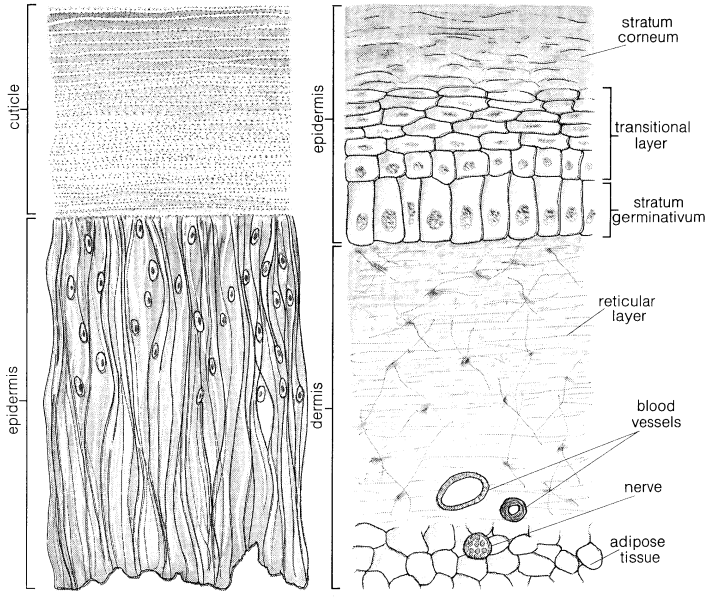


Figure 1: Comparison of (left) invertebrate and (right) vertebrate integuments.

From C.K. Weichert, *Anatomy of the Chordates*, 3rd ed., © 1965; used with permission of McGraw-Hill Book Co.

with a cuticle that may be proteinaceous (as in *Monocystis*) or composed of cellulose (as in the plantlike flagellates). Other protozoans have definite shells, composed of protein incorporating various foreign bodies, such as siliceous plates or calcium carbonate (in most foraminiferans), or cellulose (in the resting stages of slime molds). The radiolarians have an internal lattice of silica that is laid down inside the cell—a kind of internal skeleton, or endoskeleton.

Sponges. Sponges have a simple epithelium, known as the pinacoderm, which both covers the external surfaces and lines the internal waterways. Some sponges deposit needlelike spicules of calcium carbonate in the jelly (mesoglea) beneath this outer epithelium.

Cnidarians. In the cnidarians the epidermis provides all the basic features of an integument. It may contain not only epithelial cells, some of which may be contractile, but also gland cells, pigment cells, stinging cells, and sensory cells with projecting hairs. The outer surface often bears flagella or microvilli. The surface secretion may help in capturing food, adhering to substrates, cleaning away settling debris, or providing support and protection. Some hydrozoans produce a horny covering for the polyps, and

External skeleton

others have an external skeleton that is calcareous (containing calcium carbonate). Anthozoans show the same diversity. In the common reef-building corals, the calcareous skeleton is secreted by a part of the epidermis that forms a basal disk. This secretory process is continuous, and the polyp raises itself progressively upon a constantly growing stem of calcium carbonate.

Flukes and roundworms. The parasitic flukes have a relatively thick integument, which bears many spines and sensory papillae, an apical membrane that is thrown into ridges and pits, and microvilli. The outer part of the integument contains secretory bodies, which are continuously released at the surface to renew the apical membrane. This appears to be a protective device for the parasite related to the immune reaction of the host. Roundworms have a thick, flexible cuticle, with three distinguishable zones, covered by an epicuticle.

Annelids. Annelids have a thin, horny cuticle pierced by pores through which epidermal glands secrete mucus. In some marine annelids, glands are also present that secrete materials constituting a parchmentlike or calcareous tube within which the worm dwells. Earthworms and leeches secrete cocoons from a specialized epidermis in a region of the body known as the clitellum. A major feature of all annelids except leeches is the possession of bristles, or chaetae, of which there are many varieties. The bulk of each chaeta is secreted by a single cell, though the surrounding lateral cells may contribute materials that bring about its hardening.

Arthropods. The exoskeleton attains its most elaborate forms in the arthropods (for example, crustaceans and insects). The insect epidermis (Figure 2) lies on a basement membrane and secretes a tough cuticle, the bulk of which is composed of fibres of a material known as chitin embedded in a matrix of protein. Peripheral to this is an epicuticle. Chitin is a high-molecular-weight polysaccharide containing amino groups. It is synthesized within the epidermis from sugars and amino sugars.

In the integument of caterpillars chitin forms a cuticle that is tough but flexible. But in most arthropods the segments of the body or of the limbs are in the form of rigid plates that form a true exoskeleton linked to adjacent segments by flexible membranes. Such cuticles are hard and may be dark in colour. They are said to be tanned, or sclerotized, and in some species they are also mineralized.

The cuticle of caterpillars

Sclerotization involves the molecular stabilization of the protein chains of the cuticles by establishment of cross-links. Sclerotin, the product of sclerotinization, is a kind of natural plastic. In its horny consistency it closely resembles keratin; both are cross-linked, or polymerized, proteins, but the chemical nature of the linkage is different in the two substances. It is probable that other skeletal proteins in invertebrates, such as the spongin of sponges and the conchiolin of mollusks, are also tanned proteins allied to sclerotin.

In many crustaceans—crabs and lobsters, for example—much of the cuticle is rendered hard by the incorporation of calcareous substances such as aragonite or calcite. But sclerotin is actually harder than calcite, and those parts of crustaceans that need to be of maximum hardness, such as the mandibles and the tips of the claws, are in fact composed of sclerotin.

Besides functioning as a skeleton, the cuticle of terrestrial arthropods must act as a waterproof covering in order to prevent these small animals from drying up. This waterproofing is effected by the secretion of a layer of wax on the surface of the cuticle. Such a wax layer, if exposed in an unprotected state, would be excessively fragile. It is commonly protected by a thin layer of a cementlike substance that is poured over its surface by small dermal glands.

The cuticle of arthropods, pierced by ducts of dermal glands that pour out secretions over the surface, is a living structure; it can produce tactile bristles, pigment-bearing scales, claws, wings, and other structures. In some insects it shows brilliant metallic colours that result from the presence of multiple thin plates or ridges in the cuticle. In order that the arthropod may grow, the old cuticle is shed from time to time after a new and larger cuticle has been laid down beneath it. This process is termed molting, or ecdysis. During the time when the new cuticle is hardening, the arthropod is in a very vulnerable condition.

Molting in insects is hormonally controlled. A molting hormone, known as ecdysone, is mainly a product of the

From H. Weber, *Lehrbuch der Entomologie* (1933, reprinted 1968); Gustav Fischer Verlag

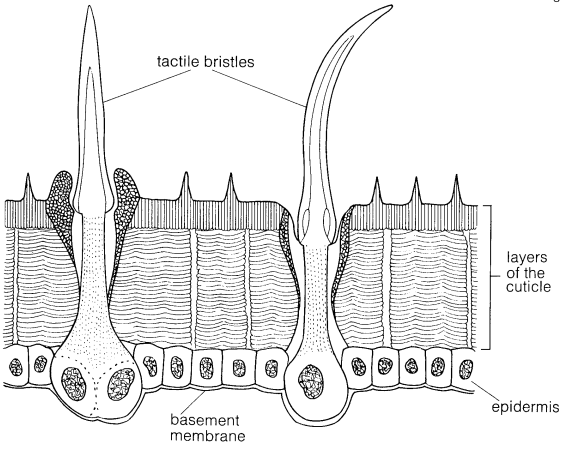


Figure 2: The integument of an insect.

thoracic glands, and its secretion is influenced by a prothoracicotropic (or ecdysiotropic) hormone produced by certain cells of the brain. Larvae also possess a juvenile hormone, which decreases in concentration until the imago (adult) emerges. Crustaceans and spiders possess analogous hormones, though their systems are not identical.

Mollusks. The epidermis of mollusks is capable of a variety of functions. Ciliated epithelium is of particular importance for feeding in bivalves and for the gliding movement of snails. Abundant gland cells secrete mucus, which protects the animal from predators and from dessication. Complex glands secrete the quinone-tanned proteins of the byssus threads, by which mussels anchor themselves, and of the operculum, with which some sea snails stopper their shells. The secretion of pallial glands enables the date mussel to bore into calcareous rock. Some cephalopods (squids, cuttlefish, octopuses) have luminous glands, although it is disputed whether the luminous material is produced by the epithelium itself or by bacteria. Cephalopods also have pigment cells that can be expanded by muscle contraction and can change colour very rapidly.

Shell

The shell of mollusks is secreted by the epithelium of the mantle and consists of an outer layer of the horny substance conchiolin, an intermediate prismatic layer composed of calcite, and a smooth inner layer (the nacreous layer) also composed mainly of calcium carbonate. The first two layers are secreted by a marginal band of cells, so that the shell grows at its outer edge. The nacreous layer is secreted by the general surface of the mantle and is the material of which pearls are formed around foreign bodies introduced into the mantle cavity.

Echinoderms. The echinoderms are characterized by a calcareous exoskeleton, which may be a rigid armour, as in echinoids (sea urchins), or of a leathery consistency, as in holothurians (sea cucumbers). The epidermis lies outside of this skeleton. The apical plasma membrane is capable of taking up dissolved organic molecules from the surrounding seawater in amounts that are at least enough for the nourishment of the epidermal layer. Many sea urchins have projecting spines on which the epidermis is worn away to expose the calcareous material.

The vertebrate integumentary system

SKIN LAYERS

In all vertebrates the skin has two major layers. The outer, relatively thin epidermis is composed of closely packed cells with little intercellular material; it provides the barrier against attack by chemicals, radiation, or microbes. The underlying dermis (cutis, corium) is thicker and tougher, and its bulk is formed by extracellular materials manufactured by scattered cells. One of its major functions is physical protection. The sensory functions of skin are shared by both epidermis and dermis. The various structures or appendages such as scales, feathers, claws, glands, and hair follicles are derived from either layer or from both. There are considerable differences between the skins of different vertebrate classes, and they are closely related to the environments of the various groups.

The epidermis is the product of the deepest layer of its cells, those that lie immediately over the dermis. From this generative layer, known as the stratum germinativum, cells move outward and become progressively flattened. The surface cells of terrestrial vertebrates, mere remnants of once living cells, are scaly and compressed; they constitute the horny layer, or stratum corneum. The cell fragments of the stratum corneum are composed largely of keratin, a tough insoluble protein. In most land vertebrates the stratum corneum is shed or molted, either periodically and in large fragments or sheets, as in reptiles, or continuously in small patches or scales, as in mammals.

The dermis, which is best developed in mammals, consists largely of fibrous connective tissue (composed of collagen fibres), blood and lymph vessels, smooth muscle cells, and nerve endings. It gives rise to so-called membrane bones—the bony scales of fishes, the bony plates in certain reptiles and mammals, and the membrane bones of the vertebrate skull. Through its blood network the dermis supplies nourishment to the overlying epidermis.

SKIN DERIVATIVES AND APPENDAGES

Among the notable changes that have taken place during the course of evolution is the development in vertebrates of a variety of glands, pigmentary structures, scales, claws, nails, horns, feathers, and hairs as adaptations to their changing environments.

Skin glands. The glands of the skin are all exocrine, that is, they secrete their products, usually through ducts, to the epidermal surface. They may be unicellular, as are the goblet cells of fishes, or multicellular, as are the sweat glands of humans. Some multicellular glands are tubular and extrude their secretion into a central space or lumen; some, like the oil-producing sebaceous glands of mammals, form their product by complete breakdown of the cells, a method of secretion known as holocrine. Glands may consist of tubes or sacs, and they may be singular, clustered, or branched; some even contain units of more than one type (Figure 3). They may secrete their product continuously, periodically, or only once.

Exocrine secretion

From C.K. Weichert, *Anatomy of the Chordates*, 3rd ed. © 1965; used with permission of McGraw-Hill Book Co.

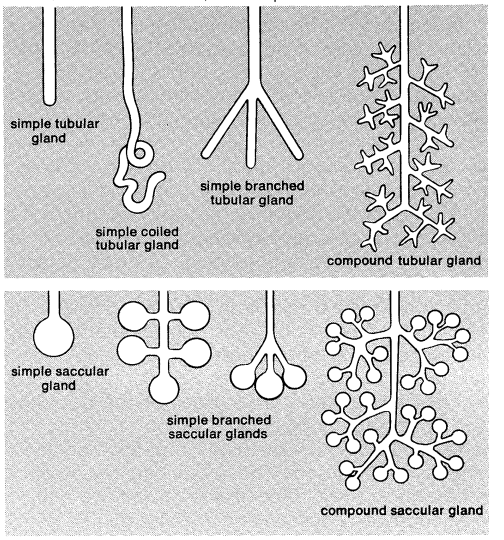


Figure 3: Types of vertebrate integumentary glands.

Mucous glands secrete a protein called mucin, which with water forms the substance known as mucus; this slimy material serves to lubricate the body, thus lessening friction and aiding locomotion in swimming animals. Serous glands produce a watery secretion; sweat glands of mammals are of this type. Sebaceous glands secrete oil, ceruminous glands secrete wax, mammary glands secrete milk, poison glands secrete various toxins, and scent glands secrete a variety of odoriferous substances. Further, certain epidermal glands may be modified into light-producing structures called photophores, seen in the skin of many deep-sea fishes.

Pigment cells. In fishes, pigment is produced in branched cells known as chromatophores, which can be found in both epidermis and dermis. Rapid colour change, by which some fishes can adapt to a change of background, is brought about by redistribution of the pigment within the cell boundaries (Figure 4). Slow, long-term changes in-

Chromatophores

From C.K. Weichert, *Anatomy of the Chordates*, 3rd ed., © 1965; used with permission of McGraw-Hill Book Co.

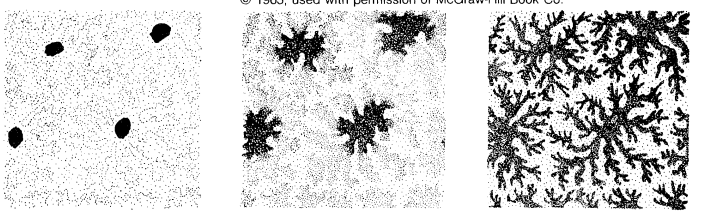


Figure 4: *Pigment dispersion and resulting change in skin colour.*
(A) Pigment concentrated in centre of cell. (B) Intermediate condition. (C) Pigment dispersed throughout cytoplasm of chromatophores.

volve alterations in the numbers of cells or in the amount of pigment they contain.

Chromatophores are also present in amphibians and reptiles, but not in birds or mammals, which possess pigment cells called melanocytes. Melanocytes are found mainly in the epidermis, though they occur elsewhere. They also are branched, or dendritic, and their dendrites are used to transfer pigment granules to adjacent epidermal cells. A number of different pigments are produced in the different vertebrate groups, but in mammals only brown eumelanin and yellow or red phaeomelanin are important. Pigment cells—chromatophores and melanocytes alike—are influenced by melanocyte-stimulating hormones of the pituitary.

Epidermal scales. Epidermal scales are horny, tough extensions of the stratum corneum. Well developed in reptiles, they are also common on exposed skin in birds and mammals (Figure 5). Such scales are periodically molted or shed gradually along with the rest of the stratum corneum. Epidermal scales are absent in fishes, but dermal, or bony, scales are abundant. Clawlike epidermal scales are present in certain amphibians, including a few toads, certain burrowing, wormlike caecilians, and the salamander *Hynobius.* The so-called horns of the horned lizard are specialized epidermal scales; and the rattle of rattlesnakes is a series of dried scales loosely attached to each other, the last one always remaining despite molting of the rest of the stratum corneum (Figure 5). Epidermal scales cover the bony scales of the carapace (top) and plastron (bottom) of turtles' shells. The beak of turtles is composed of a modified epidermal scale covering the jawbone.

Birds' bills

Bills of birds are similarly constructed. In birds, epidermal scales are confined to the lower legs, feet, and base of the bill. The spurs of some birds are bony projections covered with a scalelike sheath. The skin of the webs in aquatic birds is also scaly. In mammals, except for a few cases, epidermal scales are largely restricted to the tails and paws. The overlapping horny plates of the pangolin are modified epidermal scales (Figure 5).

Claws, nails, and hooves. In many animals, hardened corneal growths occur at the end of the digits, growing parallel to the skin surface. True claws—found in reptiles, birds, and mammals—consist of a dorsal scalelike plate (unguis) covering a ventral plate (subunguis), the whole capping the bony tip of a digit. Nails—found only in mammals—consist of a broad and flattened unguis, with the subunguis reduced to a vestige under the outer tip. Hooves, the characteristic feature of the hoofed mammals, or ungulates, are exaggerated nails, with the unguis curved all around the end of the digit and surrounding the subunguis (Figure 5).

Horns and antlers. Horns are hardened corneal projections of several types. Except for certain lizards, horns are found only in mammals. The keratin fibre horn is unique to the rhinoceros. It consists of a cone of keratinized cells that grows from an epidermis covering a cluster of dermal bumps (papillae). The fibres, somewhat resembling thick hair, grow from the papillae, and cells between the papillae produce a cement that binds the fibres together.

Hollow horns are found in cattle, sheep, buffalo, goats, and other ruminants. In certain species only the males display them. Such horns consist of an extension of the frontal bone, a permanent part of the cranium covered by a horny layer. The horn of the pronghorn antelope is unique in that the horny covering is shed periodically and a new one is formed from the epidermis that persists over the bony extension (Figure 5).

Antlers, which are characteristic features of the deer family, are not integumentary derivatives at all. Fully developed antlers are solid bone, without any epidermal covering. The young antlers, however, are covered with skin having a velvety appearance. When the antler is fully developed, the drier skin cracks and is rubbed off by the animal. Antlers in giraffes are small and remain permanently covered.

Feathers and hair. Birds and mammals display remarkable elaborations of the epidermis in the form of feathers and hair, respectively. These distinctive features are dealt with below (see *Birds; Mammals*).

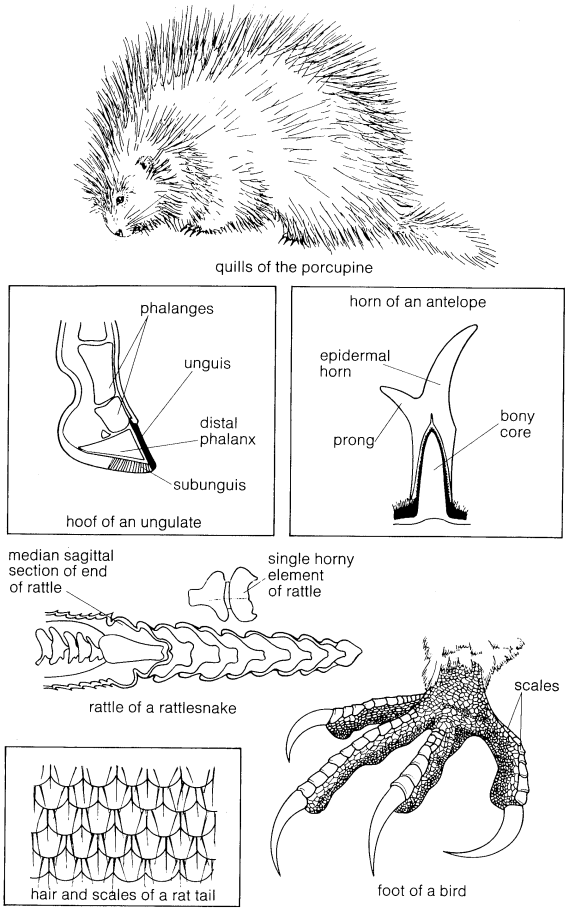


Figure 5: Vertebrate epidermal appendages.

From (quills of the porcupine) W. Montagna, "The Skin," Copyright © 1965 by Scientific American, Inc., all rights reserved; (foot of a bird) A.L. Thomson, *A New Dictionary of Birds* (1964), Nelson, London; (hoof of an ungulate) after Ellenberger in Leisering's Atlas, based on Sisson, *The Anatomy of the Domesticated Animals,* W.B. Saunders Company, from C.K. Weichert, *Anatomy of the Chordates,* Copyright 1965, used with permission of McGraw-Hill Book Company; (others) C.K. Weichert, *Anatomy of the Chordates,* Copyright 1965, used with permission of McGraw-Hill Book Company

Dermal derivatives. Dermal scales are found almost exclusively in fishes and some reptiles (Figure 6). They are bony plates that fit closely together or overlap and form the dermal skeleton. Highly developed dermal scales are seen in turtles, where the bony plates form a rigid dermal skeleton that is attached to the true skeleton. In other reptiles, dermal scales are small and localized on parts of the body, as in crocodilians, certain lizards, and a few snakes.

The armadillo's scales

Birds lack dermal scales, and only a single living mammal—the armadillo—displays them. Associated with the

From (A) W. Montagna, "The Skin," © 1965 by Scientific American, Inc., all rights reserved; (B) from a painting by Tom Dolan; (C) after Flower, from Kingsley, *Comparative Anatomy of Vertebrates,* The Blakiston Division, McGraw-Hill Book Company, by permission

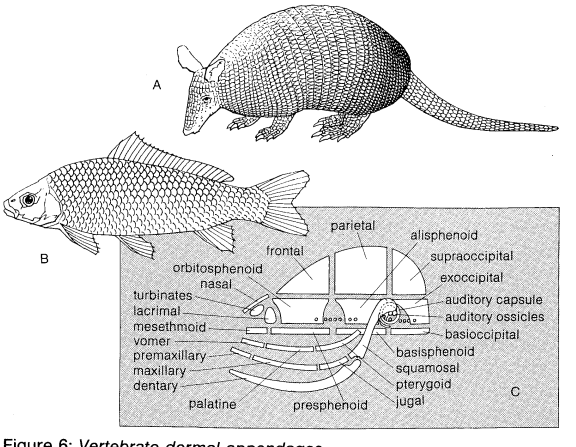


Figure 6: *Vertebrate dermal appendages.*
(A) Armadillo showing scales. (B) Carp with large scales. (C) Membrane bones of mammalian skull.

evolutionary tendency toward elaboration of epidermal extensions in birds and mammals, there has been a corresponding reduction in dermal derivatives. The membrane bones of the skull, the mandible (lower jaw), and the clavicles (collarbones) are the remaining vestiges of dermal plates in these groups (Figure 6).

VARIATIONS AMONG VERTEBRATES

The vertebrates belong to the phylum Chordata and are closely related to a small, fishlike, almost transparent invertebrate called amphioxus. Amphioxus represents chordate integument at its simplest: an epidermis, consisting of one layer of columnar or cuboidal epithelial cells and scattered mucous cells, covered by a thin cuticle, and a thin dermis of soft connective tissue. Beginning with the simplest vertebrates, the cyclostomes (lampreys and hagfishes), the integument becomes complex and pigmented; in successive evolutionary stages a wide array of derivatives appears among the various classes of vertebrates.

The lamprey

Cyclostomes. In the lamprey the surface of the skin is smooth, with no scales. The epidermis consists of several cell layers that actively secrete a thin cuticle. Gland cells producing slime are mixed with the epidermal cells, as in most aquatic vertebrates. The dermis is a thin layer of connective tissue fibres interwoven with blood vessels, nerves, muscle fibres, and chromatophores.

Fishes. Fishes have a more or less smooth, flexible skin dotted with various kinds of glands, both unicellular and multicellular. Mucus-secreting glands are especially abundant. Poison glands, which occur in the skin of many cartilaginous fishes and some bony fishes, are frequently associated with spines on the fins, tail, and gill covers. Photophores, light-emitting organs found especially in deep-sea forms, may be modified mucous glands. They may be used as camouflage or to permit recognition, either for repulsion to delimit territory or for attraction in courtship.

Also formed within the skin of many fishes are the skeletal elements known as scales (Figure 7). They may be divided into several types on the basis of composition and structure. Cosmoid scales, characteristic of extinct lungfishes and not found in any fishes today, are similar to the ganoid scales of living species. Placoid scales (or denticles) are spiny, toothlike projections seen only in cartilaginous fishes. Ganoid scales, sometimes considered a modification of the placoid type, are chiefly bony but are covered with an enamel-like substance called ganoin. These rather thick scales, present in some primitive bony fishes, are well developed in the gars.

Cycloid scales appear to be the inner layer of ganoid or cosmoid scales. Found in carps and similar fishes, they are thin, large, round or oval, and arranged in an overlapping pattern; growth rings are evident on the free edges. Ctenoid scales are similar to cycloid, except that they have spines or comblike teeth along their free edges; these scales are characteristic of the higher bony fishes—perches and sunfishes, for example. Some fishes, such as catfishes and some eels, have no scales.

Among the cartilaginous fishes, sharks have a very tough skin. Scattered over it are denticles, each with a pulp cavity, around the edge of which is a layer of odontoblasts. These cells secrete the dentine, or calcareous material, of the scale. Outside the dentine is the enamel, secreted by the overlying ectoderm. When the denticles pierce through the ectoderm, no more enamel can be added.

Bony fish scales

The dominant modern fishes, teleosts, are characterized by bony scales covered with skin. The epithelium of a trout's epidermis provides the animal with an inert covering of keratin. The scales lie in the dermis as thin, overlapping plates with the exposed part bearing the pigment cells. The scale is deposited in a series of annual rings, since its growth occurs rapidly in spring and summer and rarely in winter.

Amphibians. Most modern amphibians lack horny scales or other protective devices. An exception is seen in the caecilians, a small group that has fishlike scales similar to those possessed by ancient and extinct forms. The amphibian epidermis has five to seven layers of cells formed from a basal stratum germinativum. At the skin surface, in contact with the external environment, the cells are keratinized to form a stratum corneum, which is best developed in amphibians that spend most of their time on land. The cells of this horny layer are not continuously shed but are periodically molted in sheets. Molting is controlled by the pituitary and thyroid glands but is unaffected by sex hormones. The wartiness of toads results from local thickenings.

Some amphibian families have disklike pads on their digits for adherence to underlying surfaces. During the breeding season the males of anurans (frogs and toads) and urodeles (salamanders and newts) develop nuptial pads on some digits of the forelimbs, which facilitate firm gripping of the females; the pads are induced to form by androgenic (male) hormones.

The dermis is two-layered, having an outer and looser stratum spongiosum and an inner stratum compactum. Although some amphibians have external gills or internal lungs, for many the skin is a vital respiratory organ, and the dermis is richly supplied with blood vessels and lymph spaces. Chromatophores are located just below the junction of the dermis with the epidermis. The numerous mucous and poison glands originate from nests of epidermal cells that grow down into the dermis.

Reptiles. In the evolutionary sense, reptiles are the first truly terrestrial vertebrates, since they have dispensed with an aqueous environment for their larval development.

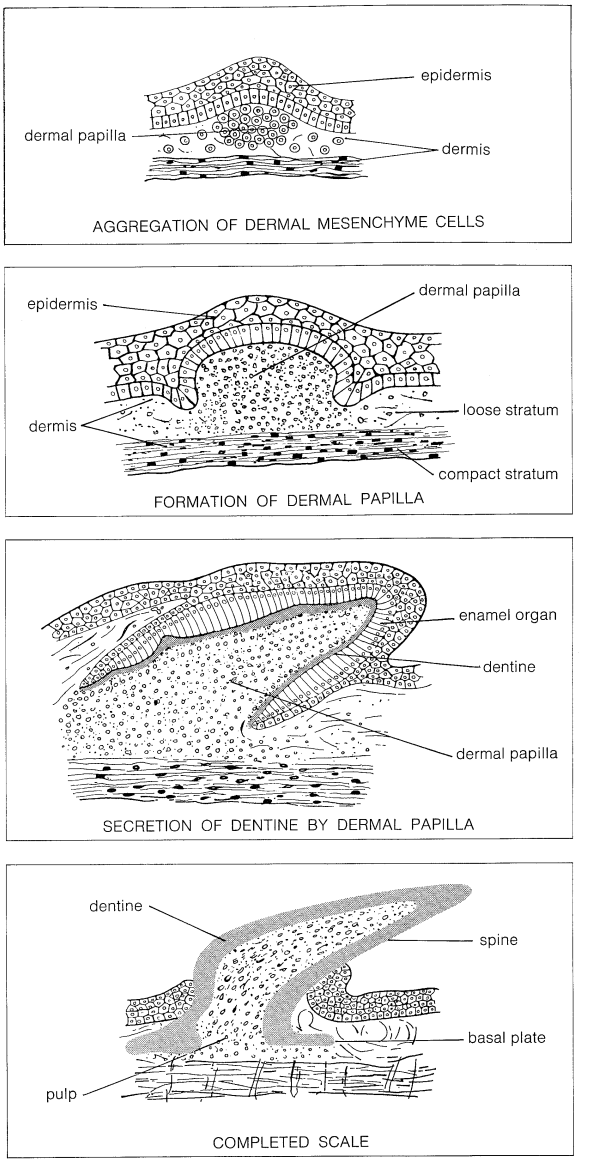


Figure 7: Fish scale development.

From L. Hyman, *Comparative Vertebrate Anatomy*, (© 1922, 1942) by the University of Chicago; all rights reserved

Their main problem is to prevent dessication by water loss through the skin. This is solved by the possession of a thick stratum corneum in which waxes are arranged in membranelike layers between the keratinized cells. Reptilian scales are overlapping folds of skin, each scale having an outer surface, an inner surface, and a hinge region. All the epidermal and dermal surfaces of each scale are continuous with those of the next scale.

Keratin

The cornified part of the epidermis is strengthened by a stiff material, beta keratin, which is present in place of or in addition to pliable alpha keratin. In crocodiles and many turtles the outer scale surface consists of beta keratin only, while the hinge region contains only alpha keratin. In lizards and snakes, however, both keratins form continuous layers, the alpha keratin lying below the beta keratin. In crocodiles and turtles there is continuous cell division in the stratum germinativum and exfoliation of cells at the skin surface. In snakes and lizards the germinal layer forms a complete new epidermal surface before the whole of the old cornified epidermis is sloughed, either in a single sheet or in portions.

The shape and size of the scales vary in the different families and with the mode of life. Maximum flexibility of the skin is achieved in some forms by reduction of the scales to small, nonoverlapping granules. Among desert dwellers there is a tendency for some scales, particularly those on the head and tail, to be enlarged to form spines. Burrowing and secretive forms have a slippery body surface because of the presence of smooth, highly polished scales. The skin is often reinforced by bony plates, which lie beneath the superficial scales (though corresponding with them in size and shape); these plates may form a continuous protective armour. Other defensive, or sometimes offensive, devices associated with the skin and scales are the occasional development of horns or fringing folds that break up the animal's outline and colouring.

The colours of reptiles are produced by both melanocytes in the epidermis and three types of chromatophores in the dermis: melanophores, which contain melanin; xanthophores, which contain yellow pigments; and iridophores, which contain reflecting platelets of colourless guanine. The pattern may be fixed, for concealment by camouflage, or the chromatophores may provide for rapid colour change.

Reptilian glands

Reptilian skin possesses glands, but they are usually small. Most are holocrine; some are tubular. Lizards and snakes have small glands that are related to the sloughing cycle, and all groups of reptiles appear to communicate by scent glands. For example, chelonians (turtles and tortoises) have glands in the throat, inguinal, and axillary regions, and snakes have saclike scent glands at the base of the tail.

Birds. The avian epidermis is thin, delicate, and clothed in feathers, except on the obviously naked areas of the legs, feet, beak, comb, and wattle. On the legs and feet, and sometimes elsewhere, the cornified layer is thickened to form scales of several types. The dermis, also thin, consists mostly of a network of connective tissue fibres and muscle fibres that help to adjust the feathers. In larger birds, such as the ostrich, the skin is thick enough to allow it to be processed into leather. The scales resemble those of reptiles in possessing layers containing beta keratin and alpha keratin.

Feathers, which consist of beta keratin, are considered to have evolved from reptile scales (Figure 8). They are periodically molted, and other keratinized structures such as the bill and claws may be molted as well. Pigment is primarily restricted to feathers and scales. Specialized nerve endings are present throughout the skin. Various holocrine and tubular glands have been observed, but nearly all are small and inconspicuous. The exception is the holocrine uropygial gland, or preen gland, located on the back just in front of the tail, which secretes oil for grooming the feathers. It is largest in aquatic birds.

Feathers are unique to birds. Those of adults are admirably engineered to be lightweight yet strong. They are of three basic types, each associated with certain functions. Contour feathers (including the flight and tail feathers) define the body outline and serve as aerodynamic devices;

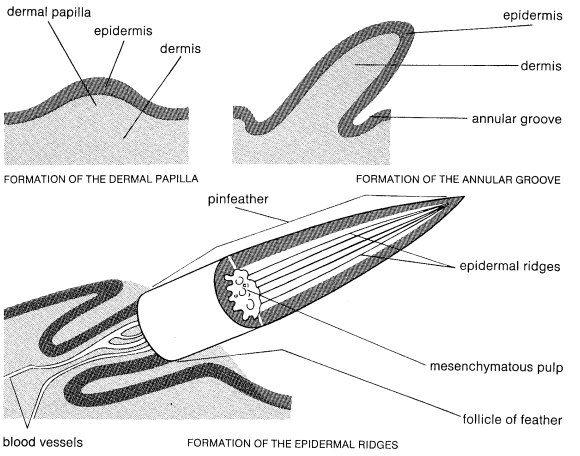


Figure 8: Development of a typical down feather. The annular groove is the beginning of the feather follicle. The epidermal ridges give rise to the barbs of the feather.

From C.K. Weichert, *Anatomy of the Chordates*, 3rd ed., © 1965; used with permission of McGraw-Hill Book Co.

filoplumes (hair feathers) and plumules (down feathers) are used principally as insulation, to conserve body heat. Colours and patterns in feathers serve as protective coloration or for sexual display.

In most birds contour feathers are not uniformly distributed over the surface of the body but are arranged in feather tracts (pterylae) separated from one another by regions of almost naked skin (apteria). The only exceptions are the ostrichlike birds, the penguins, and the South American screamers, in which the even distribution of plumage has probably been secondarily acquired. Feather tracts differ in arrangement in different species and hence are useful in the classification of birds.

The wing tract includes the flight feathers proper (remiges) and their coverts (tectrices). The remiges include the primaries, arising from the "hand" and digits and attached to the hand's skeleton; the secondaries, arising from the forewing and attached to the ulna; and the tertials (when present), arising from the upper wing and attached to the humerus. The tectrices cover the bases of the remiges, overlapping and decreasing in size toward the leading edge of the wing.

The spinal (dorsal) tract extends the whole length of the bird, excepting the head, along and on both sides of the spinal column. In gallinaceous birds this tract may be subdivided from front to back (though not separated by apteria) into the regions of the hackle, the cape, the back, and the saddle. Each region is distinguished by the form and pattern of its constituent feathers.

Regional feather patterns

On the ventral surface of the bird are paired breast tracts, with a ventral tract between them. The tail tract includes the tail feathers (rectrices) and their coverts. Other tracts cover the head, base of the wings, and legs.

A contour feather of an adult bird tends to be almost bilaterally symmetrical. It consists of a tapering central shaft, the rachis, to which are attached a large number of tapering parallel barbs. These in turn carry many minute elongated barbules on both their distal and proximal faces. The distal barbules bear tiny hooklets (hamuli) that fit into grooves on the proximal barbules of the next higher barb. In this way the barbules overlap and interlock to form the coherent web, or vane, of the feather. Barbules in the basal portions of feathers are long, delicate threads and do not bind successive barbs together; consequently, this part of the feather is fluffy.

The filoplumes, which arise at the bases of contour feathers, are inconspicuous hairlike feathers bearing a small tuft of barbs at their apexes. Filoplumes appear to be present in all birds, but only in certain species do they project beyond the contour feathers—on the thighs of cormorants, for example.

Plumules are present in young birds before they develop

the adult plumage. In adults the plumules are generally scant and are concealed by contour feathers; however, in many birds, such as gulls and ducks, they form a thick, insulating undercovering comparable to the underfur of seals. Their barbs do not form coherent vanes but are long, loose, soft, and fluffy. Their structure is much simplified, and a rachis may be entirely lacking. In herons and some hawks the tips of the plumules disintegrate into a fine scaly powder that becomes distributed over the plumage, providing protection against wetting and giving it a peculiar sheen; accordingly, these specialized down feathers are called powder down.

Coloration

Feathers get their colours from a number of pigments. Melanin is responsible for black, gray, brown, and related tints; yellow or reddish brown granules of phaeomelanin and dark brown granules of eumelanin are transferred to the epithelial cells of the feather from melanocytes. Some feathers are coloured bright yellow, vivid red, green, violet, or blue by carotenoids and other rare pigments. Cosmetic coloration of the feathers by the secretion of the preen gland is exploited by pelicans. Not all coloration requires pigments. The striking white of sea gulls and swans is a "structural colour" produced by the reflection of light by irregularly distributed air-filled cavities. Blue, green, and violet can also be structurally produced, as, for example, in kingfishers and parrots.

Mammals. An important distinguishing character of mammals is their hair. They also possess many other horny derivatives of the epidermis, including nails, claws, hooves, quills, and horns. All mammalian hard keratin, as well as the soft keratin of the stratum corneum, is of the alpha type. Bony dermal plates are found in the armadillo. Antlers, too, are made of bone and derived from the dermis, but they have an epidermal covering—the velvet—when newly grown.

Skin structure. The mammalian epidermis has several layers of cells, known as keratinocytes, which arise by cell division in a basal stratum germinativum. This rests on a basement membrane closely anchored to the surface of the dermis. Newly formed cells move outward, and at first form part of the prickle cell layer (stratum spinosum), in which they are knit together by plaquelike structures called desmosomes. Next they move through a granular layer (stratum granulosum), in which they become laden with keratohyalin, a granular component of keratin. Finally the cells flatten, lose their nuclei, and form the stratum corneum. The dead cells at the skin surface are ultimately sloughed, or desquamated. In thick, glabrous skin lacking hair follicles, such as that on human palms and soles, a clear layer, the stratum lucidum, can be distinguished between the stratum granulosum and the stratum corneum.

The important barrier to outward loss of water or inward passage of chemicals lies in a compact zone of the lower stratum corneum. There the spaces between the layers of the cornified cells are tightly packed with lipid (waxy) platelets that have been produced inside so-called membrane coating granules within the underlying epidermal cells. As well as the clear horizontal stratification of the epidermis, a vertical organization is also apparent, at least in nonglabrous skin, in the sense that the ascending keratinizing cells appear to form regular columns.

In the basal layer, groups of keratinocytes are each associated with a single dendritic (branching) pigment cell to form "epidermal melanocyte units." In addition to keratinocytes and melanocytes, the mammalian epidermis contains two other cell types: Merkel cells and Langerhans cells. Merkel cells form parts of sensory structures. Langerhans cells are dendritic but unpigmented and are found nearer the skin surface than melanocytes. After a century of question about their purpose, it is now clear that they have a vital immunologic function.

The dermis

The dermis forms the bulk of the mammalian skin. It is composed of an association of connective tissue fibres, mainly collagen, with a ground substance of mucopolysaccharide materials (glycosaminoglycans), which can hold a quantity of water in its domain. Two regions can be distinguished—an outer papillary layer and an inner reticular layer. The papillary layer is so called by reason of the numerous microscopic papillae that rise into the epidermis, especially in areas of wear or friction on the skin. These papillae, not to be confused with the "dermal papillae" of the hair follicles (see below), are arranged in definite patterns beneath epidermal ridges. In humans these external ridges are responsible for the fingerprints, or dermatoglyphs. The reticular layer has denser collagen than the papillary layer, and it houses the various skin glands, vessels, muscle cells, and nerve endings.

Hair. In evolution, the overriding importance of hair is to insulate the warm-blooded mammals against heat loss. Hairs have other uses, however. Their function as sensory organs may, indeed, predate their role in protection from cold. Large stiff hairs (vibrissae), variously called whiskers, sensory hairs, tactile hairs, feelers, and sinus hairs, are found in all mammals except humans and are immensely helpful to night-prowling animals. Vibrissae are part of a highly specialized structure that contains a mass of erectile tissue and a rich sensory nerve supply. These specialized hairs are few in number, their distribution being confined chiefly to the lips, cheeks, and nostrils and around the eyes; they occur elsewhere only occasionally. Human eyelashes consist of sensory hairs that cause reflex shutting of the eyelid when a speck of dust hits them.

Hair may also be concerned in sexual or social communication, either by forming visible structures, like the mane of the lion or the human beard, or by disseminating the product of scent glands, as in the ventral gland of gerbils or the human axillary organ. Hair is important as well in determining the coloration and pattern of the mammalian coat, serving as either camouflage or self-advertisement.

In essence, each hair is a cylinder of compacted and keratinized cells growing from a pit in the skin—the hair follicle. The follicle consists mainly of a tubular indentation of the epidermis that fits over a small stud of dermis—the dermal papilla—at its base. Indeed, it is formed in the embryo by just such as interaction between its constituents, the epidermis growing inward as a peg that ultimately invests a small group of dermal cells.

The epidermal components of an active hair follicle consist of an outer layer of polyhedral cells, forming the outer root sheath, and an inner horny stratum, the inner root sheath. This inner sheath is composed of three layers, known respectively as Henle's layer (the outermost), consisting of horny, fibrous, oblong cells; Huxley's layer, with polyhedral, nucleated cells containing pigment granules; and the cuticle of the root sheath, having a layer of downwardly imbricate scales (overlapping like roof tiles) that fit over the upwardly imbricate scales of the hair proper. The outer root sheath is surrounded by connective tissue. This consists internally of a vascular layer separated from the root sheath by a basement membrane—the hyaline layer of the follicle. Externally, the tissue has a more open texture corresponding to the deeper part of the dermis that contains the larger branches of the arteries and veins.

A small muscle, the arrector pili, is attached to each hair follicle, with the exception of the small follicles that produce only fine vellus hairs. If this muscle contracts, the hair becomes more erect and the follicle is dragged upward. This creates a protuberance on the skin surface, producing the temporarily roughened condition that is popularly called gooseflesh.

"Gooseflesh"

The hair shaft is composed chiefly of a pigmented, horny, fibrous material, which consists of long, tapering fibrillar cells that have become closely impacted. Externally, this so-called cortex is covered by a delicate layer of imbricated scales forming the cuticle. In many hairs the centre of the shaft is occupied by a medulla, which frequently contains minute air bubbles, giving it a dark appearance. The medullary cells tend to be grouped along the central axis of the hair as a core, continuous or interrupted, of single, double, or multiple columns.

The cuticular scales of mammalian hairs are predominantly of the overlapping, imbricate type, with edges that are rounded, minutely notched, or flattened. They vary in size, shape, and edge structure and are distinctive for each species. Among the higher primates, for example, those of chimpanzees are slightly oval, those of gorillas and humans have shallowly notched edges, and those of orangutans have edges that are deeply notched.

In many deer the cortical substance can hardly be distinguished; almost the entire hair appears to be composed of thin-walled polygonal cells. In the peccary the cortical envelope sends radial projections inward, the spaces between being occupied by medullary substance; and this, on a large scale, is the structure of the porcupine's quills.

One of the most remarkable mammalian hairs is that of the Australian duckbill, or platypus, where the lower portion of the shaft is slender and woollike, while the free end terminates as a flattened, spear-shaped, pigmented hair with broad imbricate scales. In the three-toed sloth a microscopic alga grows between the cuticular scales of the hairs and appears to be symbiotic; its presence gives a curious greenish gray hue to the coat of the sloth and helps to disguise the animal among the trees.

The activity of hair follicles is cyclic. After an active period (known as anagen), the follicle passes through a short transition phase (catagen) to enter a resting phase (telogen). In this process, cell division ceases, and the dermal papilla is released from the epidermal matrix, which becomes reduced to a small, inactive, secondary germ. The base of the hair expands and becomes keratinized to form a "club," which is held in the follicle until the next cycle begins. A new period of anagen starts with cell proliferation of the secondary germ, which then extends inward to reinvest the dermal papilla. After the new hair is formed, the old club hair is shed, or molted. The events of early anagen are, in effect, a reenactment of the early development of the hair follicle.

The final length of any hair depends mainly on the duration of anagen and varies between body sites and from animal to animal. Hairs on the back of a rat take three weeks to grow fully, whereas the follicles on the human scalp may be continuously active for three years or more.

Molting

The cyclic activity of hair follicles is the mechanism by which mammals molt; it thus enables animals to alter their coats as they grow or as they adjust to changing temperature-control or camouflage requirements. In some mammals molting takes place in a pattern, so that the follicles act in synchrony in a particular area of the body. In the human scalp the follicles are out of step with each other, and there is continuous loss of club hairs.

Glands. The skin glands of mammals are of three major types. Associated with hair follicles are oil-secreting sebaceous glands as well as tubular glands, which produce an aqueous secretion. Sebaceous glands are termed holocrine because their secretion involves complete disintegration of their cells, which are constantly replaced. Tubular, or merocrine, glands extrude their secretion into a central lumen. The tubular glands of the hair follicle are usually classified as apocrine because it is believed that, in some glands at least, secretion involves a breaking off of part of the gland cells. A second type of merocrine gland, not associated with hair follicles, is termed eccrine because the cells remain intact during secretion. Eccrine glands occur in hairy skin only in humans and some primates; but the footpad glands, which increase friction and thus prevent slipping in many mammalian species, are of a similar type.

A major function of skin glands is the production of odours for sexual or social communication. Many species in all but a few mammalian orders have specialized aggregations of glandular units for this purpose. These occur in almost every area of the body. Some, like the chin and anal glands of the rabbit, contain only tubular units; others, like the abdominal gland of the gerbil, are purely sebaceous; still others, like the side glands of shrews, contain batteries of both holocrine and tubular units.

In some large mammals an important function of merocrine glands is temperature control. Horses and cattle, for example, have apocrine glands for this purpose, but the superbly effective cooling system of humans is served by eccrine sweat glands.

EMBRYOLOGY AND EVOLUTION

The skin of vertebrates begins to form early in embryonic development, from a superficial germ layer, the ectoderm. The middle germ layer, or mesoderm, proliferates cells rapidly from segmental building blocks, called somites; these cells then migrate in order to lie directly under the outer ectodermal covering. These two embryonic layers—ectoderm and mesoderm—ultimately give rise to the adult skin; the ectoderm produces the epidermis and its derivatives, and the mesoderm produces the dermis.

The human fetus, at least, produces a specialized temporary embryonic skin, known as the periderm. For much of the second trimester of gestation, the periderm consists of cells with projecting globules covered with small protrusions, or microvilli. These cells are subsequently sloughed off as the stratum corneum is formed underneath them.

Differentiation of embryonic tissues proceeds rapidly during the early course of development, and much of what will become adult skin structures—including the glands and appendages—is laid down before the animal is born, often in a latent stage, to resume development later.

Evolutionary changes of the epidermis

As a surface constantly exposed to the environment, the epidermis has undergone more adaptive changes during evolution than any other portion of the skin. Ancestral vertebrates, aquatic and fishlike, were buffeted by water, which kept the living surfaces moist.

The movement to land was gradual and fraught with risk. Amphibians were among the first vertebrates to explore the terrestrial environment. Many evolved a semiaquatic lifestyle, exploiting the land for most of their activities but returning to the water for reproduction. Some remained entirely aquatic, and others adapted to a strictly terrestrial life. Their epidermises reflected such habits: aquatic amphibians developed a thin, slimy, dull skin densely covered with mucous glands; terrestrial forms acquired a thicker, horny, heavily pigmented skin dotted with poison glands.

The reptiles became even more independent of the water. Their skins grew tough, horny, and dry and sometimes received bony contributions from the dermis. Birds evolved a loose, dry skin covered with feathers for insulation and for airfoils and water foils. Finally, mammals adopted a dry, elastic skin, more or less covered with hair. The range of mammalian skin, from smooth (glabrous), as in the cetaceans (whales, dolphins, and porpoises), to densely hairy, as in Arctic bears, is associated with the dispersion of mammals into a wide range of habitats.

BIODYNAMICS

The vertebrate skin—despite its variety—serves the two common functions of protection from, and communication with, the environment. In all land vertebrates the uppermost layers of the skin are dead, but the dermis is richly endowed with living tissue that can respond rapidly to change. A variety of nerve endings constantly report current conditions, and the body makes continuous adjustments in response.

It has been said that the skin is the largest and most versatile organ of the animal body. It shields against injury, against foreign matter and disease organisms, and against potentially harmful rays of the Sun. It also regulates internal body temperature through its insulating ability and its influence on the blood flow. Further, it embodies the sense of touch and adorns the body. Its contours, colour, patterns, and composition aid in species recognition and sexual attraction.

The effectiveness of the skin as a barrier, however, is not complete. Noxious substances that can gain entry evoke an immune response, and the dermis reddens with the rush of blood to the site. Heat also causes expansion of the dermal blood vessels—and in humans and in horses stimulates the sweat glands to heightened activity—thus increasing the loss of body heat. Conversely, cold causes contraction of the vessels and initiates shivering, thereby conserving heat in the first instance and generating it in the second.

Skin microorganisms

The skin is host to a number of microorganisms, especially bacteria and fungi. It is, however, an unstable environment for this population, which lives on the dead epidermal surface that is periodically sloughed off. A normal microcosm exists on most epidermal surfaces. Over the course of evolution an alliance has been established between the skin biota and the epidermal "host," which tends to stabilize the surface; anything that disrupts the skin biota encourages an imbalance and a potential flare-up of certain microorganisms over others. (F.J.G.E./Ed.)

The human skin

DISTINCTIVE FEATURES

The apparent lack of body hair immediately distinguishes human beings from all other large land mammals. Regardless of individual or racial differences, the human body seems to be more or less hairless, in the sense that the hair is so vestigial as to seem absent; yet in certain areas hair grows profusely. These may be referred to as epigamic areas, and they are concerned with social and sexual communication, either visually or by scent from glands associated with the hair follicles.

The characteristic features of skin change from the time of birth to old age. In infants and children it is velvety, dry, soft, and largely free of wrinkles and blemishes. Children younger than two years sweat poorly and irregularly; their sebaceous glands function minimally. At adolescence hair becomes longer, thicker, and more pigmented, particularly in the scalp, axillae, pubic eminence, and the male face. General skin pigmentation increases, localized pigmented foci appear mysteriously, and acne lesions often develop. Hair growth, sweating, and sebaceous secretion begin to blossom. As a person ages, anatomical and physiological alterations, as well as exposure to sunlight and wind, leave skin, particularly that not protected by clothing, dry, wrinkled, and flaccid.

Human skin, more than that of any other mammal, exhibits striking topographic differences. An example is the dissimilarity between the palms and the backs of the hands and fingers. The skin of the eyebrows is thick, coarse, and hairy; that on the eyelids is thin, smooth, and covered with almost invisible hairs. The face is seldom visibly haired on the forehead and cheekbones. It is completely hairless in the vermilion border of the lips, yet coarsely hairy over the chin and jaws of males. The surfaces of the forehead, cheeks, and nose are normally oily, in contrast with the relatively greaseless lower surface of the chin and jaws. The skin of the chest, pubic region, scalp, axillae, abdomen, soles of the feet, and ends of the fingers varies as much structurally and functionally as it would if the skin in these different areas belonged to different animals. Figure 9 shows the structures typically found in the skin of humans.

The skin achieves strength and pliability by being composed of numbers of layers oriented so that each complements the others structurally and functionally. To allow communication with the environment, countless nerves—some modified as specialized receptor end organs and others more or less structureless—come as close as possible to the surface layer, and nearly every skin organ is enwrapped by skeins of fine sensory nerves.

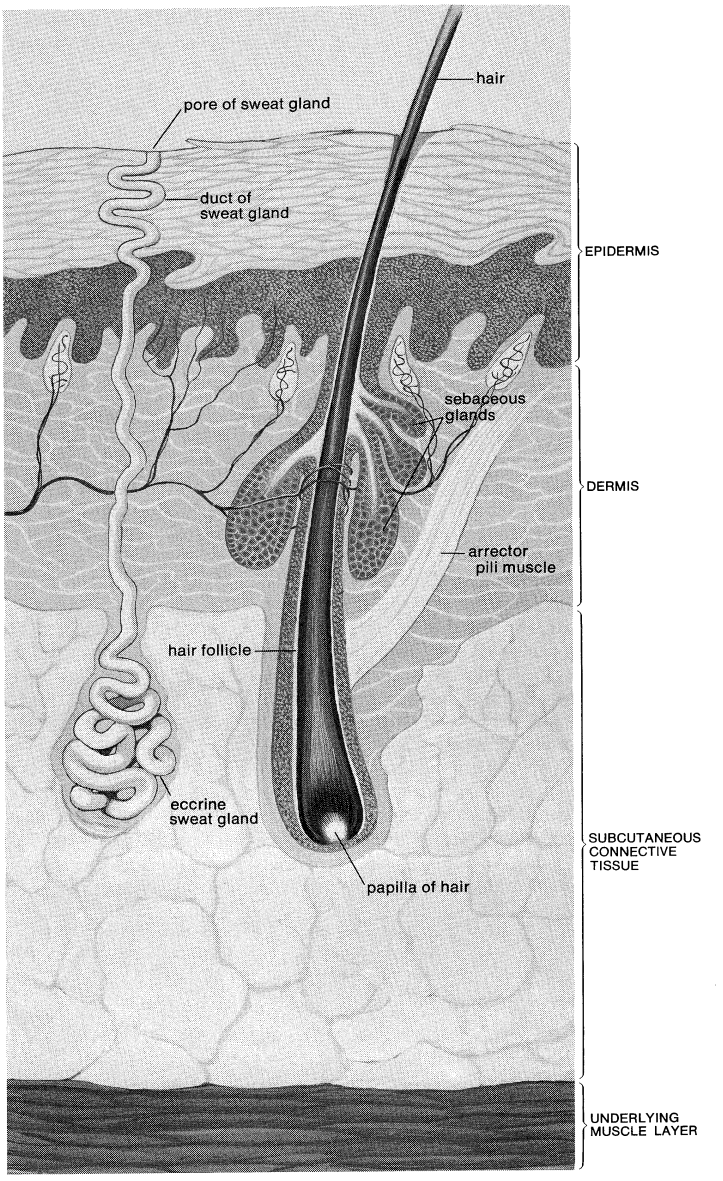


Figure 9: Section through human skin and underlying structures.

Adapted from *Cunningham's Textbook of Anatomy* edited by G.J. Romanes and published by Oxford University Press as an Oxford Medical Publication

THE DERMIS

The dermis makes up the bulk of the skin and provides physical protection. It is composed of an association of fibres, mainly collagen, with materials known as glycosaminoglycans, which are capable of holding a large amount of water, thus maintaining the turgidity of the skin. A network of extendable elastic fibres keeps the skin taut and restores it after it has been stretched.

The hair follicles and skin glands are derived from the epidermis but are deeply embedded in the dermis. The dermis is richly supplied with blood vessels, although none penetrates the living epidermis. The epidermis receives materials only by diffusion from below. The dermis also contains nerves and sense organs at various levels.

BLOOD AND LYMPH VESSELS

Human skin is enormously well supplied with blood vessels; it is pervaded with a tangled, though apparently orderly, mass of arteries, veins, and capillaries. Such a supply of blood, far in excess of the maximum biologic needs of the skin itself, is evidence that the skin is at the service of the blood vascular system, functioning as a cooling device. To aid in this function, sweat glands pour water upon its surface, the evaporation of which absorbs heat from the skin. If the environment is cold and body heat must be conserved, cutaneous blood vessels contract in quick, successive rhythms, allowing only a small amount of blood to flow through them. When the environment is warm, they contract at long intervals, providing a free flow of blood. During muscular exertion, when great quantities of generated heat must be dissipated, blood flow through the skin is maximal.

Temperature control

In addition to its control of body temperature, skin also plays a role in the regulation of blood pressure. Much of the flow of blood can be controlled by the opening and closing of certain sphincterlike vessels in the skin. These vessels allow the blood to circulate through the peripheral capillary beds or to bypass them by being shunted directly from small arteries to veins.

Human skin is permeated with an intricate mesh of lymph vessels. In the more superficial parts of the dermis, minute lymph vessels that appear to terminate in blind sacs function as affluents of a superficial lymphatic net that in turn opens into vessels that become progressively larger in the deeper portions of the dermis. The deeper, larger vessels are embedded in the loose connective tissue that surrounds the veins. The walls of lymph vessels are so flabby and collapsed that they often escape notice in specimens prepared for microscopic studies. Their abundance, however, has been demonstrated by injecting vital dyes inside the dermis and observing the clearance of the dye.

Because lymph vessels have minimal or no musculature

in their walls, the circulation of lymph is sluggish and largely controlled by such extrinsic forces as pressure, skeletal muscle action, massaging, and heat. Any external pressure exerted, even from a fixed dressing, for example, interferes with its flow. Since skin plays a major role in immunologic responses of the body, its lymphatic drainage is as significant as its blood vascular system.

THE SKIN SURFACE

The intact surface of the skin is pitted by the orifices of sweat glands and hair follicles—the so-called pores—and is furrowed by intersecting lines that delineate characteristic patterns. All individuals have roughly similar markings on any one part of the body, but the details are unique. The lines are oriented in the general direction of elastic tension. Countless numbers of them, deep and shallow, together with the pores, give every region of the body a characteristic topography. Like the deeper furrows and ridges on the palms and soles, the skin lines are mostly established before birth. The fine details of each area of body surface are peculiar to each individual. Fingerprints are used as a means of personal identification because they have a high relief, more evident patternings, and can be easily obtained.

Some of the lines on the surface of the skin are acquired after birth as a result of use or damage. For example, furrows on the forehead are an accentuation of preexisting congenital lines that become strongly emphasized in old age. As the skin becomes less firm with aging, it also forms wrinkles. Certain occupations leave skin marks that, depending upon duration and severity, may be transient or permanent.

Dermatoglyphics

The ventral surfaces of the hands and feet are etched by distinct alternating ridges and grooves that together constitute dermatoglyphics. The ridges follow variable courses, but their arrangement in specific areas has a consistent structural plan. Though apparently continuous, the ridges have many interruptions and irregularities, branching and varying in length. Every small area of surface has ridge details not matched anywhere in the same individual or in any other individual, even in an identical twin. This infallible signature makes dermatoglyphics the best-known physical characteristic for personal identification.

THE EPIDERMIS

The epidermis is thicker on the palms and soles than it is anywhere else and is usually thicker on dorsal than on ventral surfaces. Omitting the fine details, it is divisible everywhere into a lower layer of living cells and a superficial layer of compact dead cells.

General structure. All the cells, living or dead, are attached to one another by a series of specialized surfaces called attachment plaques, or desmosomes. Thus, instead of being completely fused, the membranes of adjacent cells make a zipperlike contact, with fluid-filled spaces between the contact areas. This structural pattern ensures a concatenation of cells to one another so that they cannot be sloughed off easily; at the same time it allows nutrient fluids to seep in from the vessels in the dermis. Epidermal cells, which multiply chiefly at the base in contact with the dermis, gradually ascend to the surface, manufacturing keratin as they go. They finally die in the upper part, forming a horny layer.

The epidermis is thickest on friction surfaces and thinnest over the eyelids, on the lower parts of the abdomen, and around the external genitalia. Unlike that of most other mammals, it has an intricately sculptured underside and does not lie flat upon the dermis. Seen from beneath, there are straight and branching ridges and valleys, columns and pits, all finely punctuated.

Because of this unevenness, it is almost impossible to state the exact thickness of epidermal tissue. Furthermore, individual differences, sex, and age have an enormous influence on the structure of the underside. Such labyrinthine patterns give human epidermis two unique advantages: it attains a more intimate connection with the subjacent dermis than if the surface were flat, and its source of dividing cells, the building blocks of the horny layer, is greatly increased.

Major layers. The clear stratification of the epidermis is the result of well-defined changes in its major constituent cells—the keratinocytes, or corneocytes—as they move peripherally from the basal layer, where they are continuously formed by mitosis, to the skin surface, where they are lost. In essence, the epidermis consists of a living malpighian layer, in contact with the basement membrane (which is attached to the dermis), and a superficial cornified (horny) layer of dead cells.

The innermost cells of the malpighian layer, next to the basement membrane, make up the basal layer, or stratum basale. Immediately peripheral to the basal layer is the spinous, or prickle-cell, layer—the stratum spinosum. Its cells have a spiny appearance due to the numerous desmosomes on their surface. Studies with the electron microscope have revealed that desmosomes are symmetrical, laminated structures in which some layers are contributed by the plasma membranes of adjoining cells and some form an intercellular component.

Desmosomes

The spinous layer is succeeded by the granular layer, or stratum granulosum, with granules of keratohyalin contained in the cells. These small particles are of irregular shape and occur in random rows or lattices. The cells of the outer spinous and granular layers also contain much larger, lamellated bodies—the membrane-coating granules. They are most numerous within the cells of the spinous layer. In the granular layer they appear to migrate toward the periphery of each cell and to pass into the intercellular spaces, where they discharge their waxy lipid components.

Peripheral to the granular layer is the stratum corneum, or horny layer, in which the keratinocytes have lost their nuclei and most of their organelles and contents, including the keratohyalin granules. They become progressively flattened and filled with keratin and are ultimately desquamated. Between the granular layer and stratum corneum, an unstainable stratum lucidum, or hyaline layer, can be recognized in palmar and plantar epidermis and some other regions.

Dynamics and organization. Horizontal stratification is the most obvious histological feature of the epidermis. There is also, however, distinct evidence of vertical organization. In thin epidermis, though not in thick plantar skin, the cornified cells can be shown to be arranged in regular stacks, which must reflect the underlying dynamic mechanisms. It appears that several living spinous cells are precisely and symmetrically stacked beneath each cornified column and that these are related to their own basal cells; cells do not pass from one stock to another.

All keratinocytes are formed by mitosis in the lower

William Montagna

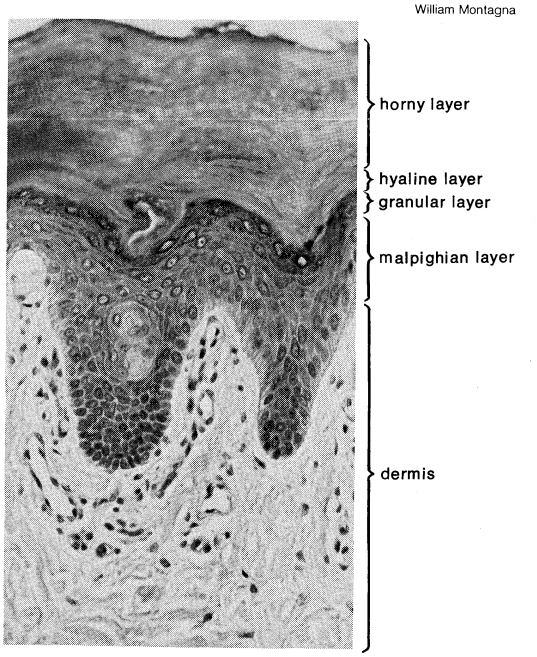


Figure 10: Photomicrograph of a vertical section of epidermis from a finger.

region of the malpighian layer. Most of the dividing cells are found in the basal layer, although it is likely that about one-third of the divisions occur above this level. Proliferating cells undergo a cycle: mitosis is followed by an interphase, this in turn is followed by a phase of DNA synthesis, and then another short resting phase occurs before mitosis begins again. The complete mitotic cycle takes about 12 to 19 days. The time for the passage of cells through the epidermis, from formation to desquamation, has been variously estimated at one to three months.

In normal skin the production and loss of cells must be finely balanced; otherwise the thickness of the epidermis would fluctuate. When the epidermis becomes abnormally thick, as in the plaques of psoriasis, this balance is altered. Either the production of cells in the malpighian layer must be abnormally high or their time of passage must be decreased. It is now generally agreed that such conditions result from a greatly increased production of cells; in fact, the cells move more, not less, rapidly through the epidermis.

There is, however, a further controversial problem. If all the basal cells were continuously cycling, greater production could be achieved only by a substantial reduction in the duration of the cell cycle. An alternative hypothesis is that not all the cells are undergoing cycles at any one time, so that greater cell production can be achieved by recruiting resting cells into activity. It seems likely that the epidermis does indeed contain noncycling cells, which can become activated, and that the cell cycle in psoriatic epidermis is speeded up only about twofold, not twelvefold, as once proposed.

Variations in rate of cell growth

When skin is wounded, there is a burst of epidermal mitotic activity about 40 hours later. It is evident, therefore, that local mechanisms of control must come into play; either inhibitors are dispersed by wounding, or stimulating hormones are released, or both. There is, on the one hand, some evidence of the existence of inhibitors, or chalones, but they have not been characterized. On the other hand, an epidermal growth factor (EGF) has been isolated from the salivary glands of mice and its chemical structure determined (a single-chain, folded polypeptide with 53 amino acid residues and three intramolecular disulfide bonds). It is not, however, extractable from skin, though the receptor protein to which it attaches in order to perform its action is present in many skin cells, and a closely similar molecule has been isolated from human urine.

The horny layer. The final product of the epidermis is the keratin that packs the cornified cells. The term keratin is applied generally to the hard keratins of hair, horn, and nails, and to the soft keratin of the epidermis. They are all insoluble filamentous proteins, composed of polypeptide chains that are stabilized by links using two atoms of sulfur. The source of the keratin of the stratum corneum has been a subject of controversy; but it is now generally accepted that about a third of its total mass is made up of proteins synthesized in the granular layer and the remainder from so-called intermediate filaments, which are present in keratinocytes from the basal layer outward.

The barrier that prevents water loss from the body is situated in the lower part of the horny layer. In this region the spaces between the compacted layers of keratin-filled cells contain lamellae of lipid (wax) that has been formed within the membrane-coating granules of the live epidermal cells below.

Pigmentation. The human skin is variously coloured and shows remarkable individual variations even within racial groups. The appearance of the skin is partly due to the reddish pigment in the blood of the superficial vessels. In the main, however, it is determined by melanin, a pigment manufactured by dendritic cells called melanocytes, found among the basal cells of the epidermis. Their numbers in any one region of the body, which range from about 1,000 to more than 2,000 per square millimetre, are roughly the same within and between races; the blondest whites have as many as the darkest blacks. Colour differences are due solely to the amount of melanin produced and the nature of the pigment granules. When the skin becomes tanned on exposure to sunlight, the melanocytes do not increase in number, only in activity.

All melanocytes, whether resident in the basal epidermis or in the matrix of the hair, have migrated there during embryonic life from a region known as the neural crest. Each epidermal melanocyte is associated with a group of neighbouring keratinocytes into which it transfers granules of pigment by way of long, branching dendrites. The whole has been termed an epidermal melanocyte unit. Once inside the epidermal cells, the melanin granules tend to move above the nucleus, forming a shroud over it. Such an orientation of melanin suggests that it is there to protect the cells from damaging ultraviolet rays, and experiments with tissue cultures support this view.

Protective function of melanin

Melanin is of two kinds: dark brown eumelanin and pale red or yellowish phaeomelanin. Both are formed within the melanocytes by the initial oxidation of the amino acid tyrosine with the aid of the enzyme tyrosinase; subsequently their synthetic pathways diverge. In addition to protecting the skin from ultraviolet radiation, epidermal pigmentation forms epigamic markings. The heavy pigmentation of the nipples and areolae of breasts, as well as that in the labia minora, penis, and scrotum, is related to sexual communication.

Immunoregulation and Langerhans cells. Although synthesis of protective keratin is clearly a major function of the epidermis, the discovery of an immunoregulatory role for the epidermis has revolutionized concepts of its importance in the immune defense systems of the host. In addition to melanocytes, human epidermis contains another system of dendritic cells, which do not manufacture pigment. Their distribution extends farther toward the skin surface than that of the pigment cells. After their discovery by the German physician Paul Langerhans in 1868, their function remained obscure until it was realized that they are a vital part of the immunologic mechanism.

Electron microscopic examination has revealed that the morphological hallmark of the Langerhans cell is a unique tennis-racket-shaped organelle, the Birbeck granule. Langerhans cells can be looked upon as "sentinel" cells of the immune system. By virtue of their situation, they are among the first cells to come into contact with foreign particulate substances encountering the skin. Their function is aided by the large surface area created by the dendritic processes of the cell. By means of specialized receptors on the cell membrane, the Langerhans cell recognizes invading as opposed to host molecules. By conveying this information to the lymphoid system, the body is able to mount a defensive immunologic response to the foreign material.

The concept that the role of keratinocytes themselves is confined to the synthesis of a horny protective outer covering for the skin has also become outmoded. Keratinocytes secrete a number of immunostimulatory high-molecular-weight peptides, collectively termed epidermal cytokines. One of these, the epidermal-cell-derived thymocyte-activating factor (ETAF), is also secreted by epidermal Langerhans cells. It has the function of enhancing the immune responsiveness of the lymphoid system as well as apparently being involved in the body's systemic reaction to infection and injury. Because the whole blood volume circulates through the skin every few minutes, immunoregulatory substances released by the cells of the epidermis may have a profound influence on the body's capacity to mount immune responses to viral or bacterial infections or to cancerous growths.

HAIR

Human hair has little protective value, even in hirsute persons. Eyelashes, eyebrows, and the hairs inside the external ears and nostrils have obviously useful functions, and scalp hair may be thick enough to provide some protection from the midday Sun. The beard and mustache, though, are embellishments, which establish maleness and are likely to be concerned with sexual or social communication; and axillary and pubic hair probably form part of scent-disseminating mechanisms.

An important role for hair, however, is its participation in the body's sensory apparatus. All hair follicles are surrounded by sensory nerves, and pressure on the hair shaft is transmitted to these nerves. Other mammals, including

subhuman primates, have highly specialized sensitive hair follicles around the eyes, lips, and muzzle. These produce "tactile" hairs, known as vibrissae or whiskers, which are particularly large in nocturnal mammals. The follicles from which these hairs emerge are rich in nerves and are surrounded by a sinus filled with blood. Humans are the only animals with no sinus hair follicles; but human hair follicles, particularly those of the face, are well supplied with nerves, and human skin is probably more sensitive than that of any other mammal.

Hair follicles

Hairs are manufactured by follicles. Essentially, these are tubelike pockets of the epidermis that extend through most or all of the depth of the skin and enclose a small papilla of dermis in their base. They lie at an angle to the skin surface. Two-thirds of the way up is a bulge, and attached to it are wisps of smooth muscle fibre that, on contracting, pull the follicle to a more or less perpendicular position. This action also puckers the skin into a mound on the surface—a so-called goose pimple.

Exactly as in other mammals, the human hair is formed by division of cells in the region known as the bulb, at the base of the follicle. Pigment is incorporated from melanocytes in this region. Human hair follicles also go through cycles of activity. After a period of growth, the hair becomes clubbed, rather than cylindrical, in shape. Fibrous rootlets anchor the club to the surrounding follicular tissue. While forming the club, the follicle shrivels up, the lower part becoming largely dissipated. A resting follicle can be recognized at once by the clubbed hair and by the follicle's short size and unique structure. Follicles remain dormant for variable periods of time. When they become active again, they reconstruct a bulb that manufactures a new hair. As the new hair works its way to the surface, the club hair is loosened from its moorings and shed.

The activity of the hair follicles in the scalp is not synchronized, so that there is a small but steady molt of about 50 to 100 hairs a day from a total of around 150,000 follicles. There is, nevertheless, evidence of seasonal fluctuation, with the greatest hair loss in late summer and fall. A follicle may continue its activity for a long time, and hairs sometimes grow for several years and attain considerable lengths. Even in the human scalp—where the hair follicles are dense and vigorously productive—baldness occurs in a large number of individuals. Baldness is not a disease but is a systematic involution of hair follicles, culminating in organs similar to the primitive embryonic follicles; the numbers of follicles do not necessarily diminish.

Until late in fetal life the forehead is indistinguishable from the scalp, with no line of demarcation between the two. After the fifth month of gestation the follicles in the rest of the scalp grow larger, but those of the forehead do not. After birth the hairs on the forehead become even smaller and nearly invisible. The hairline of newborn infants is usually indistinct; the familial hairline pattern is defined late in childhood through a process that is identical with that of baldness. When male pattern baldness sets in, in the late 20s or earlier, the follicles affected undergo exactly the same changes as do those that establish the hairline.

Male pattern baldness and its female equivalent, which is usually more diffuse, are hereditary conditions. Paradoxically, since male hormones (androgens) stimulate the growth of most other hair, this type of baldness occurs only if androgens are present.

Hairs vary in colour, diameter, and contour. The different colours result from variations in the amount, distribution, and type of melanin pigment in them, as well as from variations in surface structure that cause light to be reflected in different ways. Hairs may be coarse or so thin and colourless as to be nearly invisible. Straight hairs are round, while wavy hairs are alternately oval and round; very curly and kinky hairs are shaped like twisted ribbons.

Human hair grows at the rate of about one-third of a millimetre a day, and once keratinized it is inert. If the colour or shape of a hair is altered as it is formed, several days must elapse before the effect becomes visible. Hairs become white with aging because of the failure of the melanocytes to inject pigment into the cells as they are formed. Tales of hair becoming white overnight may perhaps arise from cases of rapid differential shedding of pigmented hairs from a mixed population of white and dark ones, but the suggestion that individual dark hairs can somehow rapidly turn white is not true.

Androgen-dependent hair

The beard and mustache are the most obvious examples of hair that requires male sex hormones, or androgens, for its growth. Facial hairs begin to develop at puberty, about two years after the start of pubic hair growth. The rate of growth of the beard initially increases with age, but levels out after 35. Hair on the chest—a traditional sign of masculinity—and that on the limbs are also androgen-dependent. Androgens cause longer hairs to be formed, partly by making them grow faster, but mainly by increasing the length of anagen, the growing phase. Fully formed hairs on the thigh are over three times longer in young men than in women; and the duration of anagen is around 54 days in males, compared with 22 days in females.

Pubic and axillary hair are also dependent on androgens, but they differ from other body hair in that they are luxuriant in females as well as in males. Their growth requires lower levels of hormone. The lower triangle of pubic hair is present in persons with a rare disease known as male pseudohermaphroditism. These individuals are genetic males who remain female in form until puberty because they lack an enzyme necessary to bind two atoms of hydrogen to the male hormone testosterone, which is responsible for male sex characteristics. Since persons with male pseudohermaphroditism lack facial hair, even when adults, it seems that beard growth requires the conversion of testosterone to dihydrotestosterone but that pubic hair growth does not.

SEBACEOUS GLANDS

The sebaceous glands are usually attached to hair follicles and pour their secretion, sebum, into the follicular canal. In a few areas of the body, disproportionately large sebaceous glands are associated with very small hair follicles; in other areas there are glands that are altogether free of follicles.

The outstanding feature of sebaceous glands is their holocrine mode of secretion, involving complete disintegration of the sebaceous cells. The glands consist of a series of lobes, or acini, each with a duct running toward the main sebaceous duct. The cells are generated by cell division around the periphery of each lobe. As they move toward the centre of the lobe and toward the duct, they synthesize and accumulate fat globules and become progressively larger and distorted. Ultimately they disintegrate to form the secretion.

Human sebum is a complex mixture of lipids—triglyceride fats (57.5 percent), wax esters (26 percent), squalene (12 percent), cholesterol esters (3 percent), and cholesterol (1.5 percent). The triglycerides are largely hydrolyzed by bacteria by the time the sebum reaches the skin surface, so that about a third of the surface fat consists of free fatty acids.

The activity of the sebaceous glands is mainly controlled by androgens. The glands are quite large at birth because of the influence of maternal hormones during development, but they regress soon afterward. They become active again at, or somewhat in advance of, the first signs of puberty. Their rate of secretion is a little higher in adult men than in women, and it falls off gradually with age in both sexes. It is very low in eunuchs but has been shown to increase when they are treated with androgens. That other factors—for example, pituitary hormones—may also influence secretion is suggested by the observation that sebum production is abnormally high in acromegaly, a disorder resulting from excessive secretion of growth hormone.

Problematic function of sebum

The function of sebum has been greatly debated. Some scientists have theorized that it is important as an emollient to prevent too rapid loss of water from the superficial layers of the stratum corneum; others have held that it is a functionless product of now useless, or vestigial, organs. Yet humans have more and larger glands than most mammals, and there is a specific plan in their distribution: they are largest and most numerous on the face and around the anogenital surfaces.

The skin around the nose, mouth, and forehead and over

the cheekbones has beds of gigantic glands, the secretion of which keeps these surfaces constantly oily. The sebaceous glands evenly spaced in rows at the border of the eyelids—the meibomian glands—are so large that they are easily seen with the naked eye when the eyelids are everted. The glands on the genitalia produce copious amounts of sebaceous matter called smegma. Only humans have rich populations of sebaceous glands on the hairless surfaces of the lips; these glands increase in number and size as persons mature. The inside of the cheeks also has many large sebaceous glands, and occasionally there are glands even on the gums and tongue.

It seems highly unlikely, in view of these facts, that sebum is functionless. While its significance is certainly not established, it is possible that it is concerned with subtle chemical communication by smell or taste. Such a function would bring human beings into line with other mammals.

SWEAT GLANDS

Sweat glands are coiled tubes of epidermal origin, though they lie in the dermis. Their secretory cells surround a central space, or lumen, into which the secretion is extruded. There are two distinct types: eccrine glands open by a duct directly onto the skin surface; apocrine glands usually develop in association with hair follicles and open into them.

Most other mammals have numerous apocrine glands in the hairy skin; eccrine glands are usually absent from the hairy skin and limited to friction surfaces. In nonhuman primates there is a tendency for the number of eccrine sweat glands over the body to increase in progressively advanced animals at the same time that the number of apocrine glands becomes reduced. Prosimians have only apocrine glands in the hairy skin; eccrine glands begin to appear in some of the higher forms. The great apes either have equal numbers or have more eccrine than apocrine glands. Humans have the most eccrine glands, with apocrine glands restricted to specific areas.

Strictly speaking, apocrine glands have nothing to do with sweating. They appear late in fetal development (five to 5½ months) nearly everywhere on the body. Most of these rudiments disappear within a few weeks except in the external ear canals, in the axillae, on the nipples of the breasts, around the navel, and on the anogenital surfaces; single glands may be found anywhere. From this, one might speculate that the ancestors of humans had apocrine glands widely distributed over the body, and the embryonic rudiments may be reminders of the history of a once widespread organ system.

Where they appear, the apocrine glands are large and numerous. In the axilla they are so large that the coils press upon each other, forming adhesions and cross-shunts of such complexity that the glands are more spongy than tubular. The complex of these large apocrine glands commingled with an equal number of eccrine sweat glands in the axilla composes what is known as the axillary organ, one of the most characteristic features of human skin. Other than humans, only chimpanzees and gorillas have axillary organs. In spite of their large size, apocrine glands secrete only small amounts of a milky, viscid fluid—pale gray, whitish, yellow, or reddish—which contributes very little to axillary sweat. If eccrine glands were not there, the axillae would be relatively dry.

Axillary secretion

The odour of axillary secretion becomes more intense as it is decomposed by bacteria. Although axillary odours frequently seem unpleasant, they are not invariably so. The odour of individual human beings comes mostly from apocrine secretion, with some contribution from sebum. Since the body odours of all other animals have a social or sexual significance, it can be assumed that this is the archetypal purpose of apocrine secretion, even in humans. The view that the axillary organs are scent glands is supported by the finding that androsterones—the compounds that are responsible for the odour of the boar to which the sow responds—also occur in human axillary secretions.

Humans have 2,000,000 to 5,000,000 eccrine sweat glands, with an average distribution of 150 to 340 per square centimetre. They are most numerous on the palms and soles and then, in decreasing order, on the head, trunk, and extremities. Some individuals have more glands than others, but there is no difference in number between men and women.

The specific function of sweat glands is to secrete water upon the surface so that it can cool the skin when it evaporates. The purpose of the glands on the palms and soles, however, is to keep these surfaces damp, to prevent flaking or hardening of the horny layer, and thus to maintain tactile sensibility. A dry hand does not grip well and is minimally sensitive.

The eccrine glands, then, can be divided into those that respond to thermal stimulation, the function of which is thermoregulation, and those that respond to psychological stimuli and keep friction surfaces moist. This makes a clear-cut distinction between the glands on the hairy surfaces and those on the palms and soles. In addition to thermal and psychological sweating, some individuals sweat on the face and forehead in response to certain chemical substances.

The glands on the palms and soles develop at about 3½ months of gestation, whereas those in the hairy skin are the last skin organs to take shape, appearing at five to 5½ months, when all the other structures are already formed. This separation of events over time may represent a fundamental difference in the evolutionary history of the two types of glands. Those on palms and soles, which appear first and are present in all but the hooved mammals, may be more ancient; those in the hairy skin, which respond to thermal stimuli, may be more recent organs.

The sweat glands in the hairy skin of subhuman primates possibly function subliminally, although they are structurally similar to those of humans. The skin of monkeys and apes remains dry even in a hot environment. Profuse thermal sweating in humans, then, seems to be a new function. Eccrine sweat glands respond to a variety of drugs with different properties. They often respond differently in different individuals under nearly identical conditions and sometimes even respond inconsistently in the same individual. Notwithstanding these apparent vagaries, the eccrine glands function continuously, although their secretion may be imperceptible. Sweating is essential for keeping the human body from becoming overheated.

Essential function of sweating

NAILS

A major characteristic of primates is that their fingers and toes terminate in nails rather than in claws. One can speculate that the development of nails into flattened plates reflects the discontinuation of their use for digging or for defending and attacking. In a broad sense, nails are analogous to hair, having similar composition (keratin) and some common structural features. Even their genesis and mode of growth are comparable, but not identical, to those of hair.

Although apparently simple structures, nails are formed by complex and still poorly understood structural entities referred to as nail organs. Unlike hair, nails grow continuously, with no normal periods of rest; if their free edges were protected from wear, they would extend to prodigious lengths, growing in a twisted fashion like a ram's horns. Nails grow about 0.1 millimetre per day, or roughly one-third as rapidly as hair. Growth is somewhat slower in winter than in summer and slower in infants and old people than in vigorous young adults. It requires about three months for a whole nail to replace itself.

A number of factors can alter normal nail growth, among them age, trauma, poisons, and organic disorders. Habitual nail biting speeds up growth, and certain occupational practices can cause an increase in thickness. The nail-forming organ is particularly sensitive to physiological changes. During stressful periods or prolonged fever, or in response to noxious drugs, nails may become cracked, thinner, thicker, furrowed, or otherwise deformed, or they may be shed. Such sensitivity of response should make nails relatively good indexes of the health of individuals. But because of their ready response to so many internal and external factors, and because changes in them often occur without a known reason, signs of abnormality can be misleading or difficult to interpret. Like hair, the visible part of the nail plate is a dead structure. Defects

inflicted upon it by mechanical means that do not disturb the underlying living tissue are eventually cast off at the free border.

Nails have a root, buried beneath the skin; a plate that is firmly attached to a nail bed underneath; and a free edge. Depending upon its thickness and the quality of its surface, the nail plate may be pink or whitish; the nail itself is translucent and colourless, allowing the colour of the blood in the superficial capillaries of the nail bed to show through. At its base the nail plate may have a whitish, arched marking called a lunule. Always present on thumbnails, lunules may be present or absent on the other fingers and are nearly always absent on the little finger. There are variations in different individuals and even between the two hands of the same person; such variations are probably controlled by genetic factors.

The nail itself consists of firmly cemented keratinized cells, flattened horizontally to the surface. Whereas the surface of nail plates may appear to be smooth, it is lined by parallel, longitudinal furrows, more strongly etched in some persons than in others and typically more prominent in the aged. These markings have some correspondence to the more pronounced grooves and ridges on the undersurface of the plate.

Nail growth

Nails grow from a matrix at the base of the nail root. During the early part of their journey, matrix cells multiply and move forward, synthesizing keratin, underneath the fold of skin (eponychium) at the base of the nail. Once exposed to the surface, the nail is fully formed. The nail plate seems to glide over the nail bed, but it is firmly attached to it; the entire tissue, nail bed and plate, most likely moves forward as a unit. The nail bed has often been called sterile matrix, since it adds little or nothing to the nail plate. Yet under certain pathologic conditions, it assumes keratinizing activities that result in a variably thickened or deformed nail plate.

Although it is less effective than claws for digging or gouging, the flattened nail is still an excellent adaptation that has added much to the development of manipulative skills. Nails not only protect the tips of fingers but also give them firmness and the ability to pick up or make contact with minute objects. Claws would be useless for such functions.

CUTANEOUS SENSE ORGANS

The skin has both free nerve endings and so-called corpuscular endings, which include nonnervous elements. The corpuscular endings are further differentiated as encapsulated or nonencapsulated receptors.

Free nerve endings occur in the epidermis, in the superficial dermis, where they are arranged in tufts, and in hair follicles. Merkel cells, which are found in the basal layer of the epidermis, are an example of nonencapsulated corpuscular receptors. The most striking example of an encapsulated receptor is the Pacinian corpuscle, an ovoid structure that is about one millimetre in length and lamellated in section, like an onion; these receptors can be found deep in the dermis. Various other dermal sense organs—for example, Golgi-Mazzoni corpuscles, Krause end bulbs, Meissner corpuscles, and Ruffini endings—have also been described.

It can easily be demonstrated that touch, cold, warmth, and pain are each perceived in separate points on the skin surface. The various end organs were at one time, therefore, somewhat arbitrarily assigned as monitors of one or another of these qualities. A difficulty was that many of the receptors are present only in glabrous skin, even though hairy skin in similarly perceptive. These earlier ideas were undoubtedly too simple, but electrophysiologists have confirmed the view that the various end organs respond to specific stimuli. The functional existence of mechanoreceptors, thermoreceptors, and pain receptors has been established, though only some of these can be identified with classical end organs. The Merkel cells and Ruffini endings, for example, are "slowly adapting" mechanoreceptors; while the Meissner, Pacinian, and Golgi-Mazzoni corpuscles and the hair follicle receptors are "rapidly adapting" mechanoreceptors.

(W.M./F.J.G.E.)

Skin diseases

GENERAL FEATURES

Although most diseases affecting the skin originate in the layers of the skin, such abnormalities are also important factors in the diagnosis of a variety of internal diseases. There is some truth in the belief that the skin mirrors a person's internal health. Often, the visibility and accessibility of skin make it the first organ of the body to show detectable signs of underlying disease. Abnormalities of the skin frequently suggest metabolic, malignant, and glandular diseases.

A mirror to health

Like other tissues, skin is afflicted by all types of pathological changes, including hereditary, inflammatory, benign and malignant (neoplastic), endocrine, hormonal, traumatic, and degenerative processes. Emotions affect the health of the skin as well. The reaction of the skin to these diseases and disorders differs from that of other tissues in many ways. For example, extensive inflammation of the skin may affect metabolism within other organs and systems of the body, causing anemia, circulatory collapse, disorders of body temperature, and disturbance of water and electrolyte balance in the blood. The skin has such vigorous healing properties, however, that widespread injury, as in thermal burns, may be followed by a marked degree of regrowth of the injured or diseased areas, with a disproportionally small degree of scarring.

Diagnosis. The skin has an inherent region-specific anatomical diversity that may profoundly modify the appearance of a rash. This is apparent when skin transplanted from one area of the body to another (other than a symmetrically opposite area) retains the morphological characteristics of the donor area. Thus the morphology of eczema or lichen planus on the palms and soles may bear little or no resemblance to the same disease in the same individual on the face or scalp. In these instances a biopsy shows the abnormalities of the cells of the skin and the pattern and distribution of any invading cells. The ability to identify immunoreactants (immunoglobulins, or antibodies, that react with specific invading agents, or antigens) in skin biopsies has greatly increased the accuracy of the diagnosis of inflammatory disorders and has clarified their immunologic basis, especially in the blistering disorders.

The classification of hereditary skin disorders generally has been based upon gross morphological, histological, and electron microscopic findings; however, because a skin disease may not always have a characteristic presentation, the specific diagnosis sometimes has been in doubt. Better understanding of the biochemical defects underlying hereditary skin disorders now allows these conditions to be diagnosed with more precision. One subset of the ichthyoses, a group of sometimes disabling genetic skin disorders, may thus be delineated from other members of the group, based on biochemical detection of a specific enzyme defect (reduced steroid sulfatase enzyme).

Distribution. The distribution of a rash depends on factors both intrinsic and extrinsic to the body. Mechanical factors (such as trauma, environmental agents, fungal or viral infections, and drugs) are among the most common extrinsic determinants of distribution. Environmental influences, such as sunburn and light-sensitive, drug-induced reactions, may also play a major role. Psoriasis and the rare hereditary blistering disorders collectively called epidermolysis bullosa owe their distributions to local trauma; lesions that show a predilection for the elbows, knees, and lower back are common in psoriasis, and those found in the hands, feet, knees, and mouth of children are indicative of epidermolysis bullosa. A lesion of an eruption that subsequently develops where a mechanical or other physical trauma was applied is termed an isomorphic reaction. Skin diseases in which isomorphic reactions are characteristic include lupus erythematosus (sunlight), psoriasis, lichen planus, and viral warts (mechanical trauma).

Factors

Intrinsic rather than extrinsic influences explain the characteristic systematic pattern of a number of rashes. Some rashes with a symmetrical and segmental distribution may owe their pattern to the segmental arrangement of the embryo. In other rashes, it is the distribution of the cutaneous nerve supply; for example, the pattern of the rash

of herpes zoster (shingles) is determined by the cutaneous distribution of the infected sensory nerve dorsal root ganglion. The blood supply to the skin has a characteristic anatomical distribution that influences the pattern of certain skin eruptions in which cutaneous vascular narrowing or blood stasis is a major feature.

Appearance. Skin diseases in which there is an overproduction of epidermal cells or a disorganization of their differentiation (development of their mature form and function) often show scaling. Simple benign hyperplasia (overgrowth) of the epidermis such as is commonly seen in infantile eczema often appears as lichenification, a term used to describe a thickening of the epidermis in which the normal surface markings of the skin are greatly exaggerated. Chronic benign or malignant proliferative dermatoses involving the epidermis often have a rough warty surface caused by overproduction by the epidermal cells of keratin.

Scaling

There are relatively few skin diseases in which inflammation, including responses to physical injury (such as sunburn), allergy, and infections (such as boils and cold sores), does not play a part. Even cancerous lesions of the skin frequently show some degree of inflammatory response. Inflammation is usually the result of the release of chemical mediators in the skin. These mediators include histamine, which is stored in a preformed state in the skin, and peptides (kinins) and fatty acids (prostaglandins and leukotrienes), which are formed enzymatically in response to injury. In some skin diseases inflammation is the major factor in the morphological appearance of the rash (for example, acute allergic contact dermatitis).

The most important features of inflammation include redness, caused by dilation of blood vessels; heat, caused by increased blood flow; swelling, caused by leakage of fluid from the small blood vessels; and pain or itching. In addition, the affected skin may feel indurated (hardened) because of the deposition of the coagulation protein fibrin and the infiltration by inflammatory blood cells (lymphocytes, histiocytes, and polymorphonuclear leukocytes). Medications designed to counteract inflammation in the skin may antagonize the effects of mediators (*e.g.*, antihistamines). Nonsteroidal anti-inflammatory drugs of the type effective in arthritis, including aspirin and indomethacin, have proved to be of little or no value in the treatment of inflammatory skin diseases.

Whealing reactions (hives, nettle rash, urticaria) constitute a special form of inflammation in which vascular changes predominate, with little or no inflammatory cellular reaction. Localized fluid accumulation (edema) causes the development of a short-lived wheal associated with intense itching. Its appearance frequently indicates an underlying allergic cause. Intense inflammation may occasionally lead to the formation of blisters, caused by the effects of enzymes released from inflammatory cells, from the resident cells of the skin, or from blood plasma components. These enzymes cause the breakdown of proteins responsible for the structural integrity of the skin. In some blistering (bullous) skin diseases (such as pemphigus), the development of large blisters is the predominant morphological feature.

Inflammatory diseases of the dermis often evoke a secondary proliferative and scaly response in the epidermis. Dermal inflammatory disorders may be acute, as in hemolytic (blood-cell destroying) streptococcal infections leading to cellulitis (diffuse inflammation of the connective tissues), or chronic and granulomatous, as in chronic cutaneous tuberculosis (lupus vulgaris). In the first instance the changes of acute inflammation discussed above are present but there is normally no epidermal change. In the second, the skin is hardened and thickened, and it has a brownish appearance under the microscope.

Pigmentation

The intrinsic colour of normal skin is pale yellow, but this colour is modified by pigments—both by melanin and by hemoglobin. The intensity of the brown hue conferred by melanin depends on the distribution of the melanin granules within the cells. All races have the same number of melanocytes, although melanin is produced and distributed through the epidermis more efficiently in blacks. Increased melanin pigmentation is a common reaction to prolonged inflammation of the skin. The intensity and shade of pink depend on the state of the cutaneous circulation. When blood vessels are dilated with increased blood flow, as in an inflammatory disorder, the skin is bright red. The vasodilation associated with stagnant blood flow, as in excessive cooling of the skin, gives the skin a purplish hue.

The predominant symptom of skin disease is itching. Although readily distinguishable from pain, itching appears to be transmitted along the same sensory neural pathways. There are no specialized sensory receptors for itching in the skin; both itching and pain are transduced by free nerve endings located at the dermo-epidermal junction. Although an itching sensation may result in healthy skin from a direct conversion of very light pressure (light stroking of the skin of the lips is a good example), in most skin diseases itching is caused by the release of pharmacological mediators, such as histamine. When itching occurs in normal-appearing skin, as is the case in liver or kidney disease, the itching is probably due to an abnormal accumulation of metabolic products in the skin.

HEREDITARY SKIN DISEASES

The formation of almost all components of the skin (for example, hair texture and colour and skin pigmentation and thickness) is under genetic control. A large number of common skin diseases also are directly or indirectly determined by a person's genotype (genetic constitution), but their expression may require an external influence or an altered hormonal milieu. The hereditary diseases psoriasis and atopic eczema are examples of skin disorders in which sunlight (as an extrinsic factor) or stress (as an intrinsic factor) activate the condition. Even when heredity has an immediate determining role, other factors may influence the expression of disease. The hereditary blood vessel diseases of the skin, for example, many of which are under direct genetic control, sometimes do not become evident until the hormonal changes of puberty create conditions optimal for disease expression.

Nevi

A common genetic abnormality is the nevus, often called a mole or birthmark. Nevi are due to primary abnormalities in the structure or number of skin cells. A local increase in the concentration of melanocytes is termed a melanocytic or pigmented nevus; an area of increased capillary concentration, a capillary nevus. In nevus anemicus, an area of skin is pale because of reduced blood flow, even though there is no evidence of a structural disorder. Although most nevi are benign, malignant transformation may cause melanocytic nevi.

If an abnormality is inherited as an autosomal dominant trait and one parent is affected, then, from the laws of Mendelian inheritance, each child has a 50 percent risk for the disease. Amniocentesis, often performed between the 12th and 20th weeks of pregnancy, entails little risk to the mother or fetus and makes it possible to diagnose prenatally severe disorders such as epidermolysis bullosa, a blistering and scarring disorder; albinism, or a generalized lack of pigment from the skin; and xeroderma pigmentosum, a precancerous condition in which sunlight-induced skin damage cannot be enzymatically corrected by the affected skin cells.

GENERALIZED SKIN DISEASES

Eczema and dermatitis. The terms eczema and dermatitis are often used interchangeably to denote an inflammatory process in the skin that involves the upper dermis and epidermis. The epidermis exhibits swelling of the keratinocytes and accumulation of fluid between them (spongiosis). In the severe form of spongiosis, blisters form within the epidermis. Childhood eczema in black children frequently is seen as a follicular eruption of pinhead-sized papules. In chronic forms of eczema or dermatitis the prominent changes are thickening of the epidermis and marked hyperkeratosis (thickening of the outer horny layer of the epidermis). These changes lead to lichenification (see above), roughening and scaling of the skin surface, and itching. The function of the horny layer as an impervious barrier may be seriously impaired, with two important consequences: loss of water from the skin leads

to desiccation of the horny layer, which in turn leads to cracking, increased scaling, and soreness; and loss of the barrier function causes increased absorption of medications applied to the surface of the skin. For example, the enhanced absorption of topically applied corticosteroids may cause toxic changes in the skin and in distant organs and tissues.

Types of dermatitis

Dermatitis is classified into several different types, including contact dermatitis, atopic dermatitis, and seborrheic dermatitis. Contact dermatitis may be further classified as allergic or nonallergic. Nonallergic contact dermatitis occurs in response to skin irritants, is common in industrial settings (occupational dermatitis), and usually affects the hands. Acids, alkalis, dirty oils, detergents, solvents, and even repeated exposure to water are among the numerous causes. Genetic predisposition undoubtedly plays a part, however, which explains why some workers, but not others, suffer from occupational dermatitis, despite equal exposure.

Allergic contact dermatitis is a less common cause of occupational skin disease but is frequently found among the general population. Like primary irritant contact dermatitis, it can be produced and studied under controlled conditions, and therefore more is known about the underlying pathogenic mechanisms. Allergic contact dermatitis occurs only in persons who have, after previous exposure, become sensitized to the offending antigen. Common antigens include nickel, chromium, rubber, and paraphenyline diamine. The dendritic Langerhans cells of the epidermis are responsible for recognizing the antigen, which is usually bound to a homologous protein or polypeptide. This information, conveyed to the local lymph glands via the lymphatic vessels, leads to the formation of specifically sensitized lymphocytes. On subsequent exposure to the offending antigen, these and other secondarily recruited cells release chemical mediators (lymphokines) that evoke dermatitis in the exposed skin. Sensitization usually persists, although it may decline with old age.

Allergic contact dermatitis is usually suspected from the distribution of the rash, which corresponds to the areas of skin exposed to the antigen. For example, red-tipped match contact allergic dermatitis is caused by phosphorous sesquisulfide released as a vapour following the striking of a match. In pipe and cigarette smokers who use this type of match, the area affected is the side of the face, explained by the smoker's habit of cupping the match and matchbox in the hands during the process of striking. The diagnosis of allergic contact dermatitis is confirmed by a patch test, in which minute concentrations of suspect antigens are placed on areas of the skin. An accurate diagnosis of allergic contact dermatitis is important because the condition can be cured by avoiding contact with the offending substance.

Atopic dermatitis, which affects a small number of infants, is one of several genetically linked disorders that include asthma, hay fever, and urticaria. This group of atopic diseases is characterized by sensitization of the skin to a wide range of common antigens. In addition to eczematous changes, persons with atopic dermatitis may also often have diminished, absent, or paradoxical cutaneous vascular reactions to vasodilating and vasoconstricting drugs; impaired immunity to fungal and viral infections; and cataracts in the lenses of the eyes. Although there is some controversy over the causes of atopic dermatitis, the role of genetics is undisputed. About 70 percent of patients have a family history of eczema, asthma, or hay fever. Most patients begin to suffer symptoms during the first six months of life, and there is evidence that ingestion of dairy products during this period may precipitate development of atopic dermatitis in a predisposed infant. There is no real evidence that food plays a significant role in the development of atopic dermatitis in older children and adults. The disease usually improves during the second decade of life, although it sometimes persists in adults and may even appear then for the first time.

Seborrheic dermatitis

Seborrheic dermatitis is a less common form of chronic dermatitis that characteristically affects the scalp and other hairy areas, the face, and flexural areas (groin, armpits, skin behind the ears, the cleft between the buttocks). It is frequently associated with infection of the skin by yeasts and bacteria, but a causal relationship with these organisms has not been established. There is no evidence of increased greasiness of the skin (seborrhea). Dandruff is a mild form of seborrheic dermatitis that affects most people at some time during their lives. The scaliness of the scalp associated with dandruff is caused by the intermittent shedding of dead stratum corneum cells, which in health are shed continuously.

Psoriasis. Although the mechanisms of inheritance are not clear, historically psoriasis, like atopic dermatitis, has been thought to be inherited as an autosomal dominant trait that pursues a chronically remitting and relapsing course. Psoriasis, however, is less common than atopic dermatitis, affecting about 2 percent of the population, and is both a proliferative and an inflammatory disorder. The most important feature of psoriasis is an accelerated proliferation of the keratinocytes, which results in the formation of raised scaly plaques in areas of injury, notably the knees, elbows, buttocks, and knuckles. Streptococcal tonsillitis frequently precedes relapses, especially in children, and the mechanism in these cases is probably immunologic. The isomorphic reaction (see above) is a feature of psoriasis. Persons with the disease may also have a characteristic form of arthritis that affects joints in the fingers and spine. Whether the increased rate of keratinocyte proliferation is due to increased activity of a growth-promoting factor or to lack of a growth inhibitor is unknown.

As an inflammatory disorder, psoriasis is characterized by nests of neutrophil leukocytes in the epidermis called microabscesses. Possible mediators that attract this cell type are leukotrienes and fragments derived from activated blood peptide components called complement. The skin blood vessels are also abnormal in psoriasis, with increased twisting. The relationship of this dermal change to the epidermal abnormalities is not clear. Although psoriasis is neither infectious nor contagious it is disfiguring.

Skin cancer. Skin cancer is one of the most common forms of cancer in humans and, although visible and therefore recognizable at an early stage, it results in significant mortality. The incidence and prevalence of skin cancer can be greatly reduced by simple preventive measures, such as avoidance of exposure to the Sun and to excessive ionizing radiation.

Primary skin cancers can be divided into two types: epidermal cancers, which originate in keratinocytes, melanocytes, or skin appendages (*e.g.*, sweat glands, the pilosebaceous apparatus); and dermal cancers, which originate in neural, vascular, mesenchymal, or lymphoreticular tissues. Malignant tumours arising from keratinocytes or melanocytes are the most frequent skin cancers.

Basal cell carcinoma

Basal cell carcinoma, rare in Negroes and Asians, is the most common malignant skin tumour in Caucasians. It arises from the undifferentiated basal keratinocytes of the epidermis. Although multiple basal cell carcinomas may develop early in life as an inherited trait (nevoid basal cell carcinoma syndrome, or as a complication of xeroderma pigmentosum), most arise in middle age and later. These cancers rarely metastasize but may be highly invasive locally; they are then known as rodent ulcers. The lesions occur in fair-skinned persons and on areas of skin that receive the greatest exposure to sunlight. Treatment with inorganic arsenical drugs and exposure to ionizing radiation (X rays, radium) may also contribute to some cases. Avoidance of unnecessary sunlight and careful control of ionizing radiation significantly lowers the incidence of basal cell carcinoma. Although metastases are rare, the cancer may spread locally and invade surrounding tissues. When this occurs, treatment may be difficult and lengthy.

Squamous cell carcinoma is less common than basal cell carcinoma but has a higher rate of metastasis. It is common in children with xeroderma pigmentosum, who are unable to repair DNA damage caused by ultraviolet irradiation. In most persons this inability is due to the deficiency of an endonuclease enzyme. Incomplete repair of damaged DNA causes mutations that appear as basal or squamous cell carcinomas, malignant melanomas, and keratoacanthomas.

In adults, squamous cell carcinoma rarely occurs in the absence of an external cause. Protracted exposure to sunlight is the usual cause, but chronic scarring from burns, as well as reactions to vaccinations, radiation dermatitis, and chronic ulceration, may contribute to some cases. Squamous cell carcinoma is also an occupational hazard, as was noted at the end of the 19th century in regard to chimney sweeps who contracted cancer after exposure to tars. Tar-induced squamous cell carcinoma occurs today in workers who distill tar vapour in the manufacture of coal gas and in machinery operators whose clothes and skin become soaked in mineral oil.

Because of the high rate of metastasis of squamous cell carcinomas, early diagnosis is important, especially in a middle-aged or elderly person with a skin ulcer that fails to heal. Skin lesions that precede the development of squamous cell carcinoma include white patches in the mucous membranes of the mouth, genitalia, or anus; warty lesions called keratoses (which are especially common after chronic exposure to the Sun, when they are called solar keratoses); and the lesions of Bowen's disease—persistent red scaly plaques that on microscopic examination are found to contain grossly abnormal keratinocytes. When squamous cell carcinoma follows Bowen's disease, there is often a history of treatment with an inorganic arsenical drug.

Malignant melanoma

In Western countries, the mortality from the skin cancer malignant melanoma is increasing by about 4 percent per year. This type of skin cancer arises from the melanocytes of the skin, and the tumour is therefore often, but not invariably, pigmented with melanin. In adults malignant melanoma arises either as a new lesion or as a change in a benign pigmented mole. Malignant melanoma metastasizes frequently, and excision of the tumour together with a collar of surrounding healthy skin is curative if done early.

Primary cancers arising in the dermis are much less common than epidermal malignant tumours. Mycosis fungoides is a malignant tumour of the T lymphocytes of the dermis. Despite the name, fungal infection does not cause the cancer. Mycosis fungoides (which is rare in children) is ultimately fatal, but the tumour grows slowly enough that many patients die of unrelated causes.

Extrinsic causes of dermal cancers are rare. An exception is the malignant cutaneous vascular tumour called Kaposi's sarcoma. Although there are many varieties of Kaposi's sarcoma, the development of one particular form is a common terminal event in patients with acquired immune deficiency syndrome, and the tumour may therefore result from infection by a retrovirus known as the human immunodeficiency virus (HIV).

Skin infections and infestations. Healthy skin harbours a resident, harmless bacterial population that includes Micrococcaceae, Propionibacteriaceae, and aerobic diphtheroids. In addition, pathogenic strains of staphylococci may inhabit the skin of healthy persons, who then act as carriers. These strains are often resistant to antibiotics, so that carriers pose a serious health hazard to those with diminished resistance to infection, such as newborn infants and elderly surgical patients. Most infections by resistant organisms occur nosocomially, or in hospitals. Although local measures, such as cleansing of the involved skin, are often effective for minor skin infections, serious infections require antibiotics.

Bacterial infections

Signs of bacterial infection appear after the organisms have been introduced into the dermis. Some organisms, such as staphylococci and streptococci, produce a pustular infection (impetigo, boils) in which the main features are acute inflammation and the accumulation of large numbers of white blood cells as pus. Other organisms, including *Mycobacterium tuberculosis* and *Treponema pallidum* (the spirochete that causes syphilis), may cause more chronic inflammatory changes without pus but involving a characteristic arrangement of mononuclear inflammatory cells called histiocytes. The term granulomatous is used to describe this kind of cutaneous reaction to infection. Leprosy is a cutaneous bacterial granulomatous disease in which the causative organism, *Mycobacterium leprae,* resides in the neural tissue in the skin, producing the characteristic features of nerve thickening and anesthesia. Despite its bacterial origin and popular reputation, leprosy is minimally infectious, and isolation of affected persons is rarely justifiable.

A number of different viral species are epidermotropic; that is, they invade the epidermis, either from the skin surface through a minor abrasion, as in viral warts, or from the bloodstream (viremia), as in varicella (chickenpox), or through the peripheral sensory nerves, as in herpes zoster (shingles). The human wart virus causes epidermal cellular proliferation and hyperkeratosis. When it invades the genital skin and mucous membranes it produces moist, exuberant, highly vascular warts called condylomata acuminata. These warts are transmitted through sexual contact and may become cancerous. By contrast, common viral warts of the hands, feet, and other non-mucosal surfaces are not sexually transmitted and rarely become cancerous. Viral warts on non-mucosal surfaces occur in almost all children, although most disappear spontaneously with the development of acquired immunity.

Viral infections

Herpes simplex (cold sores, fever blisters) and herpes zoster (shingles) are two epidermal viral infections that produce blisters within the epidermis. The severity of these infections is influenced by the state of the person's immune system; they are more severe in injured persons, in the elderly, and in debilitated patients, especially those with cancer of the lymphoid system. Herpes zoster (shingles) and varicella (chickenpox) are caused by the same virus, and patients with shingles have previously had chickenpox. The virus remains dormant in the sensory nerves and is later reactivated, either spontaneously or following the impairment of immune defense mechanisms. The virus, by translocating along the nerve branches, invades the epidermis; the distribution of the ensuing vesicular rash, which is painful, is confined to the cutaneous distribution of the same sensory nerve.

Fungal infection is a common cause of chronic skin eruption. Several species of dermatophyte fungi may live on the dead horny layer of the skin. The inflammatory changes they produce are partly due to irritant products of the fungus, which diffuse into the skin, and partly to an immunologic reaction by the host in an attempt to eliminate the fungus. Yeasts, of which *Candida albicans* is the most common, are also a cause of chronic skin or mucous membrane infection. In patients receiving immunosuppressive drugs for organ transplantation, yeast infections may become systemic, and both yeast and dermatophyte infections may be extensive, persistent, and unresponsive to drug treatment. The availability of systemic antifungal drugs, including griseofulvin and the broad-spectrum imidazole group, has revolutionized the treatment of these disorders, but many patients with both minor and serious fungal infections remain, for various reasons, resistant to drug treatment.

Skin infestations are frequent in persons living or working in overcrowded, unhygienic conditions. Pediculosis (crabs, lice, nits), which affects hairy areas, is diagnosed by identifying the egg capsules (nits) that are cemented to the hair shaft. Lice may also be visible near the base of the hair. Scalp, axillary, or pubic hair may be affected. Scabies, which is caused by the mite *Sarcoptes scabiei,* is usually acquired through close personal contact, especially sexual contact. It affects all areas of the body except the soles of the feet, head, and neck (in infants these areas are also involved) and causes an itchy eruption that is largely an allergic reaction to the products of the mites. Although most persons develop an immunologic reaction to scabies, acquired immunity is poor or nonexistent, and recurring attacks are common. The condition is treated with benzyl benzoate, and all persons who have had contact with the infected individual must be treated, whether symptomatic or not, because of the high rate of reinfestation.

Urticaria (hives). The skin blood vessels can also show acute, short-lived reactions. Urticaria features an area of central redness, on which is superimposed an irregular wheal, caused by local edema, and surrounded by a bright pink flare. The wheals are usually multiple and remain visible for 30 to 45 minutes. The reaction is due to release of histamine from cellular stores within the skin and is

Common causes of hives

usually accompanied by itching. Common causes include allergies to shellfish, strawberries, and nuts or to drugs such as penicillin; but physical factors such as cold, exercise, and sunlight may also produce the response. Since histamine is an important causative agent in urticaria, most cases respond to antihistamine treatment.

LOCALIZED SKIN DISEASES

Pigments. An increase in melanin pigmentation may be caused by an increased density of melanocytes, by abnormal packaging of melanin, or by increased melanin production. Pigmented birthmarks usually reflect local increases in melanocyte numbers, but in certain rare congenital pigmentary disorders, such as von Recklinghausen's neurofibromatosis, there is abnormal packaging of melanin within the melanocytes. Pigment production in the skin is regulated by a pituitary peptide hormone, melanocyte-stimulating hormone, and the increase in melanin pigmentation seen with pituitary tumours may reflect overproduction of this hormone by the pituitary.

Both suntans and postinflammatory pigmentation result from the overproduction of melanin by melanocytes. In suntans melanin overproduction occurs in response to ultraviolet light and protects the skin against sunburn and the carcinogenic actions of ultraviolet rays. Postinflammatory pigmentation follows a number of inflammatory skin disorders, including chronic dermatitis, acne vulgaris, lichen planus, and drug reactions. It is more common in dark-skinned than in fair-skinned persons. In suntans and postinflammatory pigmentation there is usually no increase in the number of melanocytes.

Loss of melanin pigmentation may be caused by the absence of melanocytes, as in vitiligo; reduced or absent synthesis of melanin by melanocytes, as in albinism and phenylketonuria; abnormal packaging of melanin in melanosomes, as in the Chediak-Higashi syndrome; or insufficient transfer of melanosomes to keratinocytes, as in hypopigmentation secondary to atopic dermatitis or psoriasis. Vitiligo is a common disorder of pigmentation that results in the widespread appearance of depigmented patches, notably around the mouth, anogenital area, eyelids, elbows, knees, and backs of the hands, due to an absence of melanocytes in these areas. Vitiligo is usually a slowly progressive disorder that is probably the result of abnormal immunologic reactions against constituents of the body's own cells (autoimmunity).

Certain chemicals found in plants, the psoralens, if taken orally or applied to the skin enhance the ability of ultraviolet light to induce pigmentation. Although rarely successful, treatment with psoralens and ultraviolet radiation has restored pigmentation. Persons with vitiligo should avoid sunlight, which accentuates the contrast between normal and depigmented skin.

Collagen

Connective tissues. Collagen is the major structural protein of skin and is responsible for its tensile strength, elasticity, and pliability. It is synthesized in the dermis by fibroblasts. Scarring is caused by excess production of collagen during healing. When the collagenous proliferation is extensive, a hypertrophic scar, called a keloid, results.

There are at least five types of collagen and numerous inherited disorders of collagen synthesis. The best known of the inherited disorders is the Ehlers-Danlos syndrome, which includes at least seven varieties of connective tissue defects. Most persons with Ehlers-Danlos syndrome have increased joint mobility and fragile skin, and they bruise easily. Depending on the nature of the defect in collagen synthesis, patients may also have spinal deformities, ocular defects, and ruptures of major blood vessels.

The connective tissue of the skin may rupture after stretching, even in healthy persons, forming purple linear discolorations. Because of the defect in the continuity of the dermal connective tissue, these striae are soft to the touch. Initially livid, after several months the striae become ivory white and firm because of the formation of scar tissue. Rapid growth during adolescence may cause striae to form on the thighs, buttocks, lower trunk, or shoulders. During pregnancy striae form on the abdomen as the elastic tissue stretches and weakens. Striae may also follow treatment with corticosteroid drugs.

The skin connective tissue is affected in a group of inflammatory disorders collectively termed the autoimmune connective tissue diseases. These diseases, which include systemic lupus erythematosus, dermatomyositis, polyarteritis nodosa, and systemic sclerosis, are characterized by the involvement of more than one organ or system, the presence in the serum of immunoglobulin autoantibodies that react with human tissue or organ components, and inflammatory damage to blood vessels of the skin and other organs. Systemic sclerosis causes hardening of the skin (scleroderma) due to an increase in the connective tissue of the dermis and marked tethering of the dermal connective tissue to subcutaneous tissues. The process is not confined to the skin, and the cause is not known, although an autoimmune basis is thought to be involved because of the associated immunologic abnormalities. There is evidence that the skin fibroblasts of patients with scleroderma produce more collagen than the fibroblasts of healthy persons. It has also been suggested that the connective tissue changes may be caused by vascular abnormalities. No satisfactory treatment is known, although improvement of the circulation with vasodilating drugs may produce temporary relief.

Blood vessels. Blood circulation in the skin serves both to supply nutrients and as a means of regulating temperature. The structural and functional organization of the skin's vascular system is complex, and there are major regional variations. Skin blood vessel disorders may arise from an increase in the viscosity of blood contents or from abnormalities of the vessels themselves. Increased viscosity may be due to an overabundance of red blood cells (polycythemia) or an increase in one or more fractions of the protein content of the plasma; hyperglobulinemia is an example of the latter. In functional disorders of the blood vessels, nervous control of the vessels may be disturbed, leading to vascular spasm or dilatation (flushing). In Raynaud's phenomenon, the vascular spasm is severe, affecting the extremities and leading to attacks of cold, white fingers. Milder degrees of spasm, as well as increases in blood viscosity, may cause a purple discoloration of the skin (cyanosis), which may assume a reticulate pattern (livedo reticularis).

Etiology

Damage to the vessel walls themselves may perturb skin circulation. Causes include atherosclerosis resulting from age-related degenerative changes, disturbance of lipid (fat) metabolism, and inflammation of the vessel wall. Inflammation is characteristic of the autoimmune connective tissue diseases (see above). When disease of the blood vessels is severe, lack of blood in the affected areas may cause skin ulcers. Chilblains (pernio) occur when the skin is persistently chilled, leading to protracted vessel spasm, cyanosis, coldness, and ultimately ulceration of the affected area.

Hair and pilosebaceous follicles. Hair is of little functional value in humans. Many systemic diseases, however, alter the appearance and growth patterns of hair. In diseases of malnutrition, such as kwashiorkor, the scalp hair becomes reddish brown, sparse, and brittle. Acute protein starvation may lead to the cessation of hair growth and constriction of the hair shafts. Chronic iron deficiency may cause diffuse hair loss (alopecia). Since the continuous hair growth of healthy persons involves the cells of the hair root in exuberant mitotic activity, hair is especially vulnerable to the cytotoxic action of drugs. Acute hair loss is an early side effect of the cytotoxic drugs used in cancer chemotherapy.

Kwashiorkor

Acute diffuse hair loss (telogen effluvium) may also occur after severe fevers and childbirth. It appears to be caused by an abrupt reduction in the proportion of actively growing hairs, and after a few months the hair regrows normally without treatment. The hair follicles may become inflamed, forming small pustules, which, in severe cases, result in bald, scarred areas with destroyed follicles.

Both pattern baldness in men and hirsutism in women are under the control of the steroid sex hormones. Male pattern baldness occurs to some extent in almost all men. It may start any time after puberty and is distinguished from other forms of alopecia by the pattern of frontal and temporal hair recession and thinning over the crown. On the affected areas of the scalp a downy hair called vellus

replaces the former long, sturdy, pigmented terminal hair. Almost as common is the reverse process, in which hair that was formerly vellus alters to the long pigmented terminal type and becomes more visible. When this occurs in the areas of the female body sensitive to male sex hormones, and especially when it follows the adult male pattern of hair distribution, it is known as hirsutism. Some affected women have endocrinologic abnormalities, including ovarian cystic disease or overactive adrenal glands, that elevate the levels of androgen (male sex hormones). These women may also have menstrual irregularities or be infertile. Hormonal treatment or, in less extensively hirsute women, electrolytic depilation may be helpful.

Among persons with curly hair, and especially among blacks, cut ends of hair may curve and bury themselves in the skin, causing acute pustular reactions. These ingrowing hairs, or pili incarnati, affect 70 percent of black men who are clean-shaven.

Acne. The hair follicle from which the hair shaft emerges serves as a duct for conducting sebum, the oily product of the sebaceous gland, to the skin surface. The hair follicle and sebaceous gland together form the pilosebaceous unit. Sebum secretion is under hormonal control and increases substantially at puberty, then continues at an increased rate into advanced adult life. An important disease of the pilosebaceous unit is acne vulgaris, which occurs on the face, shoulders, and upper torso of adolescents. The hallmark of acne vulgaris is the comedo, commonly known as the blackhead. The comedo consists of a mixture of keratinous debris and sebum that lodges in the mouth of the pilosebaceous duct. Affected areas of the skin also exhibit inflamed papules, pustules, and cysts. The inflammation is caused by leakage of sebum and keratin debris outside the distended pilosebaceous duct. The bacillus *Propionibacterium acnes,* which populates the lesions, may also contribute indirectly to the inflammation by metabolizing the sebum to produce irritant fatty acids.

Acne vulgaris

Eruptions can be brought on by contact with dirty oils or tars and halogenated hydrocarbons. Chloracne, caused by exposure to chlorine compounds, may develop in workers involved in the manufacture of cables, insecticides, fungicides, and herbicides. Industrial accidents involving the release of tetrachlorodibenzodioxin vapour have caused severe community outbreaks of chloracne. Acne rosacea, an acne of middle age that exhibits nodular swelling of tissues in the face, is a disease that occurs rarely among blacks.

Sweat glands. Most sweat glands in humans are eccrine (*i.e.,* they secrete outwardly) and are under the control of the autonomic nervous system. Eccrine sweat glands are distributed uniformly over the body and function mainly in controlling temperature, although on the hands and feet they assist grip. Inability to secrete adequate sweat (anhidrosis) is usually a congenital abnormality. Hyperhidrosis is the abnormal excessive secretion of sweat, usually on the palms and soles and in the axillae. Emotional stress normally stimulates sweating in these areas, but in hyperhidrosis excessive sweating may follow minor stimuli or no obvious stimulus at all. Hyperhidrosis may be an occupational problem for those who handle books, papers, or metallic objects that can rust.

Most serious disorders of the eccrine sweat glands result from the retention of sweat. Prickly heat (miliaria) occurs in persons living in a hot, humid environment and is due to sweat retention brought on by the combination of increased sweat production and blockage of the duct from hydration of the surrounding horny layer of skin. The gland continues to secrete, and the pressure may rupture the sweat duct at its junction with the epidermis. In heatstroke, another disorder of the eccrine sweat glands, the glands are exhausted and unable to secrete sweat. Hyperpyrexia, an excessive rise in body temperature, may ensue and is sometimes fatal. The most common disorder of the eccrine glands, however, is excessive sweating of the palms, soles, armpits, and face. Emotional stress normally stimulates sweating in these areas, but emotional hyperhidrosis (increased sweating) occurs when excessive sweating follows minor or no obvious stimuli.

Apocrine sweat glands, which are normally limited to the armpits, perineum, and genitalia, differ anatomically and physiologically from eccrine sweat glands. In the apocrine glands the secretions are concentrated at the free end of the cell, which is cast off along with the secretory products. Apocrine secretion is oily, opalescent, and odourless when secreted, but acquires an offensive smell when it mingles with the bacteria of the skin. The actions of deodorants and antiperspirants are based mainly on an antibacterial effect, on mechanical blockade of the apocrine gland pores, or both.

Nails. The most common causes of diseased nails are chronic paronychia—infection of the soft tissue of the fingers between the lateral edges of the nail plates and the lateral nail folds—fungus infections, trauma, psoriasis, impaired circulation, and dermatitis. Chronic paronychia is an occupational hazard of dishwashers, bar workers, hairdressers, and vegetable preparers. When the cuticle is lost, water, bacteria, and yeasts seep under the nail fold and cause inflammation. The lateral nail folds extrude beads of pus when they are squeezed. The infection is commonly caused by the yeast *Candida albicans,* and when this organism is identified, infection elsewhere in the body should be sought.

Like the hair, the nails may be useful indicators of skin diseases and of internal disorders. Psoriasis causes pitting on the surface of the nails and an increase in their growth rate. Lichen planus causes a characteristic scarring of the posterior nail fold, leading to fusion of the posterior nail fold with the nail plate. Club-shaped deformity of the nail plate is associated with chronic chest disease and cyanotic congenital heart disease. A reduction in the rate of nail growth, a yellow discoloration of the nail plate, and a smooth curvature to the nails are all features of the yellow nails syndrome, a rare hereditary condition in which the lymphatic drainage system of the body is defective. In chronic anemia the nail plate may become concave (koilonychia). Transverse depressions of the nail plates called Beau's lines result from temporary interruption of nail growth and are seen in a number of acute illnesses, including coronary thrombosis, pneumonia, and severe injuries. Various forms of interference with the blood supply may affect the nail matrix and produce irregular growth of the nails. Drug treatment may cause shedding of the nails. Photo-onycholysis—destruction of the nails from exposure to sunlight—occurs in patients taking photosensitizing drugs such as certain broad-spectrum antibiotics.

METABOLIC DISORDERS

Metabolic disorders may result in the deposition of excess materials in the dermis and thus cause disease. Uric-acid salts are deposited in gout, calcium salts in calcinosis, lipids in xanthomatosis, an abnormal protein in amyloidosis, mucin in myxedema, and porphyrins in porphyria. The appearances of the skin are usually sufficiently characteristic to be diagnostic.

A metabolic disorder that dramatically affects the skin is porphyria, an enzyme defect that causes hemoglobin precursors, or porphyrins, to accumulate in the skin. Porphyrins are abnormally sensitive to long wavelength ultraviolet light. The accumulation of porphyrins results in severe sun sensitivity, with blistering, increased skin fragility, and scarring on the areas exposed to light. The most common form, porphyria cutanea tarda, is usually caused by a combination of a hereditary enzyme defect and chronic alcoholism.

AGING AND THE SKIN

More than one-third of persons over 65 years of age have skin problems. Prevalent in the elderly are such common skin conditions as fungal infections, excessive dryness, various benign tumours, seborrheic dermatitis, seborrheic warts, solar keratoses, and hirsutism. Many age-related skin disorders previously viewed as inevitable accompaniments of advancing age are now known to be remediable.

Skin may age much faster than the rest of the body because of environmental effects, especially sunlight. A fair-skinned woman who habitually sunbathes, for example, may have a senile skin at age 40, whereas her coeval who spends most of her time indoors may not. The predominant features of the aging skin include skin laxity lead-

ing to wrinkles, dryness, itching, increased pigmentation, and telangiectasia (*i.e.,* visible dilation of the skin blood vessels). Moles, large seborrheic warts, and small hemangiomas (de Morgan's spots) are more common among whites, while small black warts (dermatosis papulosa nigra) of the upper chest and face and tiny white spots (idiopathic guttate hypomelanosis) are more common among blacks.

There is no evidence that the rate of reproduction of epidermal cells slows with age. There are, however, marked degenerative changes in the connective tissue, both elastic and collagenous, that cause wrinkling and loss of elasticity. So-called rejuvenating cosmetic skin preparations are ineffective, although simple emollients may provide temporary suppleness. Avoiding excessive exposure to sunlight and using sunblocks, if adopted early in life, should retard the aging process and reduce the incidence of skin cancers. None of these treatments, however, does more than slow the onset of senescence in the skin, and those who survive into the 10th decade of life have age-related changes that are largely independent of their earlier life-styles.

(M.W.G.)

BIBLIOGRAPHY

Animals: J. BEREITER-HAHN, A.G. MATOLTSY, and K. SYLVIA RICHARDS (eds.), *Biology of the Integument,* 2 vol. (1984–86), is a comprehensive account of all aspects of the integument of animals and a valuable reference source. Information on integumentary systems of different animals can be found in the appropriate sections of such comprehensive texts as ROBERT D. BARNES, *Invertebrate Zoology,* 5th ed. (1987); WILLIAM E. DUELLMAN and LINDA TRUEB, *Biology of Amphibians* (1986); and JOEL CARL WELTY, *The Life of Birds,* 3rd ed. (1982). WILLIAM MONTAGNA, "Cutaneous Comparative Biology," *Archives of Dermatology,* 104(6):577–591 (December 1971), is an advanced but readable article on the biologic properties and adaptive structure and function of mammalian skin. See also R.I.C. SPEARMAN and P.A. RILEY (eds.), *The Skin of Vertebrates* (1980).

Human skin: WILLIAM MONTAGNA and PAUL F. PARAKKAL, *The Structure and Function of Skin,* 3rd ed. (1974), a textbook; LOWELL A. GOLDSMITH (ed.), *Biochemistry and Physiology of the Skin,* 2 vol. (1983), articles by authorities in the field; and ARTHUR ROOK *et al.* (eds.), *Textbook of Dermatology,* 4th ed., 3 vol. (1986), an outstanding work on normal and pathological human skin. Special aspects of the subject are discussed in HOWARD I. MAIBACH and EDWARD K. BOISITS (eds.), *Neonatal Skin: Structure and Function* (1982); and R. MARKS and P.A. PAYNE (eds.), *Bioengineering and the Skin* (1981). For current research in the field, see *The Journal of Investigative Dermatology* (monthly).

Skin diseases: WALTER F. LEVER and GUNDULA SCHAUMBURG-LEVER, *Histopathology of the Skin,* 6th ed. (1983), a standard textbook of the histology of skin diseases; THOMAS B. FITZPATRICK *et al.* (eds.), *Dermatology in General Medicine,* 3rd ed., 2 vol. (1987), a textbook with an atlas; ETAIN CRONIN, *Contact Dermatitis* (1980); ROBERT M. ADAMS, *Occupational Skin Disease* (1983); R. BARAN and R.P.R. DAWBER (eds.), *Diseases of the Nails and Their Management* (1984); and PETER D. SAMMAN and DAVID A FENTON, *The Nails in Disease,* 4th ed. (1986). Also see GEORGE C. ANDREWS, *Andrews' Diseases of the Skin: Clinical Dermatology,* 7th ed., edited by ANTHONY N. DOMONKOS, HARRY L. ARNOLD, and RICHARD B. ODOM (1982); DAVID R. BICKERS, PAUL G. HAZEN, and WILLIAM S. LYNCH, *Clinical Pharmacology of Skin Disease* (1984); LEONARD C. HARBER and DAVID R. BICKERS, *Photosensitivity Diseases: Principles of Diagnosis and Treatment* (1981); and ARTHUR ROOK and R.P.R. DAWBER, *Diseases of the Hair and Scalp* (1982). A readable general overview of personal skin care is WILLIAM MONTAGNA, FRANK PARKER, and ANTONIO TOSTI, *Skin: Your Owner's Manual* (1986).

(W.M./F.J.G.E./M.W.G.)

Human Intelligence

Much of the excitement among investigators in the field of intelligence derives from their trying to determine exactly what intelligence is. Different investigators have emphasized different aspects of intelligence in their definitions. For example, in a 1921 symposium on the definition of intelligence, the American psychologist Lewis M. Terman emphasized the ability to think abstractly, while another American psychologist, Edward L. Thorndike, emphasized learning and the ability to give good responses to questions. In a similar 1986 symposium, however, psychologists generally agreed on the importance of adaptation to the environment as the key to understanding both what intelligence is and what it does. Such adaptation may occur in a variety of environmental situations. For example, a student in school learns the material that is required to pass or do well in a course; a physician treating a patient with an unfamiliar disease adapts by learning about the disease; an artist reworks a painting in order to make it convey a more harmonious impression. For the most part, adapting involves making a change in oneself in order to cope more effectively, but sometimes effective adaptation involves either changing the environment or finding a new environment altogether.

Effective adaptation draws upon a number of cognitive processes, such as perception, learning, memory, reasoning, and problem solving. The main trend in defining intelligence, then, is that it is not itself a cognitive or mental process, but rather a selective combination of these processes purposively directed toward effective adaptation to the environment. For example, the physician noted above learning about a new disease adapts by perceiving material on the disease in medical literature, learning what the material contains, remembering crucial aspects of it that are needed to treat the patient, and then reasoning to solve the problem of how to apply the information to the needs of the patient. Intelligence, in sum, has come to be regarded as not a single ability but an effective drawing together of many abilities. This has not always been obvious to investigators of the subject, however, and, indeed, much of the history of the field revolves around arguments regarding the nature and abilities that constitute intelligence.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 434.

The article is divided into the following sections:

Theories of intelligence 710
- Psychometric theories
- Cognitive theories
- Cognitive-contextual theories
- Biologic theories

Development of intelligence 714
- Jean Piaget's work
- Post-Piaget theories
- The environmental viewpoint

Measuring intelligence 715
- The IQ test
- The distribution of IQ scores

The malleability of intelligence 716
Bibliography 716

THEORIES OF INTELLIGENCE

Theories of intelligence, as is the case with most scientific theories, have evolved through a succession of paradigms that have been put forward to clarify our understanding of the idea. The major paradigms have been those of psychological measurement (often called psychometrics); cognitive psychology, which concerns itself with the mental processes by which the mind functions; the merger of cognitive psychology with contextualism (the interaction of the environment and processes of the mind); and biologic science, which considers the neural bases of intelligence.

Psychometric theories. Psychometric theories have gen-

erally sought to understand the structure of intelligence: What form does it take, and what are its parts, if any? Such theories have generally been based on and tested by the use of data obtained from paper-and-pencil tests of mental abilities that include analogies (*e.g., lawyer* : *client* :: *doctor* : ?), classifications (*e.g.,* Which word does not belong with the others? *robin, sparrow, chicken, bluejay*), and series completions (*e.g.,* What number comes next in the following series? 3, 6, 10, 15, 21, ?).

The psychological model for psychometric theories

Underlying the psychometric theories is a psychological model according to which intelligence is a composite of abilities measured by mental tests. This model is often quantified by assuming that each test score is a weighted linear composite of scores on the underlying abilities. For example, performance on a number-series test might be a weighted composite of number, reasoning, and possibly memory abilities for a complex series. Because the mathematical model is additive, it assumes that less of one ability can be compensated for by more of another ability in test performance. For instance, two people could gain equivalent scores on a number-series test if a deficiency in number ability in the one person relative to the other was compensated for by superiority in reasoning ability.

The first of the major psychometric theories was that of the British psychologist Charles E. Spearman, who published his first major article on intelligence in 1904. Spearman noticed what, at the turn of the century, seemed like a peculiar fact: People who did well on one mental ability test tended to do well on the others, and people who did not do well on one of them also tended not to do well on the others. Spearman devised a technique for statistical analysis, which he called factor analysis, that examines patterns of individual differences in test scores and is said to provide an analysis of the underlying sources of these individual differences. Spearman's factor analyses of test data suggested to him that just two kinds of factors underlie all individual differences in test scores. The first and more important factor Spearman labeled the "general factor," or *g*, which is said to pervade performance on all tasks requiring intelligence. In other words, regardless of the task, if it requires intelligence, it requires *g*. The second factor is specifically related to each particular test. But what, exactly, is *g*? After all, calling something a general factor is not the same as understanding what it is. Spearman did not know exactly what the general factor might be, but he proposed in 1927 that it might be something he labeled "mental energy."

The concept of *g*

The American psychologist L.L. Thurstone disagreed not only with Spearman's theory but also with his isolation of a single factor of general intelligence. Thurstone argued that the appearance of just a single factor was an artifact of the way Spearman did his factor analysis and that if the analysis were done in a different and more appropriate way, seven factors would appear, which Thurstone referred to as the "primary mental abilities." The seven primary mental abilities identified by Thurstone were verbal comprehension (as involved in the knowledge of vocabulary and in reading); verbal fluency (as involved in writing and in producing words); number (as involved in solving fairly simple numerical computation and arithmetical reasoning problems); spatial visualization (as involved in mentally visualizing and manipulating objects, as is required to fit a set of suitcases into an automobile trunk); inductive reasoning (as involved in completing a number series or in predicting the future based upon past experience); memory (as involved in remembering people's names or faces); and perceptual speed (as involved in rapidly proofreading to discover typographical errors in a typed text).

It is a possibility, of course, that Spearman was right and Thurstone was wrong, or vice versa. Other psychologists, however, such as the Canadian Philip E. Vernon and the American Raymond B. Cattell, suggested another possibility—that both were right in some sense. In the view of Vernon and Cattell, abilities are hierarchical. At the top of the hierarchy is g, or general ability. But below g in the hierarchy are successive levels of gradually narrowing abilities, ending with Spearman's specific abilities. Cattell, for example, suggested in a 1971 work that general ability can be subdivided into two further kinds of abilities, fluid and crystallized. Fluid abilities are the reasoning and problem-solving abilities measured by tests such as the analogies, classifications, and series completions described above. Crystallized abilities can be said to derive from fluid abilities and be viewed as their products, which would include vocabulary, general information, and knowledge about specific fields. John L. Horn, an American psychologist, suggested that crystallized ability more or less increases over the life span, whereas fluid ability increases in the earlier years and decreases in the later ones.

Fluid and crystallized abilities

Most psychologists agreed that a broader subdivision of abilities was needed than was provided by Spearman, but not all of these agreed that the subdivision should be hierarchical. J.P. Guilford, an American psychologist, proposed a structure-of-intellect theory, which in its earlier versions postulated 120 abilities. For example, in an influential 1967 work Guilford argued that abilities can be divided into five kinds of operations, four kinds of contents, and six kinds of products. These various facets of intelligence combine multiplicatively, for a total of $5 \times 4 \times 6$, or 120 separate abilities. An example of such an ability would be cognition (operation) of semantic (content) relations (product), which would be involved in recognizing the relation between *lawyer* and *client* in the analogy problem, *lawyer* : *client* :: *doctor* : ?. In 1984 Guilford increased the number of abilities proposed by his theory, raising the total to 150.

It had become apparent that there were serious problems with psychometric theories, not just individually but as a basic approach to the question. For one thing, the number of abilities seemed to be getting out of hand. A movement that had started by postulating one important ability had come, in one of its major manifestations, to postulating 150. Because parsimony is usually regarded as one of several desirable features of a scientific theory, this number caused some disturbance. For another thing, the psychometricians, as practitioners of factor analysis were called, didn't seem to have any strong scientific means of resolving their differences. Any method that could support so many theories seemed somewhat suspect, at least in the use to which it was being put. Most significant, however, was the seeming inability of psychometric theories to say anything substantial about the processes underlying intelligence. It is one thing to discuss "general ability" or "fluid ability," but quite another to describe just what is happening in people's minds when they are exercising the ability in question. The cognitive psychologists proposed a solution to these problems, which was to study directly the mental processes underlying intelligence and, perhaps, relate them to the factors of intelligence proposed by the psychometricians.

Cognitive theories. During the era of psychometric theories, the study of intelligence was dominated by those investigating individual differences in people's test scores. In an address to the American Psychological Association in 1957, the American psychologist Lee Cronbach, a leader in the testing field, decried the fact that some psychologists study individual differences and others study commonalities in human behaviour but never do the two meet. In Cronbach's address his plea to unite the "two disciplines of scientific psychology" led, in part, to the development of cognitive theories of intelligence and of the underlying processes posited by these theories.

Dangers in evaluating test scores

Without an understanding of the processes underlying intelligence it is possible to come to misleading, if not wrong, conclusions when evaluating overall test scores or other assessments of performance. Suppose, for example, that a student does poorly on the type of verbal analogies questions commonly found on psychometric tests. A possible conclusion is that the student does not reason well. An equally plausible interpretation, however, is that the student does not understand the words or is unable to read them in the first place. A student seeing the analogy, *audacious* : *pusillanimous* :: *mitigate* : ?, might be unable to solve it because of a lack of reasoning ability, but a more likely possibility is that the student does not know the meanings of the words. A cognitive analysis enables the interpreter of the test score to determine both the degree to which the poor score is due to low reasoning

ability and the degree to which it is a result of not understanding the words. It is important to distinguish between the two interpretations of the low score, because they have different implications for understanding the intelligence of the student. A student might be an excellent reasoner but have only a modest vocabulary, or vice versa.

Mental processes and representation

Underlying most cognitive approaches to intelligence is the assumption that intelligence comprises a set of mental representations (*e.g.,* propositions, images) of information and a set of processes that can operate on the mental representations. A more intelligent person is assumed to represent information better and, in general, to operate more quickly on these representations than does a less intelligent person. Researchers have sought to measure the speed of various types of thinking. Through mathematical modeling, they divide the overall time required to perform a task into the constituent times needed to execute each mental process. Usually, they assume that these processes are executed serially—one after another—and, hence, that the processing times are additive. But some investigators allow for partially or even completely parallel processing, in which case more than one process is assumed to be executed at the same time. Regardless of the type of model used, the fundamental unit of analysis is the same: a mental process acting upon a mental representation.

A number of cognitive theories of intelligence have evolved. Among them is that of the American psychologists Earl B. Hunt, Nancy Frost, and Clifford E. Lunneborg, who in 1973 showed one way in which psychometrics and cognitive modeling could be combined. Instead of starting with conventional psychometric tests, they began with tasks that experimental psychologists were using in their laboratories to study the basic phenomena of cognition, such as perception, learning, and memory. They showed that individual differences in these tasks, which had never before been taken seriously, were in fact related (although rather weakly) to patterns of individual differences in psychometric intelligence test scores. These results, they argued, showed that the basic cognitive processes might be the building blocks of intelligence.

Following is an example of the kind of task Hunt and his colleagues studied in their research. The experimental subject is shown a pair of letters, such as "A A," "A a," or "A b." The subject's task is to respond as quickly as possible to one of two questions: "Are the two letters the same physically?" or "Are the two letters the same only in name?" In the first pair the letters are the same physically, and in the second pair the letters are the same only in name.

The psychologists hypothesized that a critical ability underlying intelligence is that of rapidly retrieving lexical information, such as letter names, from memory. Hence,

Computation of reaction time to test questions

they were interested in the time needed to react to the question about letter names. They subtracted the reaction time to the question about physical match from the reaction time to the question about name match in order to isolate and set aside the time required for sheer speed of reading letters and pushing buttons on a computer. The critical finding was that the score differences seemed to predict psychometric test scores, especially those on tests of verbal ability, such as verbal analogies and reading comprehension. The testing group concluded that verbally facile people are those who have the underlying ability to absorb and then retrieve from memory large amounts of verbal information in short amounts of time. The time factor was the significant development here.

A few years later, the American psychologist Robert J. Sternberg suggested an alternative approach to studying the cognitive processes underlying human intelligence. He argued that Hunt and his colleagues had found only a weak relation between basic cognitive tasks and psychometric test scores because the tasks they were using were at too low a level. Although low-level cognitive processes may be involved in intelligence, according to Sternberg they are peripheral rather than central. He proposed that psychologists should study the tasks found on the intelligence tests and then determine the mental processes and strategies that people use to perform those tasks.

Sternberg began his study with the analogies tasks such as *lawyer* : *client* :: *doctor* : ?. He determined that the solution to such analogies requires a set of component cognitive processes: namely, encoding of the analogy terms (*e.g.,* retrieving from memory attributes of the terms *lawyer, client,* and so on), inferring the relation between the first two terms of the analogy (*e.g.,* figuring out that a lawyer provides professional services to a client), mapping this relation to the second half of the analogy (*e.g.,* figuring out that both a lawyer and a doctor provide professional services), applying this relation to generate a completion (*e.g.,* realizing that the person to whom a doctor provides professional services is a patient), and then responding. Using techniques of mathematical modeling applied to reaction-time data, Sternberg proceeded to isolate the components of information processing. He determined whether or not each experimental subject did, indeed, use these processes, how the processes were combined, how long each process took, and how susceptible each process was to error. Sternberg later showed that the same cognitive processes are involved in a wide variety of intellectual tasks, and he suggested that these and other related processes underlie scores on intelligence tests.

Sternberg's use of analogies

Other cognitive psychologists have pursued different paths in the study of human intelligence, including the building of computer models of human cognition. Two leaders in this field have been the American psychologists Allen Newell and Herbert A. Simon. In the late 1950s and early 1960s they worked with a computer expert, Clifford Shaw, to construct a computer model of human problem solving. Called the General Problem Solver, it could solve a wide range of fairly structured problems, such as logical proofs and mathematical word problems. Their program relied heavily on a heuristic procedure called "means-ends analysis," which, at each step of problem solving, determined how close the program was to a solution and then tried to find a way to bring the program closer to where it needed to be. In 1972, Newell and Simon proposed a general theory of problem solving, much of which was implemented on the computer.

The Newell–Simon theory of problem solving

Most of the problems studied by Newell and Simon were fairly well structured, in that it was possible to identify a discrete set of moves that would lead from the beginning to the end of a problem. For example, in logical-theorem proving the final result is known, and what is needed is a discrete set of steps that lead to that solution. Even in chess, another object of study, a discrete set of moves can be determined that will lead from the beginning of a game to checkmate. The biggest problem for a computer program (or a human player, for that matter) is in deciding which of myriad possible moves will most contribute toward winning a game. Other investigators have been concerned with less well-structured problems, such as how a text is comprehended, or how people are reminded of things they already know when reading a text.

All of the cognitive theories described so far have in common their primary reliance on what psychologists call the serial processing of information. Fundamentally, this means that cognitive processes are executed in series, one after another. In solving an algebra problem, for example, first the problem is studied, then an attempt is made to formulate some equations to define knowns and unknowns, then the equations may be used to solve for the unknowns, and so on. The assumption is that people process chunks of information one at a time, seeking to combine the processes used into an overall strategy for solving a problem.

For many years, various psychologists have challenged the idea that cognitive processing is primarily serial. They have suggested that cognitive processing is primarily parallel, meaning that humans actually process large amounts of information simultaneously. It has long been known that the brain works in such a way, and it seems reasonable that cognitive models should reflect this reality. It proved, however, to be difficult to distinguish between serial and parallel models of information processing, just as it had been difficult earlier to distinguish between different factor models of human intelligence. Subsequently advanced techniques of mathematical and computer modeling were brought to bear on this problem, and various researchers,

Computer modeling in intelligence research

including the American psychologists David E. Rumelhart and Jay L. McClelland, proposed what they call "parallel distributed processing" models of the mind. These models postulated that many types of information processing occur at once, rather than just one at a time.

Even with computer modeling, some major problems regarding the nature of intelligence remain. For example, a number of psychologists, such as the American Michael E. Cole, have argued that cognitive processing does not take into account that the description of intelligence may differ from one culture to another and may even differ from one group to another within a culture. Moreover, even within the mainstream cultures of North America and Europe, it had become well known that conventional tests, even though they may predict academic performance, do not reliably predict performance in jobs or other life situations beyond school. It seemed, therefore, that not only cognition but also the context in which cognition operates had to be taken into account.

Cognitive-contextual theories. Cognitive-contextual theories deal with the way that cognitive processes operate in various environmental contexts. Two of the major theories of this type are that of the American psychologist Howard Gardner and that of Sternberg. In 1983 Gardner proposed a theory of what he called "multiple intelligences." Earlier theorists had gone so far as to contend that intelligence comprises multiple abilities. But Gardner went a step further, arguing that there is no single intelligence. In his view, intelligences are multiple, including, at a minimum, linguistic, logical-mathematical, spatial, musical, bodily-kinesthetic, interpersonal, and intrapersonal intelligence. Some of these intelligences are quite similar to the abilities proposed by the psychometric theorists, but others are not. For example, the idea of a musical intelligence is relatively new, as is the idea of a bodily-kinesthetic intelligence, which encompasses the particular faculties of athletes and dancers. Gardner derived his listing of intelligences from a variety of sources, including studies of cognitive processing, of brain damage, of exceptional individuals, and of cognition across cultures. Gardner proposed that whereas most concepts of intelligence had been ethnocentric and culturally biased, his was universal, based upon biologic and cross-cultural data as well as upon data derived from the cognitive performance of a wide array of people.

Gardner's "multiple intelligences" theory

The "triarchic" theory

An alternative theory also taking into account both cognition and context is Sternberg's "triarchic" theory of human intelligence. Sternberg agreed with Gardner that conventional notions of intelligence were too narrow. But he disagreed as to how psychologists should go beyond traditional conceptions, suggesting that abilities such as musical and bodily-kinesthetic ones are talents rather than intelligences in that they are fairly specific and are not prerequisites for adaptation in most cultures.

According to Sternberg, intelligence has three aspects. These are not multiple intelligences, as in Gardner's scheme. Where Gardner viewed the various intelligences as separate and independent, Sternberg posited three integrated and interdependent aspects of intelligence. These aspects relate intelligence to what goes on internally within a person, to what goes on in the external world, and to experience, which mediates between the internal and external worlds.

The first aspect consists of the cognitive processes and representations that form the core of all thought. Sternberg distinguished three kinds of processes: those involved in deciding what to do and later in deciding how well it was done, those involved in doing what one had decided to do, and those involved in learning how to do it in the first place. For example, when deciding which of two brands of a product to buy, a shopper might first decide how to decide which is a better purchase (say, in terms of unit cost), then actually do the required calculation. Of course the shopper had first to have learned how to calculate unit prices.

The second aspect consists of the application of these processes to the external world. According to Sternberg, mental processes serve three functions in the everyday world: adaptation to existing environments, the shaping of existing environments into new ones, and the selection of new environments when old ones prove unsatisfactory. The theory holds that more intelligent persons are not just those who can execute many cognitive processes quickly or well; rather, their greater intelligence is reflected in knowing what their strengths and weaknesses are and capitalizing upon their strengths while remedying or compensating for their weaknesses. More intelligent persons, then, find a niche in which they can operate most efficiently.

The role of experience

The third aspect of Sternberg's triarchic theory is the integration of the internal and external worlds through experience. One measure of intelligence is the ability to cope with relatively novel situations. For example, intelligence might be measured by taking someone who is well adapted to one culture and placing him in an unfamiliar one, in order to assess his ability to cope with a new situation. Or in the case of a person for whom an automobile is of critical importance, intelligence may be gauged according to the way that person functions when the car is being repaired and unavailable for a short period. Sternberg also suggested that another facet of experience that is important in evaluating intelligence is the automatization of cognitive processing, which occurs when a once relatively novel task becomes familiar. For example, when children first learn to read, they are confronted with a relatively novel task. But one is interested not only in how well children deal with the novelty of reading but also in how well they render reading automatic and subconscious. The abilities to cope with relative novelty and to automatize cognitive processing are seen as interrelated: The more a person is able to automatize the tasks of daily life, the more mental resources there are left to cope with novelty.

Biologic theories. The theories discussed above seek to understand intelligence in terms of underlying hypothetical constructs, whether these constructs are factors (*e.g.,* verbal ability, spatial ability), as is the case with psychometric theories; cognitive processes (*e.g.,* retrieval of information from memory, inferring relations); or cognitive processes as they interact with context (*e.g.,* musical intelligence or the shaping of the environment). Some theorists, however, have taken a radically different approach, seeking to understand intelligence directly in terms of its biologic bases without intervening hypothetical constructs. These theorists, usually called reductionists, believe that a true understanding of intelligence can result only from the identification of its biologic substrates. Reductionism takes an appealing philosophical stance, and some would argue that there is no real alternative if the goal is to explain rather than merely to describe behaviour. But the case is not an open-and-shut one. In trying to discover why an automobile does not start in the morning, for example, the driver does not usually imagine that the basic problem is one involving molecules and atoms. The driver is probably better advised to analyze the performance of components, such as the starter or carburetor. Even if the automobile's molecular or atomic reactions could be analyzed, such an analysis would probably be unhelpful. The example suggests that the biologic approaches to intelligence should be looked at as complementary to, rather than as replacing, other approaches. Although relatively little is known about the biologic bases of intelligence, progress has been made on three different fronts, all involving studies of brain operation.

Hemispheric studies. One front has been the investigation of types of intellectual performance as related to the regions of the brain from which they originate. A researcher in this area, the American psychologist Jerre Levy, investigated the functioning of the two hemispheres of the brain. Levy and others found that the left hemisphere is superior in analytical functioning, of which the use of language, for instance, is a prime example. The right hemisphere, on the other hand, is superior in many forms of visual and spatial performance and tends to be more synthetic and holistic in its functioning than the left. Such patterns of hemispheric specialization are complex and cannot easily be generalized.

Functions of the brain's hemispheres

The specialization of the two hemispheres of the brain is exemplified in an early study by Levy and another American psychologist, Roger W. Sperry, who worked with split-brain patients, individuals who have had their

corpus callosum severed. Because the corpus callosum, in the normal brain, links the two hemispheres, in these patients the hemispheres function independently of each other. But, as in normal persons, the right side of the body connects with the left hemisphere of the brain, and the left side connects with the right hemisphere.

Levy and Sperry asked split-brain patients to match small wooden blocks held in either their left or their right hand (but not looked at) with corresponding two-dimensional pictures. They found that the left hand did better than the right at this task, but, of more interest, they found that the two hands appeared to use different strategies in solving the problem. Their analysis demonstrated that the right hand found it relatively easier to deal with patterns that are readily described in words but difficult to discriminate visually. In contrast, the left hand found it easier to deal with patterns requiring visual discrimination.

Brain-wave studies. A second front of research has involved the use of brain-wave recordings to study the relation between these waves and either performance on ability tests or in various kinds of cognitive tasks. Researchers in some of these studies found a relationship between certain aspects of electroencephalogram (EEG) waves and scores on a standard psychometric test of intelligence.

Blood-flow studies. A third and relatively new front of research has involved the measurement of blood flow in the brain, which is a fairly direct indicator of functional activity in brain tissue. In such studies the amount and location of blood flow in the brain is monitored while subjects perform cognitive tasks. John Horn, a prominent researcher in this area, found that older adults show decreased blood flow to the brain, that such decreases are greater in some areas of the brain than in others, and that the decreases are particularly notable in those areas responsible for close concentration, spontaneous alertness, and the encoding of new information. These findings highlight the importance not only of understanding intelligence in general but also of understanding it as a faculty that develops over time.

DEVELOPMENT OF INTELLIGENCE

There have been a number of approaches to the study of the development of intelligence. Psychometric theorists, for instance, have sought to understand how intelligence develops in terms of changes in the factors of intelligence over time and changes in the amounts of the various abilities that children have. For example, the concept of mental age was popular during the first half of the 20th century. A given mental age was held to represent an average child's level of mental functioning for a given chronological age. Thus, an average 12-year-old would have a mental age of 12, but an above-average 10-year-old or a below-average 14-year-old might also have a mental age of 12 years. The concept of mental age fell into disfavour, however, and has come to be used only rarely. Two main reasons appear to have caused this change. First, the concept does not seem to work after about the age of 16. The mental test performance of, say, a 25-year-old is generally no better than that of a 24- or 23-year-old. Furthermore, in later adulthood some test scores seem to start declining. Second, many psychologists believe that intellectual development does not exhibit the kind of smooth continuity that the concept of mental age appears to imply. Rather, development seems to come in intermittent bursts, whose timing can differ from one child to another.

The concept of mental age

Jean Piaget's work. The landmark work in intellectual development has not come out of psychometry but rather out of a tradition forged by the Swiss psychologist Jean Piaget. Over the course of a long career, Piaget formulated what became one of the monumental theories in the history of psychology. Two of its main aspects concern the mechanisms by which intellectual development takes place and the periods through which children develop. As modeled by Piaget, the child explores the world and observes regularities and makes generalizations, much as a scientist does. The first part of Piaget's theory recognizes two fundamental cognitive processes that work in somewhat reciprocal fashion. The first is what Piaget called assimilation, a process that involves incorporating new information into an already existing cognitive structure. Suppose, for example, that a child knows how to solve problems that require a percentage of a given number to be calculated. The child then learns how to solve problems that ask what percentage of a number another number is. The child already has a cognitive structure, or what Piaget called a "schema," for percentage problems and can incorporate the new knowledge into the existing structure. But suppose the child is to learn next how to solve time-rate-distance problems, never before having dealt with this type of problem. In this case, the child may need to call upon the second process, accommodation, to form a new cognitive structure that can incorporate the new information. Cognitive development, according to Piaget, represents a dynamic equilibrium between the two processes of assimilation and accommodation.

Piaget's "schema"

The second part of Piaget's theory postulates that there are four major periods in intellectual development. The first, the sensorimotor period, extends from birth through roughly two years of age. During this period, a child learns how to modify reflexes to make them more adaptive, to coordinate actions, to retrieve hidden objects, and, eventually, to begin representing information mentally. During the second, preoperational, period from about two to seven years of age, a child experiences the growth of language and mental imagery and learns to focus on single perceptual dimensions, such as colour and size. The third, concrete-operational, period from about seven to 12 years of age is the time during which a child develops an important set of skills referred to as conservation skills. For example, suppose that water is poured from a wide, short beaker into a tall and narrow one. A preoperational child, asked which beaker has more water, will say that the second beaker does (the tall, thin one); a concrete-operational child, however, will recognize that the amount of water in the beakers must be the same. Finally, children emerge into the fourth, formal-operational, period, which begins about age 12 and continues throughout life. The formal-operational child develops thinking skills in all logical combinations and learns to think with abstract concepts. For example, a concrete-operational child asked to determine all possible orderings, or permutations, of four digits, such as 3-7-5-8, will have great difficulty doing so. The formal-operational child, however, will adopt a strategy of systematically varying alternations of digits, starting perhaps with the last digit and working toward the first. This systematic way of thinking is not possible for the normal concrete-operational child.

De-emphasis on Piagetian concepts

Piaget's theory had a major impact on the views of intellectual development, but the theory no longer has the widespread acceptance it once had, particularly from the 1950s through the 1970s. One reason for this is that the theory deals primarily with scientific and logical modes of thought and much less with aesthetic, intuitive, and other modes of thought. Another reason is that Piaget tended to overestimate the ages at which children could first perform certain cognitive tasks. Despite its diminished influence, however, Piaget's theory continues to serve as a basis for other theories.

Post-Piaget theories. Later theories of intellectual development, influenced by Piaget, have taken several courses. One has been to expand on Piaget by suggesting a possible fifth, adult, period of development, such as problem finding. A second course has been to suggest quite different periods of development from those suggested by Piaget. A third course has been to accept that intellectual development occurs through the periods Piaget proposed, but to hold that the cognitive bases of development differ from those recognized by him. Some of these theories emphasize the importance of memory capacity. For example, it has been shown that children's difficulties in solving transitive inference problems (such as: If A is greater than B, B is greater than C, and D is less than C, which is the greatest?) result primarily from memory limitations rather than reasoning limitations, as argued by Piaget. A fourth course has been to focus on the role of knowledge in development. Some investigators argue that much of what has been attributed to reasoning and problem-solving ability in intellectual development is actually better attributed

to the increasing knowledge base of a child. Although the above approaches are diverse, they are all related as alternative responses to Piaget.

The environmental viewpoint. The views of intellectual development described above all emphasize the importance of the organism in intellectual development. But an alternative viewpoint emphasizes the importance of the environmental context and, particularly, the social environment. This view is related to the cognitive-contextual theories discussed above. Championed originally by L.S. Vygotsky, a Soviet psychologist, this viewpoint suggests that intellectual development may be largely influenced by a child's interactions with others: A child sees others thinking and acting in certain ways and then internalizes and models what is seen. An Israeli psychologist, Reuven Feuerstein, has elaborated upon this point of view, suggesting that the key to intellectual development is what he calls mediated learning experience. The parent mediates, or interprets, the environment for the child, and it is largely through this mediation that the child learns to understand and interpret the world.

MEASURING INTELLIGENCE

Almost all of the theories discussed above have in common the use of fairly complex tasks for gauging intelligence in both children and adults. Some of these tasks have been explicitly discussed—for example, those requiring recognition of analogies, classification of similar terms, extrapolation of number series, performance of transitive inferences, and the like. How did theorists of intelligence come to use these tasks in preference to the myriad other possibilities? That question is answered through the discussion of the measurement of intelligence.

Although the kinds of complex tasks that have been discussed above fall into a single tradition for the measurement of intelligence, the field actually has two major traditions. The tradition that has been discussed most prominently and has been most influential is that of the French psychologist Alfred Binet. But an earlier tradition, and one that still shows some influence upon the field, is that of the English scientist Sir Francis Galton.

The influence of Sir Francis Galton

The publication in 1859 of Charles Darwin's *The Origin of Species* had a profound effect on many lines of scientific endeavour. The book suggested that the capabilities of humans are, in some sense, continuous with those of lower animals and, hence, can be understood through scientific investigation. One person who was strongly influenced by Darwin's thinking was his cousin Sir Francis Galton. For seven years—from 1884 to 1890—Galton maintained a laboratory at the South Kensington Museum in London, where, for a small fee, visitors could have themselves measured on a variety of psychophysical tasks, such as weight discrimination and sensitivity to musical pitch. But Galton believed that these tests measured more than just psychophysical abilities. He believed that psychophysical abilities were the basis of intelligence and, hence, that his tasks were measures of intelligence. The earliest formal intelligence test, therefore, required a person to perform such simple tasks as deciding which of two weights was heavier or showing how forcefully he could squeeze his hand. The Galtonian tradition was taken to the United States by the psychologist James McKeen Cattell. One of Cattell's students, Clark Wissler, collected data showing that scores on Galtonian types of tasks were not good predictors of grades in college, or even of each other. Cattell continued his work in psychometric research, however, and with Edward L. Thorndike developed the principal facility in the United States for mental testing and measurement.

The IQ test. The more influential tradition of mental testing was developed by Binet and his collaborator, Theodore Simon, in France. In 1904 the minister of public instruction in Paris named a commission to study or create tests that would insure that mentally retarded children received an adequate education. The minister was also concerned that certain children were being placed in classes for the retarded not because they were retarded but because they had behaviour problems, and teachers did not want them in their classrooms. Even before Wissler's research, Binet, who was charged with developing the new test, had flatly rejected the Galtonian tradition, believing that Galton's tests measured fairly trivial abilities. He proposed instead that tests of intelligence should measure skills such as judgment, comprehension, and reasoning—the same kinds of skills measured on most intelligence tests today. Binet's early test was taken to the United States by a Stanford University psychologist, Lewis Terman, whose version came to be called the Stanford-Binet test. This test has been revised frequently and continues in use.

Stanford-Binet test

The Stanford-Binet test, and others like it, have traditionally yielded at the very least an overall score referred to as an intelligence quotient, or IQ. Some tests, such as the Wechsler Adult Intelligence Scale (Revised) and the Wechsler Intelligence Scale for Children (Revised) yield an overall IQ as well as separate IQs for verbal and performance subtests. An example of a verbal subtest would be vocabulary, whereas an example of a performance subtest would be picture arrangement, the latter requiring an examinee to arrange a set of pictures into a sequence so that they tell a comprehensible story.

IQ was originally computed as the ratio of mental age to chronological (physical) age, multiplied by 100. Thus, if a child of 10 had a mental age of 12 (that is, performed on the test at the level of an average 12-year-old), the child was assigned an IQ of $(12/10) \times 100$, or 120. If the 10-year-old had a mental age of eight, the child's IQ would be $(8/10) \times 100$, or 80. A score of 100, where the mental age equals the chronological age, is average.

As discussed above, the concept of mental age has fallen into disrepute, and few tests continue to involve the computation of mental ages. Many tests still yield an IQ, but they are most often computed on the basis of statistical distributions. The scores are assigned on the basis of what percentage of people of a given group would be expected to have a certain IQ. (See PSYCHOLOGICAL TESTS AND MEASUREMENT.)

The distribution of IQ scores. Intelligence test scores follow an approximately normal distribution, meaning that most people score near the middle of the distribution of scores and that scores drop off fairly rapidly in frequency as one moves in either direction from the centre. For example, on the IQ scale, about two out of three scores fall between IQs of 85 and 115, and about 19 out of 20 scores fall between 70 and 130. Put another way, only one out of 20 scores differs from the average IQ (100) by more than 30 points.

"Labeling" IQ scores

It has been common to associate certain levels of IQ with labels. For example, at the upper end, the label "gifted" is sometimes assigned to people with IQs over a certain point, such as 130. And at the lower end, mental retardation has been classified into different degrees depending upon IQ, so that, for example, IQs of 70–84 have been classified as borderline retarded, IQs of 55–69 as mildly retarded, IQs of 40–54 as moderately retarded, IQs of 25–39 as severely retarded, and IQs below 25 as profoundly retarded. Labeling schemes like these, however, have pitfalls and are in some ways dangerous.

First, the labels assume that conventional intelligence tests provide sufficient information to classify someone as either gifted, on the one hand, or mentally retarded, on the other. But most authorities would agree that this assumption is almost certainly false. Conventional intelligence tests are useful in providing information about some people some of the time, but the information they provide is about a fairly narrow range of abilities. To label someone as mentally retarded solely on the basis of a single test score is to risk doing a potentially great disservice and injustice to that person. Most psychologists and other authorities recognize that social as well as strictly intellectual skills are important in classifying a person as retarded. If a person adapts well to the environment, then it seems inappropriate to refer to that person as mentally retarded, a term with inescapably pejorative connotations.

Second, giftedness is generally recognized as more than just a degree of intelligence, even broadly defined. Most psychologists who have studied gifted persons agree that a variety of aspects make up giftedness. Howard E. Gruber, the Swiss psychologist, and the American psychologist Mihaly Csikszentmihalyi are among a number of researchers

Views on giftedness

who are skeptical that the giftedness observed in children has much to do with the giftedness observed in adults. Gruber believes that giftedness unfolds over the course of a lifetime and involves achievement at least as much as intelligence. Gifted people, he contends, have life plans that they seek to realize, and these plans develop over the course of many years. To measure giftedness merely in terms of a single test score would be, for Gruber, a trivialization of the concept.

Third, a given test score can mean different things to different people. An IQ score for a person who has grown up in a ghetto home and gone to an inadequate school does not have the same meaning as the same IQ score for someone who has grown up in an upper-middle-class suburban environment and gone to a well-endowed school. An IQ score also does not mean the same thing for a person whose first language is not English but who takes a test in English, as it does for a native English-speaker. Another factor is that some people are "test-anxious" and may do poorly on almost any standardized test. Based on these and similar drawbacks, it has come to be believed generally that scores have to be interpreted carefully on an individual basis.

Psychologists now believe that IQ represents only a part of intelligence, and intelligence is only one factor in both retardation and giftedness. Earlier rigid concepts in the field of intelligence measurement, which led to labeling, have had undesirable effects. The growth of a more recent concept, the malleability of intelligence, has also served to discredit labeling.

THE MALLEABILITY OF INTELLIGENCE

Intelligence has historically been conceptualized as a more or less fixed trait. This view perceives intelligence as something people are born with, and the function of development is to allow this genetic endowment to express itself. A number of investigators have taken the approach that intelligence is highly heritable, transmitted through the genes. Other investigators believe that intelligence is minimally heritable, if at all. Most authorities take an intermediate position.

Various methods are used to assess the heritability of intelligence. Notable among these is the study of identical twins reared apart. For a variety of reasons, identical twins are occasionally separated at or near birth. If the twins are raised apart, and if it is assumed that when twins are separated they are randomly distributed across environments (often a dubious assumption), then the twins would have in common all of their genes but none of their environment, except for chance environmental overlap. As a result, the correlation between their performance on tests of intelligence can provide an estimate of the proportion of variation in test scores due to heredity. Another method of computing the hereditary effect on intelligence involves comparing the relationship between intelligence test scores of identical twins and those of fraternal twins.

Studies of heritability of intelligence

Considering the large number of studies that have investigated the heritability of intelligence, it is surprising that so much disagreement exists among researchers. It has been estimated that roughly half the variation in intelligence test scores is caused by hereditary influences. But it is significant that estimates of heritability can differ among ethnic and racial groups, as well as across time within a single group. Moreover, the estimates are computed, for the most part, on the basis of intelligence test scores, so that the estimates are only for that part of intelligence measured by the tests.

Whatever the heritability factor of IQ, a separate issue is whether intelligence can be increased. Work by a New Zealand researcher, James Flynn, has shown that, in the late 20th century, scores on intelligence tests have been rising rather steadily throughout the world. Although the reason for the increase has not been satisfactorily explained, there is little doubt that this is a developing phenomenon requiring careful investigation.

Despite the general increase in scores, average IQs continue to vary both across countries and across different socioeconomic groups. For example, many researchers have found a positive correlation between socioeconomic status and IQ, although they disagree over the reason for the relationship. Most investigators agree that differences in educational opportunities play an important role, and some investigators believe that there is a hereditary basis for the difference as well. But there is simply no broad consensus on the issue of why the differences exist, and, again, it should be noted that the differences are based on IQ, not broadly defined intelligence.

It is important to understand that no matter how heritable intelligence is, some aspects of it are still malleable. Heritability of a trait is a separate issue from its malleability. A person's height, for example, is 90 percent heritable; the best predictor of height is the height of a person's parents. Yet, because of better nutrition and health care, average heights in the United States have climbed during the 20th century. Thus, with intervention, even a highly heritable trait can be modified. There is a growing body of evidence that aspects of intelligence, too, can be modified. Intelligence, in the view of many authorities, is not a fixed trait, with its level a foregone conclusion the day a person is born. To the contrary, a program of training in intellectual skills can increase some aspects of a person's level of intelligence. No training program—no environmental condition of any sort—can make a genius of a person with low measured intelligence. But some gains are possible, and a number of programs have been developed for increasing intellectual skills. A main trend for psychologists working in the intelligence field has been to combine testing and training functions in order to enable people to optimize their intelligence.

Modification of intelligence

BIBLIOGRAPHY

General works: Introductions that provide a frame of reference and the terminology necessary for understanding the study of intelligence are found in such comprehensive sources as ROM HARRÉ and ROGER LAMB (eds.), *The Encyclopedic Dictionary of Psychology* (1983); BENJAMIN B. WOLMAN (ed.), *Handbook of Intelligence: Theories, Measurements, and Applications* (1985); RICHARD L. GREGORY (ed.), *The Oxford Companion to the Mind* (1987); and RAYMOND J. CORSINI (ed.), *Encyclopedia of Psychology*, 4 vol. (1984). Applications and varieties of intelligence are discussed in ROBERT J. STERNBERG and RICHARD K. WAGNER (eds.), *Practical Intelligence: Nature and Origins of Competence in the Everyday World* (1986). A comprehensive but fairly elementary introduction to the field is ROBERT KAIL and JAMES W. PELLEGRINO, *Human Intelligence: Perspectives and Prospects* (1985). For current research in the field, articles in *Psychology Today* (monthly) provide coverage on a general level.

Theories of intelligence: For early theories, see ALFRED BINET and THÉODORE SIMON, *The Development of Intelligence in Children: The Binet-Simon Scale*, trans. from French (1916, reprinted 1983); JOAN W. REEVES, *Thinking About Thinking: Studies in the Background of Some Psychological Approaches* (1965), a summary of Binet's work; and CHARLES E. SPEARMAN, *The Nature of "Intelligence" and the Principles of Cognition* (1923, reprinted 1973), and *The Abilities of Man: Their Nature and Measurement* (1927, reissued 1970). Later theories are presented in HOWARD GARDNER, *Frames of Mind: The Theory of Multiple Intelligences* (1983); RAYMOND B. CATTELL, *Intelligence: Its Structure, Growth, and Action* (1987); ROBERT J. STERNBERG, *Beyond IQ: A Triarchic Theory of Human Intelligence* (1985), and *The Triarchic Mind: A New Theory of Human Intelligence* (1988); MICHAEL COLE and BARBARA MEANS, *Comparative Studies of How People Think* (1981); and ROGER SPERRY, *Science and Moral Priority: Merging Mind, Brain, and Human Values* (1983). Comprehensive reviews of the field are presented in ROBERT J. STERNBERG (ed.), *Handbook of Human Intelligence* (1982), and *Human Abilities: An Information-Processing Approach* (1985); and ROBERT J. STERNBERG and DOUGLAS K. DETTERMAN (eds.), *What Is Intelligence?: Contemporary Viewpoints on Its Nature and Definition* (1986).

Development of intelligence: A definitive summary of Piaget's earlier work is presented in JOHN H. FLAVELL, *The Developmental Psychology of Jean Piaget* (1963). Piaget's works on the mechanisms of intellectual development and fundamental cognitive processes are available in English translations: JEAN PIAGET, *The Psychology of Intelligence*, trans. by MALCOLM PIERCY and D.E. BERLYNE (1950, reprinted 1981; originally published in French, 1947); and *The Essential Piaget*, ed. by HOWARD E. GRUBER and J. JACQUES VONÈCHE (1977, reprinted 1982). For other influential views, see L.S. VYGOTSKY, *Mind in Society: The Development of Higher Psychological Processes*, trans. from Russian, ed. by MICHAEL COLE *et al.* (1978), the cognitive-contextual theory of intellectual development; and

REUVEN FEUERSTEIN, *The Dynamic Assessment of Retarded Performers: The Learning Potential Assessment Device, Theory, Instruments, and Techniques* (1979, reprinted 1985), an analysis of the role of learning experiences in intellectual development.

Measuring intelligence: Early traditional approaches to evaluating intelligence are presented in FRANCIS GALTON, *Hereditary Genius: An Inquiry into Its Laws and Consequences* (1869, reissued 1978), and *Inquiries into Human Faculty and Its Development* (1883, reissued 1973). For American investigations at the beginning of the 20th century and their results, see EDWARD L. THORNDIKE *et al.*, *The Measurement of Intelligence* (1927, reprinted 1973). Developments and applications of Binet's tradition of mental testing are described in LEWIS M. TERMAN, *The Measurement of Intelligence: An Explanation of and a Complete Guide for the Use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale* (1916, reprinted 1975). For later research in the field, see PHILIP E. VERNON, *The Measurement of Abilities,* 2nd ed. (1972); DAVID WECHSLER, *Wechsler's Measurement and Appraisal of Adult Intelligence,* 5th enl. ed., rev. by JOSEPH D. MATARAZZO (1972); and ANNE ANASTASI, *Psychological Testing,* 6th ed. (1988).

Malleability of intelligence: Arguments for and against heritability as an essential trait of intelligence are presented in the works of ARTHUR R. JENSEN, *Genetics and Education* (1972); JOHN C. LOEHLIN and ROBERT C. NICHOLS, *Heredity, Environment, & Personality: A Study of 850 Sets of Twins* (1976); PHILIP E. VERNON, *Intelligence, Heredity, and Environment* (1979); and H.J. EYSENCK and LEON J. KAMIN, *Intelligence, the Battle for the Mind* (1981).

(R.J.St.)

Intelligence and Counterintelligence

In a world of sovereign nations, information is a prime element of national power. Intelligence, which is best defined as evaluated information, is the vital and often pivotal foundation for national decisions. The nature of intelligence is illustrated by a U.S. presidential order (Executive Order No. 12333) of December 4, 1981, which directed the national intelligence system to "Provide the President . . . with the necessary information on which to base decisions concerning the conduct and development of foreign, defense and economic policy, and the protection of United States national interests from foreign security threats." The order directed that "all means" consistent with the law and rights of individuals "be used to develop intelligence information for the President" and specified that "emphasis should be given to detecting and countering espionage and other threats . . . by foreign intelligence services against the United States government." These official words define the functions of intelligence and counterintelligence, but they omit a third and most controversial role of intelligence agencies, that of intervening secretly in the political or economic affairs of other nations, an activity commonly referred to as covert action.

Intelligence, counterintelligence, and covert action denote a wide variety of governmental activities related to national security and defense and foreign policy. A statesman's day often begins and ends with the reading of intelligence reports. Accurate information is essential to, but does not guarantee, optimal decisions, while inadequate information has demonstrably led to disaster. The need for information must be recognized, and it must be collected efficiently, interpreted with sophistication, communicated with speed to precisely where it is most needed, and acted on with skill and courage.

This article is divided into the following sections:

Nature of intelligence 717
- Types of intelligence that are sought 717
- Types of intelligence in terms of scope 718
- Sources of intelligence 718

A brief history of intelligence activities 718
- Intelligence in premodern periods 718
- The growth of nationalism and intelligence systems 718
- Modern power politics and intelligence systems 718

National intelligence organizations 719
- The United States 719
- The Soviet Union 721
- The United Kingdom 721
- France 722
- China 722
- Israel 722

How intelligence is gathered 722
Policy considerations 723
Bibliography 723

NATURE OF INTELLIGENCE

The various categories of intelligence are potentially endless. The armed services need military intelligence, a space or Earth satellite program needs scientific information, a foreign office needs political or biographical information, and a premier or president needs an amalgam of these and many more. Consequently, intelligence activity has become a vast international industry, most of which goes forward in secrecy. Total U.S. governmental expenditures on intelligence in the mid-1980s, for example, were estimated at around $12,000,000,000 (precise figures are secret). Employed in this work were perhaps 200,000 persons in the United States and many thousands more overseas. The intelligence activities of the Soviet Union were likely of greater dimensions, and all major nations maintain large intelligence bureaucracies.

Types of intelligence that are sought. Political intelligence is at once the most sought after and least reliable of the various types. Obviously, no one can predict what the outcome of political forces in a foreign country will be, and the analyst is reduced to making forecasts of alternatives based on what he knows about political trends and patterns. His most concrete data include such things as voting trends, the details of party organization and leadership, and sophisticated analyses of political documents. A prime source of political intelligence has long been the reports of diplomats, who normally gather data from open sources in the country where they are stationed. Their work is supplemented by that of the professional intelligence apparatus. Much military intelligence is gathered by military attachés, who have formal diplomatic status but are known to be chiefly concerned with intelligence. Ironically, military information is often more difficult to obtain in peacetime than during war, although space reconnaissance technology has greatly expanded collection capacity. In wartime much is learned from captured equipment, captured enemy forces, and information gained from patrols. The kinds of information considered to be of greatest value pertain to military organization and equipment, procedures and formations, and the number of units and total personnel.

Sources of information

Economic factors are crucial to a nation's military strength, its political development, and the conduct of its foreign policy. Unemployment in Japan, a food shortage in India, rice production in Southeast Asia, a balance of payments crisis in the United States or the United Kingdom—each has obvious relevance to the conduct of foreign policy. Intelligence organizations, therefore, give major importance to the collection of data on trade, finance, natural resources, and industrial capacity, and to complex analyses such as the estimation of gross national product.

Technological advance has produced an ongoing race between military capabilities and defensive measures, and between new methods of collecting intelligence and new techniques of protecting secret information. Consequently, intelligence systems must be on guard against scientific or technological breakthroughs that may give one country a decisive advantage over another. They must be kept abreast of foreign advances in nuclear energy, in the electronic, chemical, and computer sciences, and along all other frontiers of scientific knowledge.

If an intelligence system is to make predictions of a foreign nation's future behaviour, it obviously needs detailed information about the personal characteristics of national decision makers. The scope of biographical research has broadened with the growth of international organizations, since negotiators need to be briefed with information about their foreign counterparts.

Intelligence agencies constantly work at compiling data on foreign populations, topographies, climates, and an increasingly wide range of ecological factors. The questions asked of the intelligence analyst may range from "What is the nature of the water supply in Addis Ababa?" to "What are the prospects for economic growth next year in Outer Mongolia?"

Three categories of intelligence

Types of intelligence in terms of scope. Intelligence is often classified in three categories: strategic (sometimes called national), tactical, and counterintelligence. Of these, the broadest in scope is strategic intelligence, which covers information about the capabilities and intentions of foreign nations. Tactical intelligence, sometimes called operational or combat intelligence, is information required by field commanders. The distinction between strategic and tactical may be vanishing, because in an age of rockets, missiles, and jet aircraft a major field commander may require much the same kind of information as a president and his national security council, and vice versa.

Counterintelligence is information and activity related to protecting one's own information and the secrecy of one's own intelligence operations. Its purpose is to prevent spies or other agents of a foreign power from penetrating one's government, armed services, or intelligence agencies. Counterintelligence is also concerned with protecting a nation's high technology, deterring terrorism, and coping with international narcotics traffic. Counterintelligence operations sometimes produce positive intelligence. The adversary's efforts to penetrate one's own security may give clues as to what information the adversary needs, as well as information about his techniques, equipment, and operational procedures. Opportunities are sometimes offered by counterintelligence to manipulate the adversary's intelligence service. Moles and double agents are part of the real world of counterintelligence.

Sources of intelligence. While the public sees the intelligence operative as a cloak-and-dagger secret agent, in fact the largest amount of intelligence work is an undramatic search of public sources. This includes the monitoring of foreign radio broadcasts, the analysis of the contents of foreign publications of all kinds, and the sifting of reports from diplomats, businessmen, accredited military attachés, and other observers. Most intelligence work is performed by university-trained research analysts in quiet offices.

Covert sources of intelligence are more dramatic, falling into three major classifications: aerial and space reconnaissance, electronic eavesdropping and code breaking, and the secret agent working at the classic spy trade. Broadly speaking, the value of each as a source of crucial information is probably in descending order as listed. This is because a photograph constitutes hard (*i.e.*, reliable) intelligence, whereas the report of a secret agent may be speculative and difficult to prove and therefore soft.

A BRIEF HISTORY OF INTELLIGENCE ACTIVITIES

Intelligence in premodern periods. The earliest counterparts of the modern intelligence agency were perhaps the ancient soothsayers, such as the Delphic oracle, who claimed to be able to communicate with the gods and therefore often were said to have the power of predicting the future. As in modern times, their reports were often ambiguous and often ignored by the decision makers. The Bible reports God's advice to Moses (Num. 13) that he send agents to "spy out the land of Canaan." Twelve agents were sent who returned after 40 days to report that the people in the land flowing with milk and honey were more powerful than the Israelites—for which intelligence they were punished by God.

Another ancient commentator on intelligence and counterintelligence was Sun-tzu, whose *Ping-fa* (*The Art of War*), written about 400 BC, is said to be widely read by contemporary Chinese Communist strategists. Sun-tzu refers in his book to five kinds of secret agents, corresponding to modern concepts of agents in place, double agents, deception agents, expendable agents, and penetration agents. Sun-tzu stressed the importance of good intelligence organization, and he also wrote of counterintelligence and psychological warfare.

In the Middle Ages

In the Middle Ages intelligence began to be used systematically in western Europe, although it was crudely organized. On the one hand, it was usually impossible to conceal the massing of troops or ships, making strategic surprise difficult to achieve in military operations; on the other hand, communication was slow, and thus a fine balance often existed between the information available and the ability to communicate it in time.

In the 15th century the Italian city-states began to establish permanent embassies in foreign capitals. The Venetians used such outposts as intelligence sources and even developed codes and ciphers by which information could be secretly communicated. By the 16th century other European governments had followed suit.

From the Renaissance to the Franco-Prussian War

The growth of nationalism and intelligence systems. The rise of nationalism, punctuated by the Peace of Westphalia in 1648, saw the growth of standing armies and professional diplomats. These brought a need on the part of statesmen for evaluated foreign information, and so organizations and procedures for its procurement were developed. Queen Elizabeth I maintained a notable intelligence organization. Her principal state secretary, Sir Francis Walsingham, developed a network of several score intelligence agents in foreign lands. He recruited graduates of Oxford and Cambridge, developed the black arts of espionage and foreign political intrigue, and created tools and techniques for making and breaking codes. Later the cardinal de Richelieu and Oliver Cromwell, whose intelligence chief, John Thurloe, is often cited as an early master spy, developed notable intelligence systems.

Not until the late 18th century, however, did sharp organizational distinctions arise between internal security (a counterintelligence function) and external foreign intelligence. The populace began to give its allegiance to the state rather than to dynasties or religious leaders. As a result, national leaders concerned themselves increasingly with foreign public opinions, producing both a new diplomacy and new intelligence needs.

Major innovations in organization and doctrine have been credited to Frederick the Great. Under Frederick, and later under Bismarck's aide Wilhelm Stieber, the Prussians organized the intelligence function as an integral part of the general staff. This called for a single military intelligence agency to serve as the nation's eyes on the outside world. Under Stieber perhaps the first large-scale espionage system emerged.

Modern power politics and intelligence systems. European power politics, the pursuit of imperialist foreign policies, and advances in military and communications technology required an increasing amount of strategic intelligence. Intelligence bureaus spread throughout Europe, producing a corresponding growth in counterintelligence.

During World War I

Nevertheless, most European nations entered World War I with inadequate intelligence services. That war is now often cited as one that none of the combatant nations intended, suggesting in itself a tragic intelligence failure. The French intelligence service had been torn by internal intrigue, after having already been weakened by the Dreyfus affair, and other services had been shaken by scandals. Miscalculation of German military strength in 1914 was a prime example of intelligence failure.

German intelligence had also deteriorated. The German General Staff of 1914 evidently placed little faith in the in-

formation supplied to it. For the Germans, policy seemed to dominate intelligence.

The Russians had great initial success against the Austrians because of the treason of an Austrian general staff officer, but subsequently Russian intelligence performed no better than that of other nations in World War I. The British were successful in breaking German naval codes and in their Middle Eastern intelligence operations. The Germans carried on successful activities in Persia and scored limited espionage successes in the United States. The United States had no central intelligence. At the beginning of the war army intelligence was a small section within the general staff, comprising two officers and two clerks; by the war's end, this service had grown to 1,200 officers and civilians. Patently, most were amateurs. Navy intelligence was equally deficient.

The intelligence lessons of World War I, along with advances in technology, especially electronics and aircraft, resulted in a proliferation of intelligence agencies in the 1920s and 1930s. The growth of such agencies was further stimulated by the advent of regimes with expansionist foreign policies in the U.S.S.R., Italy, Germany, and Japan. These produced, in reaction, counterintelligence organizations in the democracies.

World War II

World War II led to the creation and expansion of intelligence services everywhere. The United States, which had virtually no peacetime intelligence services, created in the Office of Strategic Services (OSS) its first full-fledged organization for intelligence and secret operations. The war imposed intelligence requirements never before faced by the major warring powers. This was primarily the result of accelerating technology. Air warfare in particular required vast new offensive and defensive intelligence operations, and the growth of radio broadcasting produced a new art of psychological warfare that demanded intelligence services for the analysis of its effects. Even with all the new developments, intelligence forecasting remained a precarious trade. The Pearl Harbor attack in December 1941, the Battle of the Bulge in 1944, the unexpected German resilience under Allied bombing attacks—in these and other instances the decision makers failed to profit from their elaborate intelligence networks.

But there was one area of enormous profit. Perhaps the most significant intelligence exploit of all times was the Ultra secret—in which the British, with prior help from the Poles and the French, and possessing a German Enigma encoding machine, broke the German military code and read most of the signals between Hitler and his field commanders throughout the war. Put simply, the Ultra operation involved intercepting German signals that had been mechanically enciphered by an Enigma machine, decoding the messages that the German high command assumed were impenetrable, and distributing the information secretly to appropriate headquarters. In essence, throughout much of the war the Allied side was reading the mind of the Axis command. As the war progressed, Hitler's increasingly centralized control of operations on all fronts made German military operations especially vulnerable. On those occasions when the Allies were caught by surprise, such as the American defeat at Kasserine Pass, the Allied defeat at Arnhem, and the Battle of the Bulge, either the Germans had used land lines for communication, or Ultra intercepts had been misused.

World War II was followed by the Cold War, in which the intelligence organizations of opposing blocs became combatants. In every major nation huge new bureaucracies were created. Usually these consisted of interlocking and often competitive secret agencies, vying for new assignments and sometimes withholding information from each other. The world soon became familiar with the Central Intelligence Agency (CIA) of the United States. Somewhat less well known were Great Britain's MI-5 and MI-6, the KGB and the GRU of the Soviet Union, the SDECE of France, China's Social Affairs Department, and the Mossad of Israel. At the same time, the exploits of spies and counterspies became a subject for mass media fiction. Books, movies, and television moved the spy to centre stage, sometimes in a comic but often in a deadly serious role. All of this tended to glamorize what is most often a painstakingly tedious, or even at times a disgustingly immoral and distasteful, occupation.

NATIONAL INTELLIGENCE ORGANIZATIONS

In a world in revolutionary ferment, the authentic intelligence officer occupies the centre of great debates over national security policy. At issue in most of the debates are questions of power, probability, and time. A prime task of the modern professional intelligence officer, military or civilian, is to try to answer questions for the policymaker about power and about behaviour probabilities, within a time scale. For a chief of state trying to decide a question about nuclear armaments, for example, an ideal intelligence system would provide precise knowledge of a potential enemy's power, the probability of that enemy's behaviour or reaction in given contingencies, and a time schedule for the most likely sequence of events.

Policy questions

These are basic, even classic problems for all intelligence services. Information as to how these services address their problems is highly uneven. More is generally known about the U.S. system than any other, a good deal about that of the Soviet Union, and comparatively less about other systems. Intelligence systems fall into three general types: the U.S., which has been followed by nations such as West Germany, Japan, and South Korea since they came under U.S. influence after World War II; the Soviet, which is imitated in large measure by most Communist-governed nations; and the British, on which are patterned the systems of most nations with true parliamentary governments.

The United States. The Central Intelligence Agency (CIA) was created in September 1947. The CIA grew out of the nation's wartime experience with the Office of Strategic Services (OSS) and a postwar decision to create a central organization for defense. This included a partially unified Department of Defense and a National Security Council (NSC), chaired by the president. The CIA is under the jurisdiction of the NSC.

World War II had shown that there was a need for a central organization to coordinate for the president all information on foreign affairs and to perform certain functions best done centrally. There were heated debates at the war's end as to how much centralization was needed. Some wanted a central intelligence setup that would eliminate Army, Navy, State Department, and other separate units. Others wanted to turn over all but technical military intelligence functions to the State Department. The outcome was a compromise in which the Central Intelligence Agency was created by the National Security Act of 1947, but other departments and agencies continued to maintain their own intelligence sections. Since then the idea of a single entity has given way to the concept of an "intelligence community" comprising the CIA, the Defense Intelligence Agency (DIA), separate Army, Navy, and Air Force intelligence staffs, State Department intelligence, the National Security Agency (NSA), a Department of Energy nuclear intelligence unit, and the Federal Bureau of Investigation (FBI). The National Security Act of 1947, which has remained the basic charter for the organization, assigned to the CIA five specific functions: (1) advising the National Security Council on intelligence matters related to national security; (2) recommending to the NSC measures for efficient coordination of the intelligence activities of departments and agencies of government; (3) correlating and evaluating intelligence and seeing that it is properly communicated within the government; (4) carrying out for existing intelligence agencies any additional services the NSC determines can best be performed centrally; and (5) carrying out such other functions and duties related to national security intelligence as the NSC may direct.

The Central Intelligence Agency

Out of this assignment of basic functions grew a large new agency with assignments around the globe. By the 1980s the CIA was believed to employ about 18,000 full-time employees in the United States, mainly in Washington, D.C., and several thousands more in overseas posts. Its annual budgets do not include functions performed by military and paramilitary units under the CIA's direction and control. Policy and operational guidelines for the CIA are contained in periodically revised presidential executive orders and some two-score secret National Security

Council Intelligence Directives defining its functions and establishing jurisdictions in areas in which other intelligence agencies might have a functional claim.

Since 1947 the CIA has come to perform three major assignments: (1) foreign intelligence collection, evaluation, and communication; (2) counterintelligence operations overseas; and (3) secret political intervention, psychological warfare, and paramilitary operations in foreign areas. The first two functions are clearly stipulated in the congressional statute creating the CIA; the third was assumed as a Cold War necessity and on the basis of a very free interpretation of the original charter.

The CIA is managed by a director and a deputy director, both appointed by the president and subject to Senate confirmation. The director of Central Intelligence in fact wears two hats. He is head of the Central Intelligence Agency as well as being the president's principal intelligence adviser, and thus theoretically the number one man in the intelligence community—theoretically, because each of the other major units, such as the Federal Bureau of Investigation, the National Security Agency, or the Defense Intelligence Agency, operates somewhat autonomously. The head of the CIA is, in fact, "first among equals," and his authority has increased over the years.

The major divisions of the CIA are those for intelligence, operations, science and technology, and administration. The intelligence division, the largest, is responsible for the production of finished intelligence in its various forms—from instant news flashes to encyclopaedic surveys. Most of the thousands of workers in the intelligence division are technical specialists or research analysts. By contrast, the operations directorate, sometimes called the "Department of Dirty Tricks," has handled various clandestine operations, such as espionage missions, counterintelligence overseas, political intervention, recruitment of defectors, and a variety of covert actions in foreign areas. The science and technology division concerns itself both with new techniques of intelligence gathering and with the technological developments of other nations that may have impact upon U.S. security. The administration directorate is concerned with recruitment and training, library services, including information storage and retrieval, and various other housekeeping activities.

The CIA performs the difficult dual role of independent intelligence producer and coordinator of a vast intelligence system comprising many units. The CIA produces intelligence in a variety of forms, from bulletins of current information, such as the National Intelligence Daily for top policymakers, to long-range forecasts. Under the CIA's leadership, the National Foreign Intelligence Board (NFIB) has responsibility for the production of National Intelligence Estimates (NIE), supposedly representing the best information available throughout the government on such questions as "What is the current and future status of Soviet civil defense?" The NFIB, representing all the major units of the intelligence community, includes what was formerly called the Board of National Estimates, which has been replaced by national intelligence officers with geographical and functional jurisdictions. National Intelligence Estimates have drawn mixed reviews for objectivity, accuracy, length, style, and timeliness. A National Foreign Intelligence Council operates as a companion to NFIB to deal with priorities and budgets.

The FBI's principal role in the intelligence system is in counterintelligence, limited primarily to internal security within the United States. The CIA's counterintelligence role is applied primarily overseas, while the FBI's jurisdiction normally ends at the water's edge. The FBI director operates under the attorney general in the Department of Justice. An assistant director of the FBI heads an Intelligence Division, the budget, personnel, and organization of which are secret. Increasing concern with terrorist activity has involved the FBI's Criminal Division in counterintelligence, and the FBI and CIA are known to cooperate increasingly in some areas. The Defense Intelligence Agency and agencies of the armed services also perform counterintelligence functions within their limited jurisdictions.

The National Security Agency is the largest, most expensive, and perhaps least known of all American intelligence organizations. Its basic functions are signals intelligence—the making and breaking of codes and ciphers. Created by presidential directive in 1952, the NSA has remained, despite its enormous size and worldwide activities, the most secret of the acknowledged U.S. intelligence units. The directive creating the agency remains secret. Headed by a high-ranking military officer, the agency is under the secretary of defense but maintains a degree of autonomy. From its headquarters near Washington, D.C., the NSA conducts an immense variety of electronic espionage activities. Estimates of its size generally set its number of employees at around 20,000, but its activities also involve thousands of armed-service personnel, supervised by the NSA. Devices for electronic espionage of various sorts are placed in planes, on ships, or at ground installations overseas. This kind of intelligence work was brought to world attention when the USS "Pueblo" was seized by the North Koreans in 1968 and its crew detained for a year. The subsequent Navy and congressional inquiries revealed that the ship was on assignment for the National Security Agency, with a mission to collect data on North Korean and Soviet naval operations in the Sea of Japan.

The National Security Agency

The Defense Intelligence Agency is the newest major member of the U.S. intelligence community. It was established in 1961 as a culmination of the increasing centralization in the Department of Defense, which began with the unification measures of 1947. The DIA was set up by the secretary of defense to improve the intelligence product available to him and to the Joint Chiefs of Staff. Another purpose was to eliminate costly duplication and a tendency of the separate armed services to use their own intelligence data as ammunition in the annual competition for budget allocations.

The DIA is the foreign military intelligence arm of the secretary of defense, the Joint Chiefs of Staff, and the unified and specified military commands. It supplies military intelligence for national reports and estimates, coordinates Defense Department collection requirements, and manages the defense attaché system. Persons assigned to the DIA come from each of the armed services, and more than half of the DIA staff are civilians.

The Department of State, through its Bureau of Intelligence and Research, collects, analyzes, and disseminates large amounts of political, economic, and cultural information about those nations in which the United States has accredited representation. The bureau, referred to as INR in the intelligence community, operates under a director whose rank since 1963 has been that of assistant secretary of state. The bureau has the double function of meeting the requirements of the intelligence community as set by the National Security Council and the State Department's own intelligence needs. Area specialists constitute the bulk of the INR staff, which is small by comparison with those of other intelligence agencies.

The Department of Energy has representation in the intelligence community through an assistant secretary for defense programs, whose responsibilities include the nuclear intelligence role of the former Atomic Energy Commission. Intelligence functions are the responsibility of a deputy assistant secretary for security affairs and a director of the Office of International Security Affairs, who oversee the collection of foreign energy intelligence and its dissemination within the government.

Military intelligence services

The creation of the DIA sharply reduced the role of the separate armed forces in the high-level intelligence system. Yet each of the armed services maintains a major intelligence division for tactical and technical intelligence and for counterintelligence activities.

As organized in 1985, Army intelligence comprised six directorates: Intelligence Resources Management; Foreign Liaison; Counterintelligence (the Army's traditional "CIC" security service); Intelligence Systems, made up of departments for long-range planning, signals intelligence, human intelligence, imagery, and intelligence systems integration; Intelligence Automation Management; and Foreign Intelligence, which includes departments for requirements, scientific and technical, and intelligence, which is subdivided along geographical lines. Also attached to the Intelligence directorate are a "Red Team," a field operating agency

giving staff support in war games, and a Current Intelligence Division.

U.S. Naval intelligence has evolved into a complex Office of Naval Intelligence (ONI), with five major divisions: Military Information; Intelligence Research and Development; Intelligence Policy and Estimates; Special Projects; and Plans, Programs, and Systems Architecture. Each division reports to the ONI secretariat, which has staff assistants for such categories as Soviet strategy and doctrine, foreign intelligence agency liaison, intelligence community coordination, and routine management duties. The deputy director of ONI for plans, to whom this entire structure reports, serves also as commander of the Naval Security Group Command, which has code-making and code-breaking responsibilities. In addition, the Naval Intelligence Command, under the other deputy director of ONI, handles staff support for all other areas except codes and ciphers.

The Air Force, like the Navy, has been accorded by technology a front-line intelligence position. The assistant chief of staff, intelligence, is charged with the responsibility for planning, directing, and supervising Air Force policies and organizations covering a worldwide range of intelligence activities. A special responsibility is to be a sentinel against technological, strategic, or tactical surprise. These various functions are performed under two main directorates. One is the Division of Estimates, which supervises departments for weapons, space, and technology; strategic threat; regional estimates; and general-purpose threat. The other is the Directorate of Plans and Systems, which oversees a complex set of activities, including groups for management, planning, and collection. There is also an Electronic Combat Intelligence Group, which reports directly to the assistant chief of staff, intelligence. Also coming under Air Force operational control is the National Reconnaissance Office (NRO), one of several highly secret units for surveillance by satellite, including photographic, signal, and electronic intelligence. The existence of NRO is not officially acknowledged, but it has functioned since the early 1960s, and its importance and size grow with advancing surveillance technology. It is perhaps the most expensive, and one of the most useful, sources for intelligence.

The Soviet Union. The Committee of State Security (KGB [Komitet Gosudarstvennoy Bezopasnosti]) resembles the American CIA and FBI and Secret Intelligence Service rolled into one. This combination of foreign intelligence, counterintelligence, and internal security roles is unusual, although the Soviet system has set the pattern for intelligence services in other Communist countries.

From Cheka to KGB

The lineage of the KGB began with the Cheka, the secret police established by the Bolsheviks in 1917. In 1922 the Cheka was reorganized into the GPU (State Political Administration) and in 1934 was renamed the NKVD (People's Commissariat for Internal Affairs). During World War II several further reorganizations occurred, out of which grew the MGB (Ministry of State Security).

The current organization, the KGB, was created in 1954. It is commonly believed to dominate the entire Soviet intelligence system, and some analysts picture the KGB chief as a man of immense potential power in the political sphere. (Yury Andropov, who had headed the KGB for 15 years, served as Soviet party chief from 1982 to his death in 1984.) Other separate intelligence agencies do exist in the U.S.S.R., including that of the Foreign Ministry, but most particularly the Central Intelligence Office (GRU [Glavnoye Razvedyvatelnoye Upravleniye])—the chief intelligence directorate of the Army general staff. The GRU deals primarily with military intelligence, and, although at times there have been indications of competition and conflict with the KGB, the latter dominates.

The KGB is organized into a number of major directorates. The First Chief Directorate carries out primarily counterintelligence missions. This pervasive section of the KGB performs an internal security function similar in some ways to that of the American FBI but not limited to domestic jurisdictions.

The Second Chief Directorate is responsible for foreign intelligence and is believed to have a wide variety of subsections dealing with particular areas (such as an American section) and specific functions (such as psychological warfare in foreign areas).

The Third Chief Directorate is believed to encompass Smersh, popularized in the 1960s by the British novelist Ian Fleming. The real Smersh—derived from the Russian *smert shpionam*, meaning "death to spies"—has as its principal assignment maintaining security within the armed forces and watching for potential traitors within the military and intelligence services.

There are a number of other chief directorates, each with specialized intelligence functions and particular geographical jurisdictions. Many Soviet officials serving abroad have some direct connection with the KGB or the GRU, and it has been alleged that some Soviet ambassadors have held rank within the KGB. Soviet diplomats assigned to the United Nations occasionally have been discovered to be intelligence agents, but the same is assumed to hold true for the diplomats of most other major nations.

Estimates of the size and annual cost of Soviet intelligence are speculative. Since the KGB encompasses internal security, border guards, and foreign-intelligence personnel, the total cost probably runs into billions of dollars, with perhaps several hundred thousand employees, not counting paramilitary detachments; clearly most of these are involved with domestic intelligence. But there is no reliable information on these matters.

The United Kingdom. British intelligence was organized along modern lines as early as the days of Queen Elizabeth I, and the long British experience has influenced the structure of most other systems. Unlike those of the United States and the Soviet Union, British intelligence agencies have preserved through most of their history a high degree of secrecy concerning their organization and operations. Even so, Britain has suffered an unusually large number of native spies within the intelligence establishment.

The two principal British intelligence agencies are the Secret Intelligence Service (SIS; also known by its wartime designation, MI-6) and the Security Service (commonly called MI-5). The labels derive from the fact that the Secret Intelligence Service was once "section six" of military intelligence and the Security Service, "section five."

Today MI-6 is a civilian organization with functions resembling those of the American CIA. The British intelligence community, however, is even more of a confederation of separate agencies than is that of the United States. The Secret Intelligence Service is charged with gathering information overseas and with other strategic services ranging from foreign espionage to covert political intervention. The director of the SIS is commonly referred to as "C" and remains an anonymous figure in British governmental life. A high wall of secrecy likewise surrounds his organization; the government barely acknowledges its existence, although an annual lump-sum appropriation request must be presented publicly to Parliament. The British services are much smaller, perhaps by a factor of 20, than those of either the United States or the Soviet Union.

Security Service

Also included in the annual budget submitted to Parliament is the Security Service (MI-5). This is roughly equivalent to the American FBI, or the internal security section (counterintelligence) of the Soviet KGB. It differs from the FBI in performing certain counterintelligence functions overseas. Its main charge is to protect British secrets at home from foreign spies and to prevent domestic sabotage, subversion, and the theft of state secrets. The Security Service is headed by a director general, whose code name in years gone by was "K"—a designation derived from the name of Sir Vernon Kell, its head from 1909 to 1940. The Security Service makes no direct arrests but rather works secretly behind the more publicized "Special Branch" of Scotland Yard. The director of the Security Service reports to the prime minister through the home secretary.

The supervision of the Secret Service is the duty of the Joint Intelligence Committee, a Cabinet subcommittee under the permanent undersecretary of the foreign office. Another layer of control is exerted by the Cabinet secretary in the case of the Secret Service and by the permanent head of the Civil Service in the case of the Security Service.

Another principal member of the British intelligence community is the Defence Intelligence Service, resembling

the American Defense Intelligence Agency. This integrates intelligence specialists of the Royal Army, Navy, and Air Force within the Ministry of Defence. Still another service is Communications Intelligence, an organization for electronic surveillance and cryptology. Its operations are concentrated in what is called the Government Communications Headquarters (GCHQ) at Cheltenham.

The Joint Intelligence Committee combines all of the British intelligence agencies and performs functions of overall policy control and of approval of "national estimates" similar to those carried out by the U.S. National Intelligence Council. The British Cabinet and parliamentary government affords a system of accountability lacking in the United States.

France. Contemporary French intelligence and counterintelligence may be described as an amalgam of units dating from Napoleon Bonaparte and a system developed by Gen. Charles de Gaulle as leader of the Free French in World War II. From 1947 until 1981 the major French intelligence service was known as the Service de Documentation Extérieure et de Contre-Espionnage (SDECE). Colloquially it was termed "The Pool," referring to its mixture of military and civilian officers.

In 1981 SDECE was reorganized as the Direction Générale de la Sécurité Extérieure (DGSE). The structure also changed, but the agency's traditional functions remained: collection and analysis of foreign information, counterespionage outside of France, and overseas covert political actions. Because of France's former colonial ties to Africa, it is widely believed that the bulk of French intelligence operations and covert action have centred in West Africa.

Another major French intelligence agency is the Second Directorate of the National Defense Staff, combining, to some degree, formerly separate Army, Navy, and Air Force specialists. The traditions and doctrines of the old French Army's Deuxième Bureau (Military Intelligence) no doubt influence the general staff's Second Directorate. Its function is to gather foreign military intelligence. A third important member of the French intelligence system is the DST (Direction de la Sécurité du Territoire), charged with internal security. It is the French equivalent of the American FBI and operates under the administrative control of the Ministry of the Interior.

The SDECE was shaken in 1968 by the publication of the memoirs of Philippe Thyraud de Vosjoli, who had been an important officer in the French intelligence system for 20 years. De Vosjoli asserted that SDECE had been deeply penetrated by the Soviet KGB in the 1950s, an allegation that the government denied. He also indicated that there had been periods of intense rivalry between the French and American intelligence systems. There have been other indications that SDECE experienced a number of major intelligence failures, structural reorganizations, and changes of leadership in the period between 1950 and 1980.

China. The People's Republic of China has an intelligence system that resembles closely the system of the Soviet Union, although not much is known about it. Less distinct lines exist in China between the professional intelligence function and Communist Party administrative functions. The best information suggests that the intelligence organization consists of four major units: those of the Party, the State Council (the governmental bureaucracy), the military establishment (Defense Ministry), and the Ministry of Foreign Affairs.

Dominance of the party

Under the title Social Affairs Department, the intelligence unit of the Central Committee of the Communist Party attempted for some years to exert control over the entire intelligence system, much in the fashion of the Soviet KGB. This department, with perhaps 25,000 employees, is thought by some to have been abolished during the Cultural Revolution of the late 1960s. It is probable that only its name was changed. As in the U.S.S.R., the Army—through the Defense Ministry's large intelligence staff—is potentially a strong competitor of the civilian intelligence agencies. The Army chief of staff in China also runs the military secret police and heads the Political Department of the Defense Ministry. Clearly there are two main functions within the military: first, to assure the reliability and loyalty of the Chinese military establishment—in effect, an internal security function—and second, to secure positive intelligence about foreign military systems. The State Council operates a Ministry of Public Security that concerns itself mainly with internal security. Finally, there is the Foreign Intelligence Department in the Ministry of Foreign Affairs, responsible for intelligence gathering and other overseas activities, including covert political action. Its agents operate, like those of most other nations, through embassies, legations, the New China News Agency, the overseas Chinese, and other obvious channels.

Outside observers have speculated about the quality of Chinese foreign intelligence, and some have suggested that Peking's leaders are poorly informed about the outside world as a consequence of a party-dominated intelligence system reporting what its leaders want to hear. If this is true, it is but an exaggeration of a fault common to all intelligence systems. China's decreasing isolation from world politics, beginning in the early 1970s, will likely increase the sophistication of Chinese foreign intelligence.

Israel. The state of Israel, since its creation in 1948, has met its obvious need for intelligence and counterintelligence with services that have gained a first-class reputation. One mark of their professionalism is that less is known about them than about other systems. They constitute five major organizations.

The Central Institute for Intelligence and Security has a director who is chairman of a committee presiding over the entire Israeli intelligence system. The Central Institute operates primarily in foreign areas. Popularly known as Mossad, it carries out foreign espionage and covert political operations. The numerous Israeli secret agents living in Arab lands, for example, are under the supervision of Mossad. Its director is responsible to the prime minister.

Mossad and Shin Bet

A second agency is the Shin Bet, after the Hebrew initials for General Security Services. Its main function is internal counterintelligence. Shin Bet is an investigative agency concerned with potential sabotage, Arab terrorist activities, and security matters having a strong political flavour. A third major Israeli agency is the Intelligence Corps of the Defense Forces, commonly referred to as Military Intelligence. Some see it as a rival to Mossad, and conflicts between them have been reported. Its chief is the military intelligence adviser to the defense minister. The Research Department of the Foreign Ministry is a fourth agency. This relatively small unit focusses on political information, including policy-oriented studies and the evaluation of information supplied through diplomatic channels. The fifth member of the system—the Special Investigations Department of the Israel Police Force—is concerned with criminal investigations inside Israel's boundaries.

Israel's successes in wars with Arab countries may be attributed in some measure to the efficiency of its intelligence system. But it has not been without its scandals. The most dramatic was the "Lavon affair," which led to the resignation of Prime Minister David Ben-Gurion. This is said to have been an attempt in 1954 to blow up American and British information offices in Egypt in order to arouse Western public opinion against the Egyptians. But for the most part the secret services of Israel have operated without exposure and with frequent success.

Israeli intelligence has the advantage of a cosmopolitan population, providing persons whose language and experience enable them to pose as citizens of Chicago, Munich, London, Rome, or even Cairo. Recruitment is thought to be from three main sources: the Haganah, the one-time illegal Jewish army in Palestine, with its allied intelligence agency; ex-members of the Irgun and Stern groups that also operated in the prestate era; and the newer immigrants.

HOW INTELLIGENCE IS GATHERED

Good intelligence management must begin with the determination of what needs to be known. Unless precise requirements are set, data will be collected unsystematically and the decision maker left, in the end, without the required information. Once data are collected they must be evaluated and transformed into a usable form (and sometimes stored for future use). Evaluation is essential because of the wide variety of sources, many of them of doubtful reliability. A standardized system is used. Source

reliability may be rated A, B, C, etc., down a scale. Information accuracy may likewise be rated 1, 2, 3, etc., with 1 representing confirmed information and 3 denoting a "possibly true" degree of accuracy. Thus the classification of a bit of information as C-3 would mean it came from a fairly reliable source and was possibly true.

With organizational sophistication, refined procedures, and the most advanced technology, intelligence still falls into two categories—the knowable and the unknowable. It is possible to obtain hard intelligence on the number of Soviet intercontinental ballistic missiles. But it is not possible to predict, with precision, the political intentions of Soviet leaders.

Most valuable information comes from overt sources, that is to say, from sources accessible without resort to illegal methods. Open sources probably contribute more than 80 percent of the input to most intelligence systems. This varies according to the subject country, since clearly there are far fewer state secrets in some countries than in others. Foreign mass media, including press, radio, and television, are a standard source, and they require a massive effort in recording, translating, and storing information.

Covert sources and clandestine collection methods provide the basis for much of the drama and romance accorded intelligence work in fiction. The classic espionage agent is now obsolescent. Some have suggested that the place of the human agent on the ground is being taken by inanimate operators such as orbiting reconnaissance satellites, long-range cameras, and a variety of sensing, detecting, and acoustical instruments. It is now possible to see and hear at a distance in the night and to photograph from an altitude of hundreds of miles. A race has developed between the use of technology to uncover secrets and its use to make secrets more secure.

Techniques of reconnaissance have advanced dramatically since the 1940s, when the Soviets accused the United States, accurately, of drifting balloons carrying special cameras across Soviet territory to photograph military and industrial installations. The Peking Military Museum has had on display American drone aircraft, downed over China, that were used for photoreconnaissance missions. The American U-2 aircraft and its higher flying successors are capable of taking photographs of ground installations that experts can read with great accuracy. Orbiting satellites, continuing advances in photographic capabilities and sensing devices, and the growth of computer technology have radically changed intelligence and counterintelligence capabilities. The USS "Pueblo," captured by the North Korean government in January 1968, might accurately be described as an electronic scavenger sent to collect information about North Korean radio communications procedures, Soviet naval equipment and procedures, and other kinds of intelligence such as Soviet submarine signatures—telltale and unique signals permitting their later identification. Computers can be used to analyze complex data on industrial production, missile launchings, and economic growth rates. High-altitude photographs taken from aircraft or satellites yield information on crops, floods, drought areas, or even plant diseases, which in turn can be used to assess the economic prospects of a country.

Use of aircraft and satellites

So vast is the intake of information that intelligence systems have been threatened by inundation. In the years since World War II great efforts have gone into the development of efficient means for cataloging, storing, and retrieving on demand the gigantic volume of data being amassed. Intelligence agencies have led in developing new indexing and codifying techniques and machinery for the electronic analysis, storage, and location of specific information. Some believe that data collection has been overemphasized at the expense of analysis, but silicon chips and microcircuitry hold promise for coping with the tidal wave of information.

POLICY CONSIDERATIONS

Since World War II, intelligence has come to constitute one of the world's largest industries, employing hundreds of thousands of professionals. Annual expenditures for governmental and nongovernmental intelligence run into many billions of dollars. Under contemporary political and technological conditions, intelligence systems are likely to grow in size and scope. They have already taken on assignments that go beyond their informational and security functions, such as secret political and paramilitary operations. These activities will probably continue, since they are preferable to overt warfare between nuclear powers.

Among the superpowers, intelligence services have probably outgrown their optimal size. Secret bureaucracies tend to grow beyond the point of efficiency, especially when they can argue that the fate of the nation depends upon their budgetary health. There is also the problem of control. In some respects this is similar to the problem of civil–military relations, aggravated by the secrecy factor. To put the matter simply: knowledge is power, and secret knowledge is secret power. Historically, secret services have often been vehicles of conspiracy and political intrigue. Control is difficult to apply, in both democracies and authoritarian systems, since a secret service can logically demand secrecy for its operations. In an age of accelerating technology, it seems quite likely that the power of intelligence systems will grow vis-à-vis political authority. Presidents and prime ministers, to avoid becoming prisoners of their intelligence systems, will have to be alert to the need for effective policy controls.

BIBLIOGRAPHY

General works and bibliographies: GEORGE C. CONSTANTINIDES, *Intelligence and Espionage: An Analytical Bibliography* (1983), an authoritative bibliography with substantial annotations, covering works published to 1981; TYRUS G. FAIN (ed.), *The Intelligence Community: History, Organization and Issues* (1977), a reference work containing most of the important documentation, dealing primarily with the 1970s; ROY GODSON (ed.), *Intelligence Requirements for the 1980s,* 5 vol. (1979–82), symposium papers and comments on the whole range of questions, including intelligence, counterintelligence, and covert action; DAVID KAHN, *The Codebreakers: The Story of Secret Writing* (1967), a standard work on cryptanalysis, and *Hitler's Spies: German Military Intelligence in World War II* (1978); WILLIAM V. KENNEDY *et al., Intelligence Warfare* (1983), an authoritative catalog of modern intelligence technology; HARRY HOWE RANSOM, *The Intelligence Establishment,* rev. and enlarged ed. (1970), an analysis of the evolution, functions, and problems of U.S. intelligence, with a chapter on the British experience; MYRON J. SMITH, *The Secret Wars,* 3 vol. (1980–81), a comprehensive bibliography of secret operations, loosely defined, covering the period 1939–80.

United States: JAMES BAMFORD, *The Puzzle Palace: A Report on America's Most Secret Agency* (1982), a resourceful journalist's detailed description of the National Security Agency; RAY S. CLINE, *Secrets, Spies and Scholars: Blueprint of the Essential CIA* (1976), a former intelligence professional's description of the evolution of the intelligence system and of his role in its development; WILLIAM COLBY and PETER FORBATH, *Honorable Men: My Life in the CIA* (1978), a candid and sometimes critical account, by a former director of the Central Intelligence Agency, of Colby's intelligence experience; LAWRENCE FREEDMAN, *US Intelligence and the Soviet Strategic Threat* (1977), an analysis of the performance of U.S. intelligence in gauging Soviet military power in the 1960s and 1970s; MORTON H. HALPERIN *et al., The Lawless State: The Crimes of the U.S. Intelligence Agencies* (1976), a detailed summary of the findings of congressional and other investigations of intelligence agency abuses; WILLIAM LEARY (ed.), *The Central Intelligence Agency: History and Documents* (1984), containing a reprint of ANNE KARAKELAS' authoritative history of the CIA and important documents for the years 1947–81; RONALD LEWIN, *The American Magic: Codes, Ciphers and the Defeat of Japan* (1982), an account of the role of signals intelligence in the Pacific in World War II; MARK M. LEWENTHAL, *U.S. Intelligence: Evolution and Anatomy* (1984), a practical organizational description of the U.S. intelligence system; VICTOR MARCHETTI and JOHN D. MARKS, *The CIA and the Cult of Intelligence* (1974, reissued 1983), an expression of disillusionment with the American intelligence system by two former intelligence officers; DAVID C. MARTIN, *Wilderness of Mirrors* (1980), an insightful view of conflict between intelligence and counterintelligence within the U.S. system: THOMAS POWERS, *The Man Who Kept the Secrets: Richard Helms and the CIA* (1979), a journalist's astute analysis not only of Helms's intelligence career but also of many of the intelligence controversies in the United States from 1945 to 1979; HARRY ROSITZKE, *The CIA's Secret Operations: Espionage, Counterespionage and Covert Action* (1977), a former CIA official's account of secret actions; BRADLEY F. SMITH, *The Shadow Warriors* (1983), perhaps the best account to date of

the Office of Strategic Services in World War II; THOMAS F. TROY, *Donovan and the CIA: A History of the Establishment of the Central Intelligence Agency* (1981), a detailed account.

United Kingdom: CHRISTOPHER ANDREW and DAVID DILKS (eds.), *The Missing Dimension: Governments and Intelligence Communities in the Twentieth Century* (1984), essays on intelligence and counterintelligence, covering World Wars I and II and the postwar years; F.H. HINSLEY *et al., British Intelligence in the Second World War: Its Influence on Strategy and Operations,* 3 vol. (1979–84), an official history based on the authors' access to secret archives; R.V. JONES, *The Wizard War: British Scientific Intelligence 1939–1945* (1978), an account of the role and significance of technical intelligence in World War II, by a major participant; RONALD LEWIN, *Ultra Goes to War: The First Account of World War II's Greatest Secret Based on Official Documents* (1978, reissued 1980), a demonstration of how Allied access to Germany's secret communications aided victory in World War II; BRUCE PAGE, DAVID LEITCH, and PHILLIP KNIGHTLEY, *The Philby Conspiracy* (1968, reissued 1981), one of the most detailed accounts, by a team of journalists, of Soviet agent H.A.R. Philby's penetration of the British secret service.

Soviet Union: JOHN BARRON, *KGB: The Secret Work of Soviet Secret Agents* (1974), a standard history, and *KGB Today: The Hidden Hand* (1983); BRIAN FREEMANTLE, *KGB* (1982), a popular descriptive analysis; HARRY ROSITZKE, *KGB: The Eyes of Russia* (1981), a general description by a former CIA professional; VIKTOR SUVOROV (pseud.), *Inside Soviet Military Intelligence* (1984), an insider's view by a former Soviet intelligence professional.

(H.H.R.)

International Law

International law is the body of legal rules that apply between sovereign states and such other entities as have been granted international personality (status acknowledged by the international community). The term was coined by Jeremy Bentham and is synonymous with the term "law of nations" and its equivalents in other languages.

Like precepts of international morality, the rules of international law are of a normative character; that is, they prescribe standards of conduct. They distinguish themselves, however, from moral rules by being, at least potentially, designed for authoritative interpretation by an independent judicial authority and by being capable of enforcement by the application of external sanctions.

International law means public international law as distinct from private international law or the conflict of laws, which deals with the differences between the municipal laws of different countries.

International law forms a contrast to municipal law. While international law applies only between entities that can claim international personality, municipal law is the internal law of states that regulates the conduct of individuals and other legal entities within their jurisdiction.

International law should also be distinguished from quasi-international law, which is the law governing relations similar to those covered by international law but outside the pale of international law because at least one of the parties lacks international personality. Concession agreements between oil companies and sovereign states fall into this category. In case of doubt, they are subject to the municipal law of the state granting the concession.

Transnational law is a purely negative term. It is intended to convey that, in accordance with the intention of contracting parties, a transaction of a consensual character is not or should not be subject to municipal law.

This article is divided into the following sections:

Fundamentals of international law 724
- Views on international law 724
  - The sociological perspective
  - The historical perspective
  - The ethical perspective
- Sources of international law 726
  - The law-creating processes of international law
  - The law-determining agencies of international law
  - International law and municipal law
  - Other basic issues of international law

Rules of international law 727
- Customary international law 727
  - Sovereignty
  - Recognition
  - Consent
  - Good faith
  - Freedom of the seas
  - International responsibility
  - Self-defense
- International rules governed by treaties and other agreements 728
  - Territory
  - Outer space
  - Diplomatic law and immunity
  - The protection of nationals abroad
  - Freedom of commerce and navigation
  - Extradition and asylum
  - Succession to international rights and obligations

The role of international organizations 730
Bibliography 731

Fundamentals of international law

VIEWS ON INTERNATIONAL LAW

A view of international law in three complementary perspectives assists in the better understanding of the subject. The sociological perspective offers an explanation of the social functions fulfilled by international law. The historical perspective provides insight into the growth potential of international law. The ethical perspective furnishes a normative measuring rod by which to test the moral adequacy of any particular system.

The sociological perspective. Law is primarily an outgrowth of a specific social environment. In particular, this applies to law in unorganized international society as well as in specific international societies organized on the confederate model, such as the League of Nations and the United Nations. The chief participants—the sovereign states, and especially the strongest among them—tend to view themselves as ultimate ends and are inclined to insist on control of the means indispensable for their survival in any crisis, especially their freedom to arm themselves. They form alliances and counteralliances for aggressive and defensive purposes, create precarious systems of balance of power, and pursue policies of involvement or isolation.

In such situations the primary function of law is to legitimate power; that is, the law assists in maintaining the supremacy of force and the hierarchies established on the basis of power and gives to such quasi-orders the respectability and sanctity of law. International law in unorganized international society serves these purposes in a variety of ways; for example, one of the cornerstones of international customary law is the independence of states, which provides for freedom of armament, access to raw materials and markets, and the admission of immigrants. Similarly, whether a state decides to participate in an international congress or conference depends on its own will. Moreover, in the absence of agreement to the contrary, unanimity is required for any decision reached in the assembly of any such international gathering. Fi-

nally, any binding third-party settlement of a dispute by reference to law or equity depends on the consent of the parties concerned.

By building international customary law on the foundation of state sovereignty, states make certain of reserving for themselves the choice between peace and war. Moreover, international customary law puts at the disposal of its subjects the right to apply measures short of war by way of reprisal against alleged breaches of international law.

In international confederations, such as the League of Nations and United Nations, the rights of sovereign states to threaten or resort to the use of force are limited by consensual undertakings. Yet, voting procedures providing for unanimity or reserving veto rights and wide escape clauses (such as those contained in articles 51 and 107 of the UN Charter) tend to reduce such peacekeeping systems to relative ineffectiveness, in particular in relation to the major world powers.

In fields less central to the systems of open power politics or power politics in disguise, international law may also fulfill the functions of a law of reciprocity and of a law of coordination. Thus, for example, on the basis of many treaties, an international customary law of diplomatic immunity, codified in the 1961 Vienna Convention on Diplomatic Relations, has developed.

Occasionally, the law of reciprocity—that is, a set of legal rules, compliance with which rests normally on the expectation of mutual advantages rather than on the fear of the application of external sanctions—intrudes even into spheres that are closer to actual power politics. This is always possible on a consensual basis, as in peace treaties, and occasionally has happened, as in the various conventions agreed upon at Lausanne in 1923 that terminated a war between Greece and Turkey. The laws of war are illustrations from international customary law. They include legal rules that apply to wars fought for limited purposes where both sides have exhausted the means of escalation available in systems of power politics but resist the temptation of total war with no legal restraints.

In an unorganized international society based on entities that tend to put their own interest before the commonweal, the scope of a law of coordination, or community law, in which the common interest overrides sectional interests, is limited. An example of this type of law is the gradual outlawry of the slave trade by bilateral and multilateral consensual undertakings, especially the Treaty of Paris (1815) and the Slavery Conventions of 1926 and 1956.

The historical perspective. Since the dawn of history inchoate systems of international law have come into existence in many parts of the world. Though of comparative interest, most have not had any influence on the evolution of contemporary international law. Leaving aside the borrowing of some Roman-law terminology and legal techniques, such continuity as exists in the practice of international law dates from early medieval international law.

Historical foundations

International law is the product of a threefold process initiated in the Western world: the disintegration of the medieval European community into a European society, the expansion of this European society, and concentration of power in a developing world society in the hands of a rapidly declining number of major world powers.

The premises of medieval international law were simple. (1) In the absence of an agreed state of truce or peace, war was the basic state of international relations even between independent Christian communities. (2) Unless exceptions were made by means of individual safe conduct or treaty, rulers saw themselves entitled to treat foreigners at their absolute discretion. (3) The high seas were no-man's-land, where anyone might do as he pleased.

Treaty law was the predominant feature of medieval international law. Sanctions varied from the exchange of hostages, the pledging of towns, castles, and territories, and the mortgaging of the personal property of kings, or their subjects, to the appointment of guardians or the addition of the signatures of powerful dignitaries representing the various estates of a prince's realm. Supernatural sanctions, such as solemn oaths or excommunications of a guilty party, were also employed. The observance of treaties and other engagements rested on the same basis as it did in subsequent phases of international law: self-interest, especially in relation to obligations of a reciprocal character, and the value attached by an obligated party to his moral credit and his respect for the principle of good faith.

International law and the expanded world of nation-states

With the expansion of European society the spiritual basis of inter-Christian international law was weakened but not eliminated. In particular, the universalist spirit that imbued the naturalist doctrine of international law gave to international law the elasticity needed to adapt itself to a constantly widening international environment. Even so, international law primarily served the purposes of assisting in the process of Western expansion.

In the process of the transformation of international law into a world law, international law exchanged its Christian foundation for that of a law among states that were civilized in a highly formal sense. Civilization was understood as compliance with the minimum requirements of the rule of law, as this term, or its continental equivalents, was used in pre-1914 days in the Western world, especially regarding the treatment of the persons and property of foreign nationals. It took merely a further step to make sovereignty the decisive test of full international personality. In the pre-1939 era of the coexistence of democratic communities with totalitarian states such as the Soviet Union, Fascist Italy, Nazi Germany, and militaristic Japan, international law had become a law among sovereign states.

While the coexistence of sovereign states in a legal system postulates equality, this equality in international law is of a purely formal character. For example, since the formation of the United Nations in 1945, membership has increased threefold. Even though the number of sovereign and equal states has sharply increased, real influence still rests with only a handful of nations. The veto power of the permanent members of the UN Security Council, the weighting of votes according to the financial interest taken in the International Monetary Fund, and the special position accorded to states of chief industrial importance (as in the International Labour Organisation) are indicative of this trend.

The development of the doctrine of international law followed only slowly in the wake of the practice of international law. In the early days of international law it sufficed to have lawyers trained in canon and civil law. They tended to apply to novel situations the concepts of municipal law with which they were familiar. This accounts for the long-continued overemphasis in the doctrine of international law on analogies from these familiar systems of internal community law to systems developed by societies with very different cultural traditions.

The beginnings of European international law and relations are to be found in the microscopic interstate system of the Italian city-states. Here may be seen the beginnings of the doctrine of international law, especially in the writings of two Italian lawyers, Bartolo da Sassoferrato (1314–57) and Baldo degli Ubaldi (1327–1400). When in the late 15th and 16th centuries Spain became the leading Western power, Francisco de Vitoria (*c.* 1486?–1546) founded the Spanish school of international law. In the 17th century it came to be rivaled by the Anglo-Dutch school, particularly in the persons of Alberico Gentili (1552–1608) and Hugo Grotius (1583–1645).

The early theorists

While neither Grotius nor any other exponent of international law was the "father" of international law, Grotius' *De Jure Belli ac Pacis* (1625; *On the Law of War and Peace*) acquired a fame far greater than that of the works of his predecessors. This was due to a combination of factors that appealed to his contemporaries and subsequent generations: he stressed the self-defeating character of war, accepted sovereign states as the basic unit of international law, and skillfully blended natural law, Roman law, and state practice in a manner that left in vital matters sufficient discretion to governments to do, without legal hindrance, what they thought opportune.

Samuel von Pufendorf (1632–94), the German publicist and jurist, espoused the priority of natural law over positive law. An extreme naturalist school following his lead attempted to identify international law with natural law.

In England, Richard Zouche (1590–1661) laid the foundations of positivism in international law, drawing a sharp distinction between the postulates of natural law and international law as actually supplied in state practice. An eclectic school, sometimes described as Grotian, tried to find a golden mean between the extremes of naturalism and positivism by relying on both natural and positive law. Christian Wolff (1679–1754) and Emerich de Vattel (1714–67) were two of its early exponents.

Contemporary approaches

The one-sidedness and subjectivity of these techniques led to new departures on inductive, interdisciplinary, and relativist lines. The essence of the inductive approach is in the ascertainment of the rules of international law exclusively by means of generally accepted and rationally verifiable evidence. In particular, this involves recognition that the principles, as distinct from the rules, of international law are normally merely abstractions from these rules but do not constitute legitimate law-creating processes. The interdisciplinary treatment of international law makes it possible to view international law from an outside perspective provided by sociology, history, and ethics. Finally, the exploration of the possible forms of the development of international law in a relativist way, that is making available, side by side, various patterns that exist for the solution of any social problem, provides a detached but constructive approach to problems of international law in the making.

The ethical perspective. In its own speculative framework, the naturalist doctrine of international law provided both sociological and ethical perspectives of the subject. Subsequently, during the reign of positivism, the consideration of these perspectives was neglected.

A less subjective ethical measuring rod than those applied by naturalist writers is that of civilization itself. Links between international law and civilization exist not only historically but also explicitly in one of the three law-creating processes that the International Court of Justice is charged to apply: the general principles of law recognized by civilized nations. Civilization in this sense is more than a mature and rational apparatus of thought and action. It is a continuous process toward, and away from, community relations; its aim is to develop relations ultimately based on free cooperation and fellowship rather than on fear.

The relations between international law and civilization have varied considerably throughout the history of international law. Ample evidence of this is furnished by state practice regarding the recognition of new governments, states, or nations; the large-scale disregard of the minimum standard regarding the treatment of foreigners; and the ambivalent attitude of states to the legality of weapons of mass destruction. Thus, the relation between international law and civilization is tenuous at the best of times, and it is advisable for any ethical evaluation of contemporary international law to err on the side of caution.

SOURCES OF INTERNATIONAL LAW

Basic issues

Certain basic issues call for discussion: the law-creating processes of international law; the law-determining agencies of international law; the relations between international law and municipal law; and the nature of and the relations between rules, principles, and standards of international law.

The law-creating processes of international law. These are the forms in which rules of international law come into existence; *i.e.,* treaties, rules of international customary law, and general principles of law recognized by civilized nations. It is the merit of article 38 of the Statute of the International Court of Justice that this exclusive list of primary law-creating processes has received almost universal consent. States that have assented to, or acquiesced in, resolutions adopted unanimously by the UN General Assembly and stated to be declaratory of existing international law may be thought to be prevented in good faith from contesting any longer the existence of a formerly controversial rule of international law. On a consensual basis and, thus, in accordance with one of the primary law-creating processes, this and other secondary law-creating processes can, and have, come into existence.

International customary law. This is essentially the international law of unorganized international society, and its rules can be summarized under the heads of seven fundamental principles. The two constitutive elements of international customary law are (1) a general practice of states on a universal, general, or regional basis and (2) the acceptance by the states concerned of this practice as law.

The origin of international customary law is frequently found in earlier treaty clauses, which subsequently were taken for granted, as with the rules regarding the minimum standard applicable to foreign nationals and their property. Occasionally, as in the law of the sea and the law of armed conflict, individual rules of international law have developed out of roughly parallel practices of the leading powers.

Treaties. Treaties and other consensual engagements are legally binding undertakings by which, without any requirements of form under international customary law, the subjects of international law may declare, modify, or develop existing international law as they see fit or agree on transactions; *e.g.,* of a territorial character. They are thus able to transform *jus strictum* into *jus aequum, jus dispositivum* into *jus cogens* (see below), and vice versa, and unorganized international society into global or regional international societies on confederate or supranational levels of integration.

The general principles of law recognized by civilized nations. Such principles must fulfill two requirements. To qualify under this heading, a legal principle must be a general principle of law, as distinct from a legal rule of a more limited functional scope. It must be recognized and shared by a fair number of civilized nations and probably include representation of at least the principal legal systems.

The general principles of law come into play only as a subsidiary law-creating agency, that is, in the absence of competing rules of international customary law or treaty law. Their existence in the background forestalls any argument that supposed gaps in international law prevent international judicial organs from deciding on the substance of any dispute submitted to their jurisdiction.

The law-determining agencies of international law. These agencies furnish the evidence for the existence of asserted rules of international law. The totality of the subjects of international law constitutes the relevant agency for any rules of universal customary international law. A convincing majority of subjects of international law provides the requisite evidence for the existence of an alleged rule of general customary international law. The ensemble of the parties to a treaty fulfills the same function in relation to any particular consensual engagement. The body of civilized nations forms the relevant law-determining agency regarding general principles of law recognized by civilized nations.

The decisions of international and national courts and tribunals plus the doctrine of international law (*i.e.,* the teachings of the most highly qualified publicists) constitute what are described in article 38 of the Statute of the International Court of Justice as "subsidiary" means for the determination of the rules of international law.

In practice, consensus scarcely ever exists in any of these law-determining agencies. Thus, it is necessary to determine the relative evidential value of any pronouncements made by the elements of law-determining agencies; *i.e.,* the views of individual parties to treaties, relevant diplomatic material, and pertinent decisions of international and national judicial organs. To obtain as objective as possible an evaluation, it is advisable to subject each case to a threefold scrutiny: the degree of generic and individual independence of the element of the law-determining agency concerned, its international outlook, and its professional attainments.

International law and municipal law. International law applies in the relations between the subjects of international law. The relations between subjects and objects, and between objects alone, of international law are governed by municipal law or quasi-international law.

While international law is a legal system that actually exists, the term municipal law is an abstraction from the

The relation between municipal and international law

multitude of legal systems that are internal to the individual subjects of international law. Thus, actual conflicts can only arise between international law and individual legal systems other than international law, such as United States law or Soviet law. How such conflicts are resolved depends on the level on which they arise. Ultimately, any municipal organ is governed by its own municipal law and must, if needs be, give priority to it. Similarly, international organs, such as the International Court of Justice, may have to give priority to international law and treat municipal law as inferior in an accepted hierarchy of interlocking legal systems. They may even view international law as being exclusive of all other law and treat municipal law as a mere set of facts, which, as the case may be, complies with or contravenes the international obligations of a subject of international law.

Other basic issues of international law. *Rules, principles, and standards.* The rules of international law are the legal norms that can be verified as the products of one or more of the three generally recognized law-creating processes. For purposes of systematic exposition and legal education, it is also valuable to abstract principles from legal rules. Such principles of international law provide the common denominator for a number of related legal rules. They must not be abused by reversing the procedure for the purpose of deriving from them additional legal rules that cannot be verified independently by reference to the primary or secondary law-creating processes of international law. The more fundamental the rules that underlie any particular principle, the more a justification exists for seeing the principle itself as fundamental. It is possible to summarize the whole of international customary law in a number of fundamental principles, and attempts even have been made to reduce all these rules to a single fundamental principle, or *Grundnorm,* such as consent, recognition, and good faith.

Underlying norms

By way of treaty, subjects of international law are free to create additional principles; *e.g.,* those of freedom of commerce or navigation, or a principle such as that of peaceful coexistence embodied in article 2 of the Charter of the United Nations. Unless parties desire to give unconditional effect to any such optional principle, they have at their disposal counterparts to compulsory rules in the form of optional standards, such as those postulated by most favoured nation and preferential treatment.

Jus dispositivum and jus cogens. In terminology derived from Roman law, a distinction is made in mature legal systems of municipal law between rules that may be altered by contracting parties (*jus dispositivum*) and others that may not (*jus cogens*). As distinct from legal systems with a centralized legal order, around which such *jus cogens* has grown, international customary law, as the law of unorganized international society, does not know of any such peremptory rules. Limitations on the freedom of states imposed by common sense, self-interest, and other pragmatic considerations must not be mistaken for *jus cogens.* Yet, nothing prevents sovereign states from creating peremptory international law by way of treaty; *e.g.,* the seven principles formulated in article 2 of the United Nations Charter.

Jus strictum and jus aequum. Also derived from Roman law, this distinction indicates differences between two other types of rule. Rules of *jus strictum* (*e.g.,* the rules of international customary law on the right of a state to request the recall of a foreign envoy as persona non grata) must be interpreted strictly and literally as embodying absolute rights. Others, such as those providing for freedom of communication, must be interpreted as rules of *jus aequum; i.e.,* in a reasonable and equitable manner.

Rules of international law

CUSTOMARY INTERNATIONAL LAW

The basic rules of international customary law can be summarized in the following fundamental principles: sovereignty, recognition, consent, good faith, freedom of the seas, international responsibility, and self-defense. In this survey are also included post-1945 codifications of the relevant rules of international customary law.

Sovereignty. Initially, a subject of international law is bound only by applicable rules of universal or general international customary law. Additional international obligations may be imposed on any subject of international law only with its consent. Unless the territorial jurisdiction of a state is excluded or limited by rules of international law, its exercise is exclusively the concern of the state in question. Subjects of international law may claim potential jurisdiction over persons or things outside their territorial jurisdiction. In the absence of permissive rules to the contrary (*e.g.,* the right of hot pursuit from the territorial sea to the high sea, or the right of reprisal) they may exercise such jurisdiction only inside their territories. It follows from the coexistence of sovereign states under international law that, in principle, they are all equal in status.

Territorial jurisdiction

Recognition. The rules governing recognition cover situations such as the co-option of new subjects of international law, the recognition of territorial claims of another state, the grant and withdrawal of nationality, and the recognition of the maritime flag of a landlocked state.

In principle, recognition is discretionary, but premature recognition of belligerents and insurgents runs counter to the exclusive domestic jurisdiction of the other state concerned and is illegal. The scope and effects of recognition must be ascertained according to the tenor of the act of recognition and its context. It may be unconditional or conditional and may be explicit or implied.

The devices of protest and reservation of rights may be used to prevent silence from being misinterpreted as an implied recognition of a situation or transaction. Notification is a means of bringing a situation or transaction to the attention of a third power with the intent to invite recognition or some other reaction.

The uses of recognition

In practice, the chief function of recognition is to acknowledge the existence of an entity as a subject of international law with whom another state can maintain diplomatic relations. The main forms of recognition are recognition of a state or government as exercising de facto or de jure authority in a territory or, as it is simply called, de facto and de jure recognition. De facto recognition implies acceptance of the claim of the recognized government to exercise jurisdiction within its own territory. De jure recognition, however, usually implies acceptance of the claim of the recognized government to exercise extraterritorial jurisdiction over, for example, nationalized companies that own ships entitled to sail under the flag of the recognized state.

Recognition, being a matter of intent, may fall short of full recognition and be limited to recognition of a group as belligerents or as insurgents, if such rebels are in de facto control of part of the territory of another state.

Sovereign states are the principal subjects of international law. Yet nothing prevents states from recognizing dependent states with limited international personality, such as international protectorates or the former mandates of the League of Nations. None of the trust territories of the United Nations has international personality. They are, however, under the control of the United Nations. Similarly, states are free to recognize, for all or limited purposes, nontypical subjects, such as the Holy See, international institutions, and even individual persons as subjects of international law. In each case, whether any entity has been so recognized is merely a question of evidence.

Consent. The rules on consent enable subjects of international law, when entering into agreement, to modify and to supplement as they see fit, but without prejudice to the rights of third parties, any of the rules of international customary law or the general principles of law recognized by civilized nations.

Sovereign states have full capacity to enter into any kind of consensual engagement. The capacity of other entities with international personality to undertake consensual commitments under international law is limited according to the scope of their international personality. In the absence of evidence to the contrary, consensual engagements between subjects of international law are governed by international law, but consensual engagements between subjects and objects, or between objects of international law, are outside the pale of international law.

The ambit of agreement

Barring prior obligations to the contrary, as contained, for example, in an undertaking to negotiate or conclude another agreement, the entry into consensual engagements is purely optional.

International customary law does not prescribe any particular form for consensual engagements, unless the parties desire not to create legal obligations. The effect of consent given in accordance with the requirements of international law is to create legal rights and duties between the contracting parties. In the absence of any contrary intention of the parties, the suspension, revision, and termination of consensual engagements depend on the consent or acquiescence of each of the contracting parties.

Excepting agreements entered into by international organizations, the subject of the law of treaties is now codified in the 1969 Vienna Convention on the Law of Treaties.

Good faith. In the early phases of the evolution of international law, good faith meant, primarily, the absence of bad faith. Gradually, however, good faith was identified with the requirements of reasonableness, common sense, and equity.

Thus, parties to consensual engagements and parties responsible for duly communicated unilateral acts, which they intend to have legal effect, must interpret and execute such engagements in good faith. If a consensual engagement that is subject to international ratification has been ratified, good faith regulates also the relations between the parties prior to final ratification. Barring more specialized provisions, acts committed contrary to good faith by any international institution, all of which derive their authority from consensual engagements, are void. Excess of jurisdiction by an international judicial organ or corruption of judges by one of the parties falls in this category. Rules of *jus aequum* must be interpreted as relative rights; that is, their arbitrary exercise is an abuse of right and a tortious act. In the case of rights derived from rules of *jus strictum,* a harsh exercise of such rights is not illegal but amounts to an unfriendly act; that is, it is open to retorsion, meaning lawful but unfriendly acts of retaliation. On the international judicial level the consensual nexus within which judges and parties operate tends to transform any absolute rights into relative rights, subject to judicial balancing processes in which considerations of good faith, common sense, and reasonableness play a prominent part.

Equitable principles

Freedom of the seas. The inclusion of the rules on the freedom of the high seas (*i.e.,* those parts of the interlinking chain of oceans that lie seaward of the territorial sea) among those of a fundamental character would be justifiable on the ground alone that they apply geographically to two-thirds of the globe. These rules preclude the appropriation by any individual subject of international law of any portion of the high seas as distinct from the subsoil and bed of the sea. The exercise of permitted jurisdiction varies according to the state of peace, intermediacy between peace and war (*status mixtus*), or war between the states concerned. Subject to a number of exceptions, in time of peace a state may exercise jurisdiction only over ships entitled to fly its own flag. In a state of intermediacy, states are free, under international customary law, to interfere with one another's shipping by way of reprisal. In time of war, permissible interference with enemy and neutral shipping is regulated by the rules of sea warfare and prize law. The use of the high seas, the airspace above the high seas, and the seabed must be exercised with reasonable regard for the interests of others. The Conventions of 1954, 1962, and 1969 for the Prevention of Pollution of the Sea by Oil provide a limited implementation of this rule. Piracy *jure gentium* (*i.e.,* illegal acts of violence, detention, or depredation for private ends committed on the high seas) and slave trading are illegal forms of the use of the high seas under international customary law.

The subject is now largely codified in the 1958 Geneva Conventions on the High Seas and on Fishing and Conservation of the Living Resources of the High Seas.

International responsibility. The rules governing the principle of international responsibility complement all other rules of international law. They transform merely admonitory precepts into legal forms and, in this sense, may also be described as sanctions of international law.

The rules on international responsibility can be reduced to two propositions: (1) the breach of any international obligation by the organ of a subject of international law constitutes an illegal act or international tort, and (2) the commission of an international tort involves the duty to make reparation. These are rules of international customary law. Thus, the obligations they create arise independently of the will of any particular subject of international law, and they may be modified by consent and acquiescence. In particular, they can be strengthened by consensual rules that provide for penalties corresponding to those in municipal criminal law (sometimes also described as international criminal law) but, actually, constituting merely internationally postulated rules of municipal law, which may be waived by acquiescence and nonprosecution of claims (also described as extinctive prescription).

Other rules, such as the powers exercised by states in relation to pirates, blockade runners, and war criminals, constitute extraordinary forms of the exercise of national jurisdiction. They are lawful because the home states of these three groups of individuals may not in good faith contest the exercise of such jurisdiction.

Self-defense. In unorganized international society the distinction between the lawful and unlawful use of force was accepted in state practice in situations of *status mixtus.* In a state of war any limitations of the right to wage war (*jus ad bellum*) remained a largely ignored postulate of naturalist doctrine on the distinction between just (and legal) and unjust (and illegal) war. The realization of this objective had to await later multilateral treaties, which, by reference to the test of self-defense, incorporated the distinction between legal and illegal wars and other use of force.

The use of force

Under international customary law measures of self-defense may be taken against illegal acts that are attributable to another subject of international law; against acts of individuals, ships, or aircraft that disentitle any other subject of international law from the grant of protection; and against acts of objects of international law that lack a subject of international law that is entitled to give them diplomatic protection.

The need for self-defense must be compelling and instant. Measures of self-defense comprise any action, including hot pursuit from the territorial sea into the high seas, that is necessary to repel an imminent or present invasion of the rights of a subject of international law.

In cases not covered by the conditions of lawful self-defense, the threat or use of force under international customary law may amount to a legitimate form of self-help. If a subject of international law has committed an international tort and refuses to make reparation, the other party may resort to acts of retorsion or reprisal.

The legal effects of resorting to war under international customary law are to bring into operation the laws of war and neutrality (*jus in bello*).

INTERNATIONAL RULES GOVERNED BY TREATIES AND OTHER AGREEMENTS

The sphere of freedom of action for subjects of international law—what, in relation to typical international persons, may also be termed unlimited state jurisdiction—is governed primarily by the rules on sovereignty. Limitations of this jurisdiction come about as the result of the interplay of the rules underlying some of the other fundamental principles with those on sovereignty.

This interaction of rules has brought about secondary rules and legally determined situations. Five of these are of especial significance: territory, diplomatic law, and immunity; the protection of nationals abroad; freedom of commerce and navigation; extradition and asylum; and succession to international rights and obligations.

Territorial control

Territory. Owing to the preponderance, in a world largely appropriated by sovereign states, of territorial over personal jurisdiction, the rules governing title to territory are of major importance. There are significant exceptions, however, such as the high seas; Antarctica, barred from further exclusive appropriation by the 1959 Antarctic Treaty; and outer space and celestial bodies, excluded under the 1967 Outer Space Treaty. The rules relating to territory

rest, first, on sovereignty: occupation, addition by natural causes of new land to riverbanks (accretion, accession, or alluvion), and assumption, under international customary law, of sovereignty over territories whose state apparatuses have been destroyed by conquest (debellatio); second, on recognition that stops third parties from contesting the validity of a recognized title; and, third, on consent—namely, consent of the cession of territory.

The legal function of frontiers is to settle the exact extent of contiguous territories by unilateral action, express consent, recognition, or acquiescence.

The airspace above, and the subsoil below, national territory, including the territorial sea, are treated as appurtenances of a state's territory.

Internal waters include ports, harbours, all waters on the landward side of the baseline of the territorial sea, and historic bays; *i.e.,* bays that, irrespective of their width, are treated, on grounds of acquiescence or recognition, as subject to the jurisdiction of the coastal state.

The normal baseline of the territorial sea is the low-water line along a state's seacoast. It is generally recognized that the minimum breadth of the territorial sea is three miles. The outer limit of the territorial sea, which constitutes also the frontier between national territory and the high sea, is drawn by reference to the baseline.

Most of the sea matters are not codified in the 1958 Convention on the Territorial Sea and Contiguous Zone. While it has proved impossible to reach agreement on the breadth of the territorial sea, it is laid down in the above convention that the contiguous zone—*i.e.,* a geographically limited zone of the high seas contiguous to the territorial sea in which coastal states exercise a limited jurisdiction over foreign ships—should not extend beyond 12 miles from the baseline of the territorial sea. (G.Sc.)

Outer space. The successful launching in 1957 of the first artificial Earth satellite marked the beginning of a new branch of international law, namely, international space law. This refers to those rules within the international legal system that regulate human activities in outer space, including the Moon and other celestial bodies, and in relation to outer space.

International space law embraces, in the first place, all existing and future rules of international customary or treaty law that contain no geographical limitations, expressed or implied, and are consequently applicable to any conduct of subjects of international law. Insofar as international customary law is concerned, these rules include virtually all those governing the principles of recognition, consent, good faith, self-defense, and international responsibility. That, in principle, international law was from the very beginning applicable to outer space has been reaffirmed by the United Nations on several occasions.

International space law includes, second, such new rules, whether of customary or treaty law, that have evolved since the beginning of the space age or that may be developed specifically to regulate the activities of states and their nationals in space or actions in relation to such activities. While a number of bilateral and multilateral agreements as well as international agencies already exist in this field, the United Nations has been particularly active in the development of general international space law. It has adopted a number of resolutions embodying recommended standards of conduct in relation to outer space and has prepared five multilateral treaties for adoption by states.

United Nations resolutions

These multilateral treaties are: (1) the 1967 Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies; (2) the 1968 Agreement on the Rescue of Astronauts, the Return of Astronauts and the Return of Objects Launched into Outer Space; (3) the 1972 Convention on International Liability for Damage Caused by Space Objects; (4) the 1975 Convention on Registration of Objects Launched into Outer Space; and (5) the 1979 Agreement Governing the Activities of States on the Moon and Other Celestial Bodies. The last one proclaims the Moon and other celestial bodies within the solar system, other than the Earth, together with their natural resources, the common heritage of mankind. All five treaties are in force among their respective contracting parties, but the most important of these are doubtless the treaties on principles and on liability. The latter lays down detailed rules governing the recovery of damages for losses caused by space objects.

Many of the provisions of the treaty on principles are not only binding on the contracting parties in terms of treaty law but also have gained wide acceptance from both parties and nonparties as expressing rules of general international law; consequently they are binding on all states. In fact, the technological advances in this sphere have been so fast and the need to adopt appropriate legal rules in response thereto has been so pressing that it has become increasingly apparent that much of the product of what has traditionally been called the law-creating process of international customary law need not rest on custom at all. Such practice as may be required as one of the two constitutive elements of a rule of international customary law needs to have existed only long enough to prove the existence of a general acceptance among states of the norm in question as a rule of international law. In fact, the law-creating process can be reduced to a single act or omission, and its role is reduced to that of merely establishing the acceptance of a given rule as law by the generality of states.

Among the norms of treaty law that appear to have acquired such general acceptance are primarily some of the basic principles found in the 1967 treaty. They include the principles that outer space, including the Moon and other celestial bodies, is not subject to national appropriation (Article II) and that the state of registry of a space object retains jurisdiction and control over such object and over any personnel thereof while in outer space or on a celestial body (Article VIII). If so, this means that under general international law there can be no territorial acquisition or exercise of territorial sovereignty in outer space and on celestial bodies. With regard to this issue, at least some of the rules underlying the principle of freedom of the high seas are capable of being applied by analogy. Two other principles from the 1967 treaty have probably also achieved the same status as rules of general international law: first, states bear international responsibility for national activities in space and must subject such activities, whether conducted by official bodies or private individuals, to authorizations and control (Article VI); and second, states that launch or procure the launching of a space object or from whose territory or facility a space object is launched are internationally liable for damage caused to another state or its nationals by such object (Article VII).

Neutrality of outer space

More problematic is Article IV of the 1967 treaty, a key provision from the standpoint of world disarmament. Article IV, in its first paragraph, prohibits the stationing of nuclear weapons or other weapons of mass destruction in outer space or on celestial bodies, while in its second paragraph it prescribes that the Moon and other celestial bodies (without mention of outer space as such) shall be "used . . . exclusively for peaceful purposes." The problem is twofold. First, it is difficult to determine, on the evidence available, whether this provision has been generally accepted as declaratory of general international law. Second, there is a wide division of opinion as to the interpretation of the word peaceful in the second paragraph. One school maintains that it means "nonmilitary," while the other argues that it means simply "nonaggressive." The latter interpretation appears to have gained currency, although it renders the stipulation supererogatory, inasmuch as aggressive activities as such are contrary to international law wherever they may occur.

The arrival of the space age has created many problems crying out for legal regulation. International agencies other than the United Nations have also contributed to the development of international space law. The International Telecommunication Union, for example, has done much to regulate, among other things, the use of radio frequencies for telecommunications and direct television broadcasting by artificial satellites. Many problems remain unresolved and others will no doubt arise. Among those being discussed are the definition of outer space and its delimitation from airspace, equitable use of the

geostationary orbit, the use of nuclear-powered satellites, international direct television broadcasting, remote sensing, and the military use of outer space. (Bi.C.)

Diplomatic law and immunity. States and international institutions can act only through individuals. Thus, relations between states—and international institutions—are based on the principle of necessary representation. The chief representative of a state is the head of state who, in principle, has plenary powers to commit his state. After a number of earlier attempts to settle continuous disputes over the precedence of diplomatic envoys, the classes of diplomatic envoys and their privileges and immunities are now codified in the 1961 Vienna Convention on Diplomatic Relations.

Protection and status of citizens and economic interests abroad

Similarly, the rights and immunities of consuls—resident officials stationed abroad with the consent of the receiving state for purposes of promoting trade and assisting nationals of the sending country—are codified in the 1963 Vienna Convention on Consular Relations.

The protection of nationals abroad. The relevant rules for such protection grew out of individual safe-conducts and innumerable bilateral treaties of commerce and navigation and, between civilized nations, were gradually taken for granted as rules of international customary law or general principles of law recognized by civilized nations.

These rules imply the application of a minimum standard that complies with the rule of law, as understood in liberal and democratic Western countries, regarding the protection of the life, liberty, dignity, and property of foreign nationals. Regarding property, the freedom of states to expropriate or nationalize private property in the public interest with full (or adequate), prompt, and effective compensation is generally accepted as a rule of international customary law. The rule has behind it the authority of the Permanent Court of International Justice and a considerable number of international tribunals. Doubts that have been raised against the continued validity of the rule (especially in Communist and capital-importing states) are related to the application of the rule in cases of doubtful titles to property rather than to the existence of the rule itself. There is also a widespread mixture of politics, trade, and aid that, on pragmatic grounds, frequently makes inadvisable an insistence by capital-exporting states on strict compliance with the rule.

Freedom of commerce and navigation. Under international customary law the right of foreign nations to trade in a country and use its means of communications, such as roads, rivers, and airspace, is within the exclusive jurisdiction of the territorial sovereign. By way of treaty such rights of commerce and navigation are granted normally on relative terms; *i.e.,* by reference to optional standards. The classical standards of international treaty law in these fields are those of most favoured nation treatment (treatment on the basis of foreign parity), national treatment (treatment on the basis of inland parity), identical treatment, equitable treatment, good-neighbourly treatment, open-door treatment (equal treatment of all concerned in a third sovereign state or a territory such as a United Nations trust territory), and preferential treatment. In state practice, some of these standards are employed cumulatively or alternatively in one and the same treaty.

Extradition and asylum. In accordance with a long-established practice, states have concluded extradition treaties enabling them to secure the return of fugitives from their own territorial jurisdiction. In states in which the rule of law in the Western sense applies, considerable care is taken to define precisely the offenses for which extradition may be granted, and extradition normally is limited to nonnationals of the country requested to grant extradition. While a number of states take a different view of political crimes, it is a liberal Western tradition to exclude political offenders from extradition unless they are charged with an attack on life.

Treatment of fugitives

In the absence of consensual undertaking to the contrary, any state may grant asylum in its own territory to any individual. This territorial asylum differs from diplomatic asylum; *i.e.,* asylum that is granted in diplomatic premises situated in another state's territory. In the absence of express treaty rights to this effect, diplomatic asylum may not be granted, but, on humanitarian grounds, the territorial sovereign often acquiesces in such action.

Succession to international rights and obligations. It is necessary to distinguish three typical situations: (1) revolution—this, in principle, is treated as a purely internal affair and does not affect the obligations of the subject of international law concerned; (2) territorial changes—if two states decide on the cession of an insignificant portion of territory, the matter is settled between the parties by the rules on consent and, in relation to third parties, by those on recognition; if a state agrees to its own truncation or if a composite state is dismembered, the legal consequences of such changes are settled by treaty, recognition, or acquiescence; (3) belligerent occupation—in the case of belligerent occupation falling short of debellatio, any territorial changes are treated as temporary while the war lasts. Furthermore, it is presumed that in the absence of any express settlement in a treaty of cession the public property of the ceding state becomes automatically the property of the cessionary state, and the public law of the ceding state is replaced by that of the cessionary state.

There is no general rule of international customary law imposing automatic succession by the cessionary state to the state debts of the ceding state. On equitable grounds, however, a rule to the opposite effect is frequently asserted regarding strictly localized debt. Cessionary states are under no obligation to assume any responsibility for tortious acts or omissions of the ceding state.

In principle, treaties are binding only between the contracting parties. Thus, if one of the parties cedes part of its territory, existing treaties are interpreted according to the rule of movable treaty frontiers; that is, the territorial scope of treaty obligations is presumed to be automatically adjusted to subsequent territorial changes. In cases in which the nonexistence of rules of international law on the automatic succession to international obligations would lead to harsh results, these are likely to be mitigated by the need of the new subject of international law concerned to be recognized and the freedom of existing subjects to make recognition dependent on compliance with justified expectations.

The role of international organizations

Global multipurpose institutions, such as the League of Nations and the United Nations, are best understood as organizational superstructures of international customary law on a consensual and confederate basis. Their impact on international law is threefold: modification by express consent of the rules underlying the fundamental principles of international law, indirect modification of these rules by acquiescence on the part of member states in the action of organs not actually authorized to exercise lawmaking functions, and initiation of the further codification and development of international law.

United Nations influence on international law

The chief modification introduced by the United Nations Charter is the limitation of the rights of subjects of international law under international customary law to threaten or resort to armed reprisals and war. This extends the duties of the former members of the League of Nations and parties to the Kellogg–Briand Pact of 1928. The prohibition covers the threat or use of force in circumstances falling short of war in the formal sense.

The principal means of indirect lawmaking in the United Nations are resolutions of the General Assembly that are adopted unanimously or with the two-thirds majority required for important questions. If such resolutions purport to be declaratory of international law, it is difficult for member states who voted for them to claim that, on the matters involved, the General Assembly is limited to the mere task of making recommendations. If the organs concerned of the United Nations act consistently on particular resolutions, eventually a time comes when even those states that have voted against them will be deemed to have acquiesced in such resolutions. Nonmember states that are admitted to membership in the United Nations after such resolutions have been adopted may find themselves in a similar situation. They have obtained their recognition on the assumption that they will abide by the

generally accepted rules of international law, and, increasingly, member states that grant recognition may equate the near-universal law and practice of the United Nations with general international customary law. Moreover, new members must expect that they join this global confederation as they find it.

A number of resolutions passed by the General Assembly fall into this in-between category of law-in-the-making—*e.g.,* those on the Nuremberg Principles that dealt with crimes against peace, war crimes, and crimes against humanity (Res. No. 95[II], 1946); genocide (Res. No. 96[I], 1946); the Universal Declaration of Human Rights (Res. No. 217[III], 1948); the right of peoples and nations to self-determination (Res. No. 637[VII], 1952); permanent sovereignty over natural resources (Res. No. 1803[XVII], 1962); denuclearization (Res. No. 1884[XVIII], 1963); and nonintervention (Res. No. 2131[XX], 1965).

In some instances, as before the adoption of the Universal Declaration of Human Rights, the almost unanimous protestations by speakers in the General Assembly regarding the purely moral character of the precepts enshrined in the declaration provide adequate evidence of the nonlegal character of the resolution in question. In others, such intention may become evident from the self-contradictions contained in the resolutions themselves. In still others, the intimation of the need for further study and the request for codification of the subject may suggest the political rather than legal character of a particular resolution. But if at any subsequent stage it can be shown that large and consistent majorities of the principal organs of the United Nations accept rules laid down in such resolutions as legally binding, the transition from law-in-the-making to new law tends to be made.

International Law Commission

The International Law Commission, an auxiliary but autonomous organ of the General Assembly of the United Nations, consists of 25 members of recognized competence in international law (Article 2[1] of the Commission's Statute). It has initiated codification and development in a number of fields of international law. In practice, the commission does not distinguish between its efforts on the codification (*i.e.,* the restatement of existing international law) and the development of international law by draft rules involving changes in existing international customary law. Thus, any of the rules proposed by the commission must be examined from this point of view.

In the field of humanitarian law (*i.e.,* the protection of the individual) the International Convention on the Elimination of All Forms of Racial Discrimination of 1965 and the International Covenants on Civil and Political and on Economic, Social, and Cultural Rights, opened for signature in 1966, were channeled from the United Nations Commission on Human Rights to the General Assembly of the United Nations through the Economic and Social Council. On a level of closer constitutional and ideological homogeneity, the Council of Europe adopted the Rome Convention for the Protection of Human Rights and Fundamental Freedoms of 1950, as subsequently amended, and in the European Commission of Human Rights and the European Court of Human Rights provided the most effective means yet put into operation for the implementation of the protection of human rights.

On specialized topics, such as the law of the sea, international labour law, and international private law, the Inter-Governmental Maritime Consultative Organization, the International Labour Organisation, and the Hague Conference of Private Law, respectively, fulfill drafting functions of a quasi-legislative character comparable to those of the International Law Commission. Yet it remains for the sovereign states concerned to decide if they want to limit their freedom of action by such further consensual commitments.

Future development

If this will exists, states are not limited to the development of international law on a confederate level. They are free to transform regional, continental, or global areas into federations of a territorial type such as the Commonwealth of Australia. They may also try functional federation on the model of such European supranational organizations as the European Economic Community. Under these conditions the wheel has come full circle, and international law turns again into municipal law, but, until such a development becomes universal, international law is likely to remain indispensable in the relations between such sectional groupings. (G.Sc.)

BIBLIOGRAPHY

Collections of treaties: CLIVE PARRY (ed.), *The Consolidated Treaty Series,* 231 vol. (1969–81); LEAGUE OF NATIONS, *Treaty Series: Publications of Treaties and International Engagements Registered with the Secretariat of the League,* 205 vol. in 211 (1920–46), superseded by UNITED NATIONS, *Treaty Series: Treaties and International Agreements Registered or Filed and Recorded with the Secretariat of the United Nations* (1946–); UNITED NATIONS, *List of Treaty Collections* (1956, reprinted 1981); and MANLEY O. HUDSON (ed.), *International Legislation,* 9 vol. (1931–50, reprinted 1970–72).

Reports of judicial case law: UNITED NATIONS. OFFICE OF LEGAL AFFAIRS. CODIFICATION DIVISION, *Reports of International Arbitral Awards* (irregular); *International Law Reports* (annual); and A.M. STUYT, *Survey of International Arbitrations, 1794–1970* (1972).

Digests of state practice: ALEXANDRE CHARLES KISS (ed.), *Répertoire de la pratique Française en matière de droit international public,* 6 vol. (1962–69); *British Practice in International Law* (irregular); JOHN BASSET MOORE (ed.), *A Digest of International Law,* 8 vol. (1906, reprinted 1970); GREEN HAYWOOD HACKWORTH (ed.), *Digest of International Law,* 8 vol. (1940–44, reprinted 1973); MARJORIE M. WHITEMAN (ed.), *Digest of International Law,* 15 vol. (1963–73).

Dictionaries: KARL STRUPP (ed.), *Wörterbuch des Völkerrechts,* 2nd ed. edited by HANS-JÜRGEN SCHLOCHAUER *et al.,* 3 vol. (1960–62); and UNION ACADÉMIQUE INTERNATIONALE, *Dictionnaire de la terminologie du droit international* (1960).

General treatises: MICHAEL AKEHURST, *A Modern Introduction to International Law,* 5th ed. (1984); GERHARD VON GLAHN, *Law Among Nations: An Introduction to Public International Law,* 4th ed. (1981); R.C. HINGORANI, *Modern International Law,* 2nd ed. (1984); RICHARD FALK, FRIEDRICH KRATOCHWIL, and SAUL H. MENDLOVITZ (eds.), *International Law: A Contemporary Perspective* (1985); J.L. BRIERLY, *The Law of Nations: An Introduction to the International Law of Peace,* 6th ed. edited by HUMPHREY WALDOCK (1963); CHARLES CHENEY HYDE, *International Law, Chiefly as Interpreted and Applied by the United States,* 2nd rev. ed., 3 vol. (1945); MYRES S. MCDOUGAL, *Studies in World Public Order* (1960); D.P. O'CONNELL, *International Law,* 2nd ed., 2 vol. (1970); L.F.L. OPPENHEIM, *International Law: A Treatise,* 8th ed. edited by H. LAUTERPACHT, 2 vol. (1955, reprinted 1974); GEORG SCHWARZENBERGER and E.D. BROWN, *A Manual of International Law,* 6th ed. (1976); MAX SØRENSON (ed.), *Manual of Public International Law* (1968); PAUL REUTER, *Droit international public,* 5th rev. ed. (1976); CHARLES ROUSSEAU, *Droit international public,* 10th ed. (1984); GEORG DAHM, *Völkerrecht,* 3 vol. (1958–61); ALFRED VERDROSS, *Völkerrecht,* 5th ed. edited by STEPHAN VEROSTA and KARL ZEMANEK (1964); D. ANZILOTTI, *Corso di diritto internazionale,* 4th ed., vol. 1 (1964); ANGELO P. SERENI, *Diritto internazionale,* 4 vol. in 5 (1956–65); CÉSAR SEPÚLVEDA, *Derecho internacional,* 13th ed. (1983); ACADEMY OF SCIENCES OF THE U.S.S.R., *International Law,* trans. from Russian by DENNIS OGDEN (1967); G.I. TUNKIN, *Вопросы международного права* (1960), available also in French and German trans.; EDWARD MCWHINNEY, *Conflict and Compromise: International Law and World Order in a Revolutionary Age* (1981); JOHN A. PERKINS, *The Prudent Peace: Law as Foreign Policy* (1981); BIN CHENG, "Outer Space: The International Legal Framework," in INSTITUTE OF INTERNATIONAL PUBLIC LAW AND INTERNATIONAL RELATIONS OF THESSALONIKI, *Air and Outer Space Law,* pp. 41–106, vol. 10 of *Thesaurus Acroasium* (1981); NANDASIRI JASENTULIYANA and ROY S.K. LEE (eds.), *Manual on Space Law,* 4 vol. (1979–81); CARL Q. CHRISTOL, *The Modern International Law of Outer Space* (1982); and A.S. PIRADOV (ed.), *International Space Law* (1976; originally published in Russian, 1974).

Periodicals and yearbooks: *The American Journal of International Law* (quarterly); *Archiv des Völkerrechts* (quarterly); *The British Yearbook of International Law* (annual); *Indian Yearbook of International Affairs* (annual); *The Japanese Annual of International Law* (annual); *Revue générale de droit international public* (quarterly); *Советский ежегодник международного права* (annual), with English summaries; *The Year Book of World Affairs* (annual); *Zeitschrift für ausländisches öffentliches Recht und Völkerrecht* (quarterly); *Journal of Space Law* (semiannual); *Annals of Air and Space Law* (annual), in English and French; and INTERNATIONAL INSTITUTE OF SPACE LAW. COLLOQUIUM, *Proceedings* (annual).

Bibliography: JACOB ROBINSON, *International Law and Organization: General Sources of Information* (1967).

(G.Sc./Bi.C.)

20th-Century International Relations

The purpose of diplomatic history is to explain the origins and the effects of foreign policies, and its focus is each country's policy-making elite. But this involves the historian at once in a complication. The statesman exists simultaneously in two realms: the domestic political system whence his authority derives, and the international system in which he represents his state to the world. Pressures and problems, temptations and opportunities arise constantly from both realms. Which one ought to command the historian's attention? The founders of modern diplomatic history, beginning with Leopold von Ranke, propounded a view known as "the primacy of foreign policy." Founded on German Idealist philosophy, Rankeanism asserted the primary influence of a state's geography and external threats in the shaping, not only of its foreign policy but of its internal military, political, and cultural institutions as well. An island kingdom like Britain, for instance, free of the constant threat of invasion, could militarily afford and commercially benefit from liberal institutions. Prussia, by contrast, relatively poor and surrounded by potential enemies, required for its survival as a state rigorous centralization and militarization.

The primacy of foreign policy was especially plausible to historians immersed in the diplomacy of medieval and early modern Europe, when foreign policy was a virtual monopoly of the prince and his advisers. The rationalist bias of the Enlightenment reinforced the notion of the international state-system as a kind of self-regulating Newtonian universe in which states revolved about each other in alliance or war according to natural laws of self-interest and balance of power. A wise ruler like Frederick II the Great of Prussia saw himself as "the first servant of the state" and made policy according to *raison d'état,* the prudent and rational dictate of dynastic interest. This model of the international system, while reductionist, was not determinist, since it made room for the wisdom and folly, courage and cowardice of individual rulers.

But the debate over the origins of World War I, and the failure of documentary reconstruction of the diplomatic narrative to resolve the question of responsibility for 1914, threw diplomatic history into a crisis. By the late 1920s historians like Sidney Fay and Pierre Renouvin were looking beyond the documents for the deeper or long-range causes of the war, such as militarism or imperialism. Historians influenced by sociology and economics, in turn, located the seeds of the fateful foreign policies preceding the war in the economic and social conflicts of prewar Europe. A young German, Eckhart Kehr, turned Ranke on his head by postulating a "primacy of domestic policy" and argued that a state's foreign policy derives from domestic social and political forces, not vice versa. In particular, imperialism and militarism were seen to be defensive strategies by which threatened elites attempted to rally their people against a foreign threat to distract them from social tensions at home. But if the old diplomatic history was simplistic and dangerous in its glamorization of the exercise of power, the primacy of domestic policy tended to the opposite extreme, ignoring the fact that governments are obliged to respond to real pressures from abroad regardless of their domestic situation. An empirical approach, therefore, is to examine the internal causes of foreign policy in all states and also the effects of those policies on all other states as they are transmitted through the international system.

Ultimately, the conduct and analysis of diplomacy and war rest on a calculus of the power of each state in the system and of its perception by others. National power is the product of all those assets, human and material, that contribute to a state's ability to influence the behaviour of other states by force, threat, or inducement. Human sources of power include population, educational level and work discipline, morale, motivation (through ideology, patriotism, or charismatic leadership), and skill in military and civil administration. Material resources include land area and climate, geographic location, raw materials, and agricultural resources. Last but not least is technology, which is a function of both human and material resources and which can alter the importance of population and geography and render once-effective administrative systems obsolete. Despite the best efforts of political scientists and military planners, these elements of national power (especially human resources) are difficult to quantify and compare. Hence misperception by one state of another's capabilities and intentions is almost the rule rather than the exception. This is why those elusive assets, prestige and intelligence, are sometimes decisive in diplomacy and war.

International relations are shaped primarily by those states perceived to be Great Powers, countries whose interests and capabilities transcend their own self-defense or region. For some 200 years after the treaties of Utrecht and Nystad (1713–14, 1721), the roster of the Great Powers included the same five states: Britain, France, Prussia (and, later, Germany), the Habsburg monarchy (Austria), and Russia. But a mere three decades after World War I only one of these venerable powers, Britain, had not undergone two or more radical changes of government, and only one, Russia, was still a Great Power. Between 1914 and 1945 the European system committed suicide, and two global superpowers rose to replace it. And four decades after 1945, the ability of the United States and the Union of Soviet Socialist Republics to control events around the world was in turn challenged from many sources, giving rise to speculation that the world might be in the early stages of a shift back into a multipolar balance-of-power system.

This article is intended to provide a single integrated narrative of world diplomacy and politics from the outbreak of World War I to the 1980s. Its twin themes are the rivalries of the Great Powers during the age of the world wars and the Cold War, and the replacement, largely through the agency of those wars, of the European state system by a world system with many centres of both power and discord. Because domestic affairs figure heavily in the analysis of each state's foreign policies, the reader should consult the histories of the individual countries for more detail.

For discussion of the military strategy, tactics, and conduct of World War I and World War II, see WORLD WARS, THE. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 963, 96/10, 96/11, and all of Part Nine, Division VII, especially sections 971, 972, 975, 976, and 978.

This article is divided into the following sections:

The roots of World War I, 1871–1914 733
The Bismarckian system, 1871–90 733
The impact of industrialism and imperialism 734
Completing the alliance systems, 1890–1907 736
Militarism and pacifism before 1914 738
The Balkan crises and the outbreak of war, 1907–14 738
The war-guilt question 740
World War I, 1914–18 742
Military stalemate and new belligerents 742
War mobilization at home and abroad 745
War aims and peace feelers 745
The crises of 1917 747
Last battles and armistice 748
The collapse of the old order 750

Peacemaking, 1919–22 750
Competing visions of stability 751
The Versailles *Diktat* 752
The West and the Russian Civil War 754
Central Europe and the Middle East 755
The new balance in East Asia 757
The postwar guilt question 758
A fragile stability, 1922–29 759
A broken world 759
Reparations, security, and the German question 760
The agreements of mid-decade 762
Italy and east-central Europe 763
The invention of Soviet foreign policy 763
The United States, Britain, and world markets 765
The Locarno era and the dream of disarmament 766
The origins of World War II, 1929–39 767
The economic blizzard 767
The rise of Hitler and fall of Versailles 769
British appeasement and American isolationism 771
Anschluss and the Munich Pact 773
Technology, strategy, and the outbreak of war 774
Hitler's war or Chamberlain's? 776
World War II, 1939–45 776
The last European war, 1939–41 777
Origins of American belligerence 780
The turning point, 1942 782
The economic and scientific wars 782
Strategy and diplomacy of the Grand Alliance 785
The defeat of Nazi Germany 786
The defeat of Japan 788
The coming of the Cold War, 1945–57 789
The Cold War guilt question 789
Wasteland: the world after 1945 790
The Cold War in Europe 791
The Cold War in the Middle East and Asia 793
The pace of European integration 796
Nuclear weapons and the balance of terror 798
Total Cold War and the diffusion of power, 1957–72 799
The world after Sputnik 799
Superpower relations in the 1960s 802
The Europe of the fatherlands 803
Asia beneath the superpowers 805
The war in Southeast Asia 806
Nixon, Kissinger, and the détente experiment 807
Dependence and disintegration in the global village 809
The decline of détente 810
The "arc of crisis" 813
Rhetorical Cold War revived 815
Latin-American upheavals 817
The Gorbachev era 818
Forces that shape world politics 819
Tour d'horizon 821
Bibliography 822

The roots of World War I, 1871–1914

Forty-three years of peace among the Great Powers of Europe came to an end in 1914, when an act of political terrorism provoked two great alliance systems into mortal combat. The South Slav campaign against Austrian rule in Bosnia, culminating in the assassination of the Habsburg heir-apparent at Sarajevo, was the immediate spark. But this local crisis rapidly engulfed all the powers of Europe through the mechanisms of the Triple Alliance and the Triple Entente, diplomatic arrangements meant precisely to enhance the security of their members and deter potential aggressors. The long-term causes of the war can therefore be traced back to the forces that impelled the formation of those alliances, increased tensions among the Great Powers, and made at least some European leaders desperate enough to seek their objectives by force, even at the risk of a general war. These long-range forces include militarism and mass mobilization, instability in domestic and international politics occasioned by rapid industrial growth, global imperialism, popular nationalism, and the rise of a social Darwinist worldview. But the question of why World War I broke out should be considered together with the questions of why peace ended, and why in 1914 rather than before or after.

THE BISMARCKIAN SYSTEM, 1871–90

The era of the Great Powers. The European map and world politics were less confused in the decades after 1871 than at any time before or since. The unifications of Italy and Germany removed the congeries of Central European principalities that dated back to the Holy Roman Empire, while the breakup of eastern and southeastern Europe into small and quarreling states (a process that would yield the term balkanization) was not far advanced. There the old empires, Russian, Austro-Hungarian, and Ottoman (Turkish), still prevailed. The lesser powers of Europe, including some that once had been great, like The Netherlands, Sweden, and Spain, played little or no role in the affairs of the Great Powers unless their own interests were directly involved. Both physical size and the economies of scale important in an industrial age rendered smaller and less developed countries impotent, while the residual habits of diplomacy dating from the Congress of Vienna of 1815, made the Great Powers the sole arbiters of European politics.

In the wider world, a diplomatic system of the European variety existed nowhere else. The outcome of the U.S. Civil War and Anglo-American settlement of the Canadian border ensured that North America would not develop a multilateral balance-of-power system. South and Central America had splintered into 17 independent republics following the final retreat of Spanish rule in 1820; but the new Latin-American states were inward-looking, their centres of population and resources isolated by mountains, jungle, and sheer distance, and disputes among them were of mostly local interest. The Monroe Doctrine, promulgated by the United States and enforced by the British Navy, sufficed to spare Latin America new European adventures, the only major exception—Napoleon's III's gambit in Mexico—occurring while the United States was preoccupied with civil war. When the United States purchased Alaska from the Russian tsar and Canada acquired dominion status, both in 1867, European possessions on the American mainland were reduced to three small Guianan colonies in South America. North Africa was still nominally under the aegis of the Ottoman sultan, while sub-Saharan Africa, apart from a few European ports on the coast, was terra incognita. The British had regularized their hold on the Indian subcontinent after the Indian Mutiny of 1857–58, while the Chinese and Japanese empires remained xenophobic and isolationist. Thus the cabinets of the European Great Powers were at the zenith of their influence.

Europe itself, by 1871, seemed to be entering an age of political and social progress. Britain's Second Reform Bill (1867), the French Third Republic (1875), the triumph of nationalism in Italy and Germany (1871), the establishment of universal manhood suffrage in Germany (1867), equality for the Hungarians in the Habsburg monarchy (1867), emancipation of the serfs in Russia (1861), and the adoption of free trade by the major European states all seemed to justify faith in the peaceful evolution of all Europe toward liberal institutions and prosperity.

International peace also seemed assured once Otto von Bismarck declared the new German Empire a satisfied power and placed his considerable talents at the service of stability. The chancellor knew Germany to be a military match for any rival but feared the possibility of a coalition. Since France would never be reconciled to her reduced status and the loss of Alsace-Lorraine imposed by the treaty ending the Franco-German War, Bismarck strove to keep France isolated. In 1873 he conjured up the ghost of monarchical solidarity and formed a Dreikaiserbund (Three Emperors' League) with Austria-Hungary and Russia. But such a combination was always vulnerable to Austro-Russian rivalry over the Eastern Question—the problem of how to organize the Balkan Peninsula as the feuding nationalities there gradually freed themselves from the decrepit Ottoman Empire.

The Three Emperors' League

After the Slavic provinces of Bosnia and Hercegovina rebelled against Ottoman rule in 1875, and Russia made war on the Ottoman Empire two years later, the Dreikaiserbund collapsed. Bismarck achieved a compro-

mise at the Congress of Berlin (1878), but Austro-Russian amity was not restored. In 1879, therefore, Bismarck concluded a permanent peacetime military alliance with Austria, whereupon the tsarist government, to court German favour, agreed to a renewal of the Dreikaiserbund in 1881. Italy, seeking aid for her Mediterranean ambitions, joined Germany and Austria-Hungary to form the Triple Alliance in 1882.

Balkan diplomacy

The next Balkan crisis, which erupted in Bulgaria in 1885, again tempted Russia to expand its influence to the gates of Constantinople. Bismarck dared not oppose the Russians lest he push them toward an alliance with vengeful France. So instead he played midwife to an Anglo-Austro-Italian combination called the Second Mediterranean Entente, which blocked Russian ambitions in Bulgaria while Bismarck himself concluded a Reinsurance Treaty with St. Petersburg in 1887. Once more the Eastern Question had been defused and Germany's alliances preserved.

The nature of the German state. The generation of peace after 1871 rested on Germany's irenic temper, served in turn by Bismarck's statesmanship. Should that temper change, or less adept leadership succeed Bismarck, Germany had the potential to become the major disrupter of European stability. For the constitution drafted by Bismarck for the Second Reich was a dysfunctional document designed to satisfy middle-class nationalism while preserving the power of the Prussian crown and the Junker class (the Prussian landed aristocracy). Apparently a federal empire, Germany was in fact dominated by Prussia, which was larger in area and population than all the other states combined. The king of Prussia was kaiser and chief warlord of the German armies; the prime minister of Prussia was the federal chancellor, responsible, not to a majority in the Reichstag, but only to the crown. Furthermore, Prussia retained a three-class voting system weighted in favour of the wealthy. The army remained, in Prussian tradition, virtually a state within the state, loyal to the kaiser alone. In sum, Germany remained a semi-autocratic military monarchy even as it blossomed into an industrial mass society. The lack of outlets for popular dissent and reform was especially damaging given the cleavages that continued to plague Germany after unification: Protestant North versus Catholic South; agriculture versus industry; Prussia versus the other states; Junkers versus middle-class liberals; industrialists versus the (increasingly Socialist) working class. Bismarck manipulated the parties and interests as he did foreign powers. But toward the end of his tenure, even he realized that German politics might someday reduce to a choice between surrender of privilege by the old elites or a coup d'état against the liberal and Socialist groups he labeled *Reichsfeinde* (enemies of the Reich).

Austria-Hungary and Russia, still overwhelmingly agrarian, faced different challenges by the end of the 19th century. Most acute for Austria-Hungary was the nationality question. An heir to the universalist vision of the Holy Roman Empire, Austria-Hungary was a multinational empire composed not only of Germans and Magyars but also of (in 1870) 4,500,000 Czechs, 3,100,000 Ruthenes, 2,400,000 Poles, 2,900,000 Romanians, 3,000,000 Serbo-Croats, and 600,000 Italians. In this age of nationalist ideology, the Habsburgs faced the challenge of accommodating the yearnings of their ethnic minorities without provoking the dissolution of their empire. In British, French, and, increasingly, Russian opinion, Austria-Hungary was simply out of step with the times, moribund, and, after Turkey, the most despised of states. Bismarck, however, saw Austria-Hungary as "a European necessity": the organizing principle in an otherwise chaotic corner of Europe, the bulwark against Russian expansion, and the keystone in the balance of power. But the progress of nationalism among the peoples of eastern Europe gradually undermined the legitimacy of the old empires. Ironically, Austria existed from 1815 to 1914 in a symbiotic relationship with her ancient enemy, the Ottoman Empire. For as the Balkan peoples gradually pulled free from Constantinople, they and their cousins across the Habsburg frontier inevitably agitated for liberation from Vienna as well.

Austria-Hungary's challenge

Russia was also a multinational empire, but with the exception of the Poles her subject peoples were too few compared to Great Russians to pose a threat. Rather, Russia's problem in the late 19th century was backwardness. Ever since the humiliating defeat in the Crimean War, tsars and their ministers had undertaken reforms to modernize agriculture, technology, and education. But the Russian autocracy, making no concession to popular sovereignty and nationality, was more threatened by social change even than the Germans. Hence the dilemma of the last tsars: They had to industrialize in order to maintain Russia as a Great Power, yet industrialization, by calling into being a large technical and managerial class and an urban proletariat, also undermined the social basis of the dynasty.

In sum, the decades after 1871 did not sustain the liberal progress of the 1860s. Resistance to political reform in the empires, a retreat from free trade after 1879, the growth of labour unions, revolutionary socialism, and social tensions attending demographic and industrial growth all affected the foreign policies of the Great Powers. It was as if, at its pinnacle of achievement, the very elements of liberal "progress"—technology, imperialism, nationalism, cultural modernism, and scientism—were inviting Europeans to steer their civilization toward calamity.

THE IMPACT OF INDUSTRIALISM AND IMPERIALISM

Patterns of population. European demographic and industrial growth in the 19th century was frantic and uneven, and both qualities contributed to growing misperceptions and paranoia in international affairs. European population grew at the rate of 1 percent per year in the century after 1815, an increase that would have been disastrous had it not been for the outlet of emigration and the new prospects of employment in the rapidly expanding cities. But the distribution of Europe's peoples changed radically, altering the military balance among the Great Powers. In the days of Louis XIV, France was the most populous—and also the wealthiest—kingdom in Europe, and as late as 1789 it numbered 25,000,000 to Britain's 14,500,000. When the French Revolution unleashed this national power through rationalized central administration, meritocracy, and a national draft based on patriotism, it achieved unprecedented organization of force in the form of armies of millions of men.

The French tide receded, at the cost of more than a million deaths from 1792 to 1815, never to crest again. Population growth in France, alone among the Great Powers, was almost stagnant thereafter; by 1870 her population of 36,000,000 was nearly equal to that of Austria-Hungary and already less than Germany's 41,000,000. By 1910 Germany's population exploded to a level two-thirds greater than France's, while vast Russia's population nearly doubled from 1850 to 1910 until it was more than 70 percent greater than Germany's, although Russia's administrative and technical backwardness offset to a degree her advantage in numbers. The demographic trends clearly traced the growing danger for France vis-à-vis Germany and the danger for Germany vis-à-vis Russia. Should Russia ever succeed in modernizing, she would become a colossus out of all proportion to the European continent.

The decline of France

Population pressure was a double-edged sword dangling out of reach above the heads of European governments in the 19th century. On the one hand, fertility meant a growing labour force and potentially a larger army. On the other hand, it threatened social discord if economic growth or external safety valves could not relieve the pressure. The United Kingdom adjusted through urban industrialization on the one hand and emigration to the United States and the British dominions on the other. France had no such pressure but was forced to draft a higher percentage of its manpower to fill the army ranks. Russia exported perhaps 10,000,000 excess people to its eastern and southern frontiers and several million more (mostly Poles and Jews) overseas. Germany, too, sent large numbers abroad, and no nation provided more new industrial employment from 1850 to 1910. Still, Germany's landmass was small relative to Russia's, her overseas possessions unsuitable to settlement, and her sense of beleaguerment acute in the face of the "Slavic threat." Demographic trends thus

Emigration as a safety valve

helped to implant in the German population a feeling of both momentary strength and looming danger.

Industry, technology, and trade. Industrial trends magnified the demographic, for here again Germany was far and away the fastest growing economic power on the Continent. This was so not only in the basic industries of coal and iron and steel but also in the advanced fields of electricity, chemicals, and internal combustion. Germany's swift development strained the traditional balance of power in her own society and politics. By the end of the century Germany had become a highly urbanized, industrial society, complete with large, differentiated middle and factory proletariat classes, but it was still governed largely by pre-capitalist aristocrats increasingly threatened by demands for political reform.

Industrialization also made possible the outfitting and supply of mass armies drawn from the growing populations. After 1815 the monarchies of Europe had shied away from arming the masses in the French revolutionary fashion, and the events of 1848 further justified their fear of an armed citizenry. But in the reserve system Prussia found a means of making possible a rapid mobilization of the citizenry without the risk to the regime or the elite officer corps posed by a large standing, and idle, army. (In Austria-Hungary the crown avoided disloyalty in the army by stationing soldiers of one ethnic group on the soil of another.) After Prussia's stunning victory over France in 1871, all the Great Powers came sooner or later to adopt the German model of a mass army, supplied by a national network of railways and arms industries coordinated in turn by a general staff. The industrialization of war meant that planning and bureaucracy, technology and finance were taking the place of bold generalship and esprit in the soldier's craft.

The final contribution to the revolution in warfare was planned research and development of weapons systems. Begun hesitantly in the French Navy in the 1850s and 1860s, command technology—the collaboration of state and industry in the invention of new armaments—was widely practiced by the turn of the century, adding to the insecurity that inevitably propelled the arms races. The demographic, technical, and managerial revolutions of the 19th century, in sum, made possible the mobilization of entire populations and economies for the waging of war.

The Pax Britannica

The home of the Industrial Revolution was Great Britain, whose priority in the techniques of the factory system and of steam power was the foundation for a period of calm confidence known (with some exaggeration) as the Pax Britannica. The pound sterling became the preferred reserve currency of the world and the Bank of England the hub of international finance. British textiles, machinery, and shipping dominated the markets of Asia, South America, and much of Europe. The British Isles (again with some hyperbole) were "the workshop of the world" and in consequence from 1846 led the world in promoting free trade. British diplomacy, proudly eschewing alliances in favour of "splendid isolation," sought to preserve a balance of power on the Continent and to protect the routes to India from Russian encroachment in the Middle East or Afghanistan.

The Pax Britannica could last only as long as Britain's industrial hegemony. But that hegemony very naturally impelled other nations somehow to catch up, in the short term by imposing protective tariffs to shield domestic industries and in the longer term by granting government subsidies (for railroads and other national development work) and the gradual replication of British techniques. First Belgium, France, and New England, then Germany and other states after 1850 began to challenge Britain's industrial dominance.

France (1860), Prussia (1862), and other countries then reversed earlier policies and followed the British into free trade. But in 1873 a financial panic, attributed by some to overextension in Germany after receipt of France's billion-franc indemnity, ended the period of rapid growth. In the depression of 1873–96 (actually years of slower, uneven growth) industrial and labour leaders formed cartels, unions, and lobbies to agitate for tariffs and other forms of state intervention to stabilize the economy. Bismarck resisted until European agriculture also suffered from falling prices and lost markets after 1876 owing to the arrival in European ports of North American cereals. In 1879 the so-called alliance of rye and steel voted a German tariff on foreign manufactured goods and foodstuffs. Free trade gave way to an era of neo-mercantilism. France, Austria, Italy, and Russia followed the new (or revived) trend toward tariff protection. After 1896 the volume of world trade rose sharply again, but the sense of heightened economic competition persisted in Europe.

The alliance of rye and steel

Social rifts also hardened during the period. Challenged by unrest and demands for reforms, Bismarck sponsored the first state social insurance plans, but he also used an attempt on the Kaiser's life in 1878 as a pretext to outlaw the Social Democratic Party. Conservative circles, farmers as well as the wealthier classes, came gradually to distrust the loyalty of the urban working class, but industrialists shared few other interests with farmers. Other countries faced similar divisions between town and country, but urbanization was not advanced enough in Russia or France for socialism to acquire a mass following, while in Britain agriculture had long since lost out to the commercial and industrial classes, and the working class participated fully in democratic politics. The social divisions attending industrialization were especially acute in Germany because of the rapidity of her development and the survival of powerful pre-capitalist elites. Moreover, the German working class, while increasingly unionized, had few legal means of affecting state policy. All this made for a series of deadlocks in German politics that would increasingly affect foreign policy after Bismarck's departure.

The New Imperialism. The 1870s and 1880s, therefore, witnessed a retreat from the free market and a return to state intervention in economic affairs. The foreign counterpart to this phenomenon was the New Imperialism. The Great Powers of Europe suddenly shook off almost a century of apathy toward overseas colonies and, in the space of 20 years, partitioned almost the entire uncolonized portion of the globe. Theories postulating Europe's need to export surplus capital do not fit the facts. Only Britain and France were capital-exporting countries in 1880, and in years to come their investors preferred to export capital to other European countries (especially Russia) or the Western Hemisphere rather than to their own colonies. The British remained free-trade throughout the era of the New Imperialism, a booming home economy absorbed most German capital, and Italy and Russia were large net importers of capital. Once the scramble for colonies was complete, pressure groups did form in the various countries to argue the economic promise of imperialism, but just as often governments had to foster colonial development. In most cases, trade did not lead but followed the flag.

Imperialism and capital

Why, then, was the flag planted in the first place? Sometimes it was to protect economic interests, as when the British occupied Egypt in 1882, but more often it was for strategic reasons or in pursuit of national prestige. One necessary condition for the New Imperialism, often overlooked, is technological. Prior to the 1870s Europeans could overawe native peoples along the coasts of Africa and Asia but lacked the firepower, mobility, and communications that would have been needed to pacify the interior. (India was the exception, where the British East India Company exploited an anarchic situation and allied itself with selected native rulers against others.) The tsetse fly and the *Anopheles* mosquito—bearers of sleeping sickness and malaria—were the ultimate defenders of African and Asian jungles. The correlation of forces between Europe and the colonizable world shifted, however, with the invention of shallow-draft riverboats, the steamship and telegraph, the repeater rifle and Maxim gun, and the discovery (in India) that quinine is an effective prophylactic against malaria. By 1880 small groups of European regulars, armed with modern weapons and exercising fire discipline, could overwhelm many times their number of native troops.

The scramble for Africa

The scramble for Africa should be dated, not from 1882, when the British occupied Egypt, but from the opening of the Suez Canal in 1869. The strategic importance of that

waterway cannot be overstated. It was the gateway to India and East Asia and hence a vital interest nonpareil for the British Empire. When the Khedive of Egypt defaulted on loans owed to France and Britain, and a nationalist uprising ensued—the first such Arab rebellion against the Western presence—the French backed away from military occupation, although with Bismarck's encouragement and moral support they occupied Tunis in 1881, expanding their North African presence from Algeria. Prime Minister William Ewart Gladstone, otherwise an adamant anticolonialist, then established a British protectorate in Egypt. When the French reacted bitterly, Bismarck further encouraged French colonial expansion in hopes of distracting them from Europe, and he then took his own country into the fray by claiming four large segments of Africa for Germany in 1884. In that year the King of the Belgians cast his eye on the entire Congo Basin. The Berlin West Africa Conference of 1884–85 was called to settle a variety of disputes involved in European colonial occupation, and over the next 10 years all the Great Powers of Europe save Austria-Hungary staked out colonies and protectorates on the African continent. But whatever the ambitions and rivalries of military adventurers, explorers, and private empire-builders on the scene, the cabinets of Europe came to agreements on colonial boundaries with surprising neighbourliness. Colonial wars did ensue after 1894, but never between two European colonial powers.

It has been suggested that imperial rivalries were a long-range cause of World War I. It has also been said that they were a safety valve, drawing off European energies that might otherwise have erupted in war much sooner. But the links between imperialism and the war are more subtle. The heyday of the New Imperialism, especially after 1894, created a tacit understanding in the European elites and the broad literate classes that the days of the old European balance of power were over, that a new world order was dawning, and that any nation left behind in the pursuit of world power would sink into obscurity. This intuition must surely have fed a growing sense of desperation among Germans, and one of paranoia among Britons, about trends in global politics. A second point, subtler still, is that the New Imperialism, while it did not directly provoke World War I, did occasion a transformation of alliances that proved dangerous beyond reckoning once the Great Powers turned their attention back to Europe.

The impact of social Darwinism

Charles Darwin published *The Origin of Species* in 1859, and within a decade popularizers had applied—or misapplied—his theories of natural selection and survival of the fittest to contemporary politics and economics. This pseudoscientific social Darwinism appealed to educated Europeans already demoralized by a century of higher criticism of religious scripture and conscious of the competitiveness of their own daily lives in that age of freewheeling industrial capitalism. By the 1870s books appeared explaining the outcome of the Franco-German War, for instance, with reference to the "vitality" of the Germanic peoples by comparison to the "exhausted" Latins. Pan-Slavic literature extolled the youthful vigour of that race, of whom Russia was seen as the natural leader. A belief in the natural affinity and superiority of Nordic peoples sustained Joseph Chamberlain's conviction that an Anglo-American–German alliance should govern the world in the 20th century. Vulgar anthropology explained the relative merits of human races on the basis of physiognomy and brain size, a "scientific" approach to world politics occasioned by the increasing contact of Europeans with Asians and Africans. Racialist rhetoric became common currency, as when the Kaiser referred to Asia's growing population as "the yellow peril" and spoke of the next war as a "death struggle between the Teutons and Slavs." Poets and philosophers idealized combat as the process by which nature weeds out the weak and improves the human race.

By 1914, therefore, the political and moral restraints on war that had arisen after 1789–1815 were significantly weakened. The old conservative notion that established governments had a heavy stake in peace lest revolution engulf them, and the old liberal notion that national unity, democracy, and free trade would spread harmony, were all but dead. The historian cannot judge how much social Darwinism influenced specific policy decisions, but a mood of fatalism and bellicosity surely eroded the collective will to peace.

COMPLETING THE ALLIANCE SYSTEMS, 1890–1907

Germany's new course. In 1890 the young Kaiser William II dismissed the aged Bismarck and proclaimed a new course for Germany. An intelligent but unstable man who compensated for a withered arm with military demeanour and intemperate remarks, William felt keenly his realm's lack of prestige in comparison with the British Empire. William rejected Bismarck's emphasis on security in Europe in favour of a flamboyant *Weltpolitik* (world policy) aimed at making Germany's presence abroad commensurate with her new industrial might. Where Bismarck considered colonies a dangerous luxury given Germany's geographic position, the Kaiser thought them indispensable for Germany's future. Where Bismarck sought alliances to avoid the risk of war on two fronts, the Kaiser (and his chief foreign policy official, Baron von Holstein) believed Germany should capitalize on the colonial quarrels among France, Britain, and Russia. Where Bismarck had outlawed the Socialists and feared for the old order in Germany, the Kaiser permitted the anti-Socialist laws to lapse and believed he could win over the working class through prosperity, social policy, and national glory.

The consequences of the new course were immediate and damaging. In 1890 Holstein gratuitously dropped Bismarck's Reinsurance Treaty with Russia, prompting St. Petersburg to overcome its antipathy to republican France and conclude a military alliance in 1894. The tie was sealed with a golden braid: Between 1894 and 1914 the Russians floated billions of francs in loans on the Paris market to finance factory building, arms programs, and military railroads to the German border. Russia hoped mainly for French support in its colonial disputes with the British Empire and even went so far as to agree with Austria-Hungary in 1897 to hold the question of the Balkans in abeyance for 10 years, thereby freeing resources for the construction of the Trans-Siberian Railroad and the penetration of northern China. The German foreign office thus did not take alarm at the alliance Bismarck had struggled so long to prevent.

The Dual Alliance

The Sino-Japanese War of 1894–95 signaled the arrival of Japan on the world stage. Having seen their nation forcibly opened to foreign influence by Commodore Matthew C. Perry in 1853, the Japanese determined not to suffer China's fate as a hapless object of Western incursion.Once the Meiji Restoration established strong central government beginning in 1868, Japan became the first non-Western state to launch a crash program of industrialization. By the 1890s its modern army and navy permitted Japan to take its place beside the Europeans as an imperial power. In the war with China, Japan won control of Korea, Taiwan, Port Arthur on the Manchurian mainland, and other advantages. European intervention scaled back these gains, but a scramble for concessions in China eventuated. Russia won concessions in Manchuria, the French in South China, the Germans at Chiao-chou Bay on the Shantung Peninsula. In 1898 the United States annexed the Philippine Islands after the Spanish–American War. The loser in the scramble, besides China, was Britain, which had previously enjoyed a near monopoly in the China trade.

The threats to Britain's empire. British fortunes suffered elsewhere during this high tide of imperialism from 1897 to 1907. The South African, or Boer, War (1899–1902) against the independent Boer republics of the South African interior proved longer and costlier than the British expected, and although they won the "dirty little war" the British saw their world position erode. Germany partitioned Samoa with the United States, and the latter annexed the Hawaiian Islands. Germany abandoned her long apathy toward the Middle East and won a concession for Turkish railroads. The Kaiser, influenced by his envy of Britain, his own fondness for seafaring, and the worldwide impact of *The Influence of Sea Power upon History* by the American naval scholar Captain Alfred Thayer Mahan, determined that *Weltpolitik* was impossible without

The Boer War

a great High Seas Fleet. The prospect of a large German navy—next to the growing fleets of France, Russia, Japan, and the United States—meant that Britain would no longer rule the waves alone.

The dawn of the 20th century was thus a time of anxiety for the British Empire as well. Challenged for the first time by the commercial, naval, and colonial might of many other industrializing nations, the British reconsidered the wisdom of splendid isolation. To be sure, in the Fashoda Incident of 1898 Britain succeeded in forcing France to retreat from the upper reaches of the Nile. But how much longer could Britain defend her empire alone? Colonial Secretary Joseph Chamberlain began at once to sound out Berlin on the prospect of global collaboration. A British demarche was precisely what the Germans had been expecting, but three attempts to reach an Anglo-German understanding, between 1898 and 1901, led to naught. In retrospect, it is hard to see how it could have been otherwise. The German foreign minister and, from 1900, chancellor, Bernhard, Fürst von Bülow, shared the Kaiser's and Holstein's ambitions for world power. If, as Germany's neo-Rankean historians proclaimed, the old European balance of power was giving way to a new world balance, then the future would surely belong to the Anglo-Saxons (British Empire and America) and Slavs (Russian Empire) unless Germany were able to achieve its own place in the sun. Bülow agreed that "our future lies on the water." German and British interests were simply irreconcilable. What Britain sought was German help in reducing Franco-Russian pressure on the British Empire and defending the balance of power. What Germany sought was British neutrality or cooperation while Germany expanded its own power in the world. Bülow still believed in Holstein's "free hand" policy of playing the other powers off against each other and accordingly placed a high price on German support and invited Britain to join the Triple Alliance as a full military partner. Understandably, the British declined to underwrite Germany's continental security.

Anglo-German naval competition

The failure of the Anglo-German talks condemned both powers to dangerous competition. The German navy could never hope to equal the British and would only ensure British hostility. But equality was not necessary, said Admiral Alfred von Tirpitz. All Germany needed was a "risk fleet" large enough to deter the British, who would not dare alienate Germany and thus lose their only potential ally in the continuing rivalry with France and Russia. In this way Germany could extract concessions from London without alliance or war. What the Germans failed to consider was that Britain might someday come to terms with its other antagonists.

This was precisely what Britain did. The Edwardian era (1901–10) was one of intense concern over the decline of Britain's naval and commercial dominance. German firms shouldered aside the British in numerous markets (even though they remained each other's best trading partners). The new German navy menaced Britain in her home waters. The French and Russian fleets, not to mention the Japanese, outnumbered the Royal Navy's Asian squadron. The French, Italian, and potential Russian presence in the Mediterranean threatened the British lifeline to India. Soon the Panama Canal would enable the United States to deploy a two-ocean navy. Accordingly, the foreign secretary, Lord Lansdowne, set about reducing the number of Britain's potential opponents. First, he cemented friendly relations with the United States in the Hay–Pauncefote Treaty (1901). He then shocked the world by concluding a military alliance with Japan, thereby securing British interests in East Asia and allowing the empire to concentrate its regional forces on India. But when growing tension between Russia and Japan over Manchuria appeared likely to erupt in war in 1904, France (Russia's ally) and Britain (now Japan's ally) faced a quandary. To prevent being dragged into the conflict, the French and British shucked off their ancient rivalry and concluded an Entente Cordiale whereby France gave up opposition to British rule in Egypt, and Britain recognized French rights in Morocco. Though strictly a colonial arrangement, it marked another step away from isolation for both Britain and France and another step toward it for the restless and frustrated Germans.

The Entente Cordiale

The Russo-Japanese War of 1904–05 was an ominous turning point. Contrary to all expectations, Japan triumphed on land and sea, and Russia stumbled into the Revolution of 1905. U.S. President Theodore Roosevelt mediated the Treaty of Portsmouth ending the war, and the Tsar quelled the revolutionary flames with promises of parliamentary government, but the war resonated in world diplomacy. Japan established itself as the leading Asian power. The example of an Oriental nation rising up to defeat a European Great Power emboldened Chinese, Indians, and Arabs to look forward to a day when they might expel the imperialists from their midst. And tsarist Russia, its Asian adventure a shambles, looked once again to the Balkans as a field for expansion, setting the stage for World War I.

The Triple Entente. In 1905 the Germans seized on Russia's temporary troubles to pressure France in Morocco. Bülow believed he had much to gain—at best he might force a breakup of the Anglo-French entente, at worst he might provoke a French retreat and secure German rights in Morocco. But at the Algeciras Conference in 1906, called to settle the Morocco dispute, only Austria-Hungary supported the German position. Far from breaking the Entente Cordiale, the affair prompted the British to begin secret staff talks with the French military. The United States, Russia, and even Italy, Germany's erstwhile partner in the Triple Alliance, took France's side. For some years Italian ambitions in the Mediterranean had been thwarted, and the attempt to conquer Abyssinia in 1896 had failed. The German alliance seemed to offer little, while Rome's other foreign objective, the Italian *irredenta* in the Tirol and Dalmatia, was aimed at Austria-Hungary. So in 1900 Italy concluded a secret agreement pledging support for France in Morocco in return for French support of Italy in Libya. The Russo-Japanese War also strengthened ties between France and Russia as French loans again rebuilt Russia's shattered armed forces. Finally, and most critically, the defeated Russians and worried British were now willing to put to rest their old rivalry in Central Asia. The Anglo-Russian Convention of 1907 made a neutral buffer of Tibet, recognized Britain's interest in Afghanistan, and partitioned Persia into spheres of influence. Foreign Secretary Sir Edward Grey also hinted at the possibility of British support for Russian policy in the Balkans, reversing a century-old tradition.

Anglo-Russian entente

The heyday of European imperialism thus called into existence a second alliance system, the Triple Entente of France, Britain, and Russia. It was not originally conceived as a balance to German power, but that was its effect, especially in light of the escalating naval race. In 1906 the Royal Navy under the reformer Sir John Fisher launched HMS *Dreadnought,* a battleship whose size, armour, speed, and gunnery rendered all existing warships obsolete. The German government responded in kind, even enlarging the Kiel Canal at great expense to accommodate the larger ships. What were the British, dependent on imports by sea for seven-eighths of their raw materials and over half their foodstuffs, to make of German behaviour? In a famous Foreign Office memo of January 1907, Senior Clerk Sir Eyre Crowe surmised that *Weltpolitik* was either a conscious bid for hegemony or a "vague, confused, and unpractical statesmanship not realizing its own drift." As Ambassador Sir Francis Bertie put it, "The Germans aim to push us into the water and steal our clothes."

For France the Triple Entente was primarily a continental security apparatus. For Russia it was a means of reducing points of conflict so that the antiquated tsarist system could buy time to catch up technologically with the West. For Britain the ententes, the Japanese alliance, and the "special relationship" with the United States were diplomatic props for an empire beyond Britain's capacity to defend alone. The three powers' interests by no means coincided—disputes over Persia alone might have smashed Anglo-Russian unity if the war had not intervened. But to the Germans, the Triple Entente looked suspiciously like encirclement designed to frustrate their rightful claims to world power and prestige. German attempts to break

the encirclement, however, would only alarm the entente powers and cause them to draw the loose strings into a knot. That in turn tempted German leaders, fearful that time was against them, to cut the Gordian knot with the sword. For after 1907 the focus of diplomacy shifted back to the Balkans, with European cabinets unaware, until it was too late, that alliances made with the wide world in mind had dangerously limited their freedom of action in Europe.

MILITARISM AND PACIFISM BEFORE 1914

Anxiety and the arms race. It is difficult to escape the conclusion that Europe before 1914 succumbed to hubris. The conventional images of "armed camps," "a powder-keg," or "saber rattling" almost trivialize a civilization that combined within itself immense pride in its newly expanding power and almost apocalyptic insecurity about the future. Europe bestrode the world, and yet Lord Curzon could remark, "We can hardly take up our morning newspaper without reading of the physical and moral decline of the race," and the German chief of staff, Helmuth von Moltke, could say that if Germany backed down again on Morocco, "I shall despair of the future of the German Empire." France's stagnant population and weak industry made her statesmen frantic for security, Austrian leaders were filled with foreboding about their increasingly disaffected nationalities, and the tsarist regime, with the most justification, sensed doom.

Whether from ambition or insecurity, the Great Powers armed as never before in peacetime, with military expenditures reaching 5 to 6 percent of national income. Military conscription and reserve systems made available a significant percentage of the adult male population, and the impulse to create large standing armies was strengthened by the widespread belief that firepower and financial limitations would make the next war short and violent. Simple reaction also played a large role. Fear of the "Russian steamroller" was sufficient to expand Germany's service law; a larger German army provoked the outmanned French into an extension of national service to three years. Only Britain did without a large conscripted army, but her naval needs were proportionally more expensive.

The peacetime armies

In an age of heavy, rapid-fire artillery, infantry rifles, and railroads, but not yet including motor transport, tanks, or airplanes, a premium was placed by military staffs on mass, supply, and prior planning. European commanders assumed that in a continental war the opening frontier battles would be decisive, hence the need to mobilize the maximum number of men and move them at maximum speed to the border. The meticulous and rigid advance planning that this strategy required placed inordinate pressure on the diplomats in a crisis. Politicians might hold back their army in hopes of saving the peace only at the risk of losing the war should diplomacy fail. What was more, all the continental powers embraced offensive strategies. The French general staff's "cult of attack" assumed that élan could carry the day against superior German numbers. Its Plan XVII called for an immediate assault on Lorraine. The Germans' Schlieffen Plan addressed the problem of war on two fronts by throwing almost the entire German army into a sweeping offensive through neutral Belgium to capture Paris and the French army in a gigantic envelope. Troops could then be transported east to meet the slower-moving Russian army. Worked out down to the last railroad switch and passenger car, the Schlieffen Plan was an apotheosis of the industrial age: a mechanical, almost mathematical perfection that wholly ignored political factors. None of the general staffs anticipated what the war would actually be like. Had they glimpsed the horrific stalemate in the trenches, surely neither they nor the politicians would have run the risks they did in 1914.

The Schlieffen Plan

Above the mass infantry armies of the early 20th century stood the officer corps, the general staffs, and at the pinnacle the supreme war lords: kaiser, emperor, tsar, and king, all of whom adopted military uniforms as their standard dress in these years. The army was a natural refuge for the central and eastern European aristocracies, the chivalric code of arms sustaining almost the only public service to which they could still reasonably lay claim. Even in republican France a nationalist revival after 1912 excited public morale, inspired the military buildup, and both fueled and cloaked a revanche aimed at recovery of the provinces lost 40 years before. Popular European literature poured forth best-sellers depicting the next war, and mass-circulation newspapers incited even the working classes with news of imperial adventures or the latest slight by the adversary.

The peace movements. Various peace movements sprang up to counter the spirit of militarism before 1914. Most numerous and disturbing to those responsible for national defense were the Socialists. The Second International took the Marxist view of imperialism and militarism as creatures of capitalist competition and loudly warned that if the bosses provoked a war, the working classes would refuse to take part. Jean Jaurès defined the proletariat as "masses of men who collectively love peace and hate war." The 1912 Basel Conference declared the proletariat "the herald of world peace" and proclaimed "war on war." Sober observers like George Bernard Shaw and Max Weber doubted that any putative sense of solidarity among workers would outweigh their nationalism, but the French government kept a blacklist of agitators who might try to subvert mobilization. Some of Germany's leaders imagined that war might provide the opportunity to crush socialism by appeals to patriotism or martial law.

A liberal peace movement with a middle-class constituency flourished around the turn of the century. As many as 425 peace organizations are estimated to have existed in 1900, fully half of them in Scandinavia and most others in Germany, Britain, and the United States. Their greatest achievements were the Hague conferences of 1899 and 1907, at which the powers agreed to ban certain inhumane weapons but made no progress toward general disarmament. The liberal peace movement also foundered on internal contradictions. To outlaw war was to endorse the international status quo, yet liberals always stood ready to excuse wars that could claim progressive ends. They had tolerated the wars of Italian and German unification, and they would tolerate the Balkan Wars against the Ottoman Empire in 1912–13 and the great war in 1914. Another solution for many peace advocates was to transcend the nation-state. Norman Angell's *The Great Illusion* (1910) argued that it already had been transcended: that interdependence among nations made war illogical and counterproductive. To Marxists this image of capitalism was ludicrous; to Weber or Joseph Schumpeter, it was correct, but beside the point. Blood was thicker than class, or money; politics dominated economics; and irrationality, reason.

The Hague conferences

The one European statesman most sympathetic to the peace movements was, not surprisingly, Britain's Liberal foreign secretary, Sir Edward Grey. Citing the waste, social discord, and international tension caused by the naval arms race, he made several overtures to Germany in hopes of ending it. When these failed, Britain had little choice but to race more quickly than the Germans. Even radical Liberals like David Lloyd George had to admit that however much they might deplore arms races in the abstract, all that was liberal and good in the world depended on the security of Britain and its control of its seas.

THE BALKAN CRISES AND THE OUTBREAK OF WAR, 1907–14

Growing tensions and German isolation. In the end, war did not come over the naval race, or commercial competition, or imperialism. Nor was it sparked by the institutional violence of the armed states, but by underground terrorism in the name of an oppressed people. Nor did it come over the ambitions of Great Powers to become greater, but over the fear of one Great Power that unless it took vigorous action it might cease to exist altogether. It began in the Balkans.

In 1897 Austria-Hungary and Russia had agreed to put their dispute over the Balkans on ice. When the agreement ran out in 1907, the Ottoman Empire still ruled Macedonia, ringed by Greece, Montenegro, Serbia, and Bulgaria. But everything else had changed. For now Austria-Hungary's only reliable ally was Germany, whose *Weltpolitik* had led it to join the competition for influence at

The thawing of the Balkans

Constantinople. Russia was looking again at the Balkans for foreign policy advantage and enjoying, for the first time, a measure of British tolerance. In Serbia, the state most threatening to Vienna because of its ethnic tie to the Serbs and Croats inside the Dual Monarchy, a fundamental political shift had occurred. In previous years Vienna had neutralized Serbia by bribing the ruling Obrenović dynasty, but in 1903 the rival Karageorgević clan seized control in Belgrade in a bloody coup d'état and shifted to a violently anti-Austrian policy. Finally, in 1908, a cabal of officers known as the Young Turks staged the first modernizing revolution in the Muslim world and tried to force the Sultan to adopt liberal reforms. In particular the Young Turks called for parliamentary elections, thereby placing in doubt the status of Bosnia and Hercegovina, provinces still under Ottoman sovereignty but administered by Austria-Hungary since 1878. Austro-Hungarian foreign minister Aloys Aehrenthal proposed to settle the Bosnian issue and to crush Serbian ambitions once and for all by annexing the provinces. To this purpose he teased the Russian foreign minister, Aleksandr Petrovich Izvolsky, with talk of a quid pro quo, Russia's acquiescence in annexation and in return Austria-Hungary's in the opening of the Dardanelles Strait to Russian warships. When instead Aehrenthal acted unilaterally, and Izvolsky's straits proposal was rejected, the Russians felt betrayed. Their response was to increase aid and comfort to their client Serbia and to determine never again to back down in the Balkans.

Annexation and the straits question

German politics were also approaching a breaking point. Chancellor von Bülow had governed, with the support of Tirpitz, the Kaiser, and the moderate and conservative parties in the Reichstag, on the basis of a grand compromise of which the navy was the linchpin. Agrarian interests continued to demand protection against foreign foodstuffs, but the tariffs imposed to that end harmed German industrial exports. A large armaments program, especially naval, compensated heavy industry for lost foreign markets. The losers in the tariffs-plus-navy-legislation arrangement were consumers, who were taxed for the defense program after they had paid higher prices for bread. Popular resentment tended to increase the Socialist vote, and the other parties could command a majority only by banding together.

Soon, however, the expensive dreadnought race provoked a fiscal crisis that cracked the Bülow bloc and, in 1909, elevated Theobald von Bethmann Hollweg to the chancellorship. He faced the choice of ending the naval race and moderating Germany's *Weltpolitik,* or making democratic concessions to the left, or somehow rebuilding the coalition of conservative agrarians and industrialists in the teeth of Socialist opposition. Bethmann showed signs of preferring the first course but was undercut by the pressure of industry, Tirpitz' naval propaganda, and the Kaiser's bravado, symbolized by his damaging *Daily Telegraph* interview (1908) in which he made inflammatory remarks about the British. When in 1912 Lord Haldane was dispatched to Berlin to discuss a suspension of the naval arms race, the Kaiser spoiled chances for an accord by introducing a new naval bill two days before Haldane's arrival. The British then accelerated their own dreadnought construction. By now the failure of German policy was apparent. Clearly the British would not permit Germany to challenge their sea power, while the Junker-dominated German army agreed in 1912 to tolerate further naval expansion only if the army were granted a sharp increase in funding as well. In the 1912 elections the Social Democrats won 110 seats and became the largest party in the Reichstag.

Domestic and foreign stalemate obsessed Germany's political and military leadership. Reform at home meant an end to the privileged positions of the various elites; retreat abroad meant the end of Germany's dreams of world power. A bold stroke, even at the risk of war, seemed the only way out of the double impasse. In 1911 Foreign Minister Alfred von Kiderlen-Wächter tried to force the issue in Morocco, where the French clearly aimed at a formal protectorate in defiance of the Algeciras accords. Germany sent the gunboat *Panther* to the Moroccan port of Agadir in defense of "German interests" there. But Britain again stood with France, and Kiderlen-Wächter acquiesced in a French Morocco in exchange for portions of French colonies in Central Africa. In France this accommodation of Germany brought down the government of Premier Joseph Caillaux, who was succeeded by Raymond Poincaré, a determined nationalist and advocate of military preparedness who quickly secured passage of an expansion of the standing army. In Britain, Winston Churchill, then first lord of the Admiralty, withdrew his fleet from the Mediterranean to home waters, making mandatory even closer military coordination with France.

The *Panther* incident

This Second Moroccan Crisis confirmed Germany's isolation, while the British, French, and Russian military buildups meant that time was on the side of the entente. Moltke had already raised the notion of preventive war, and in the Kaiser's war council of December 1912 he blustered, "War, the sooner the better." To be sure, jingoism of this sort could be found in every Great Power on the eve of the war, but only the leaders in Berlin—and soon Vienna—were seriously coming to view war as not simply a possibility but as a necessity.

The final prewar assault on the Ottoman Empire also began in 1911. Italy cashed in her bargain with France over Libya by declaring war on Turkey and sending a naval squadron as far as the Dardenelles. Simultaneously, Russian ministers in the Balkans brought about an alliance between the bitter rivals Serbia and Bulgaria in preparation for a final strike against Ottoman-controlled Europe. The First Balkan War erupted in October 1912, when Montenegro declared war on Turkey, followed quickly by Serbia, Bulgaria, and Greece. The Young Turks ended the conflict with Italy, ceding Libya, but failed to contain the Balkan armies. In May 1913 the Great Powers imposed a settlement; Macedonia was partitioned among the Balkan states, Crete was granted to Greece, and Albania was given its independence. Landlocked Serbia, however, bid for additional territory in Macedonia, and Bulgaria replied with an attack on Serbia and Greece, thus beginning the Second Balkan War in June 1913. In the peace that followed in August, Bulgaria lost all her stake in the former Turkish lands plus much of the southern Dobruja to Romania. Serbia, however, doubled its territory and, flushed with victory, turned its sights on the Austro-Hungarian provinces of Bosnia and Hercegovina.

The Balkan wars

The final crisis. How might the Habsburg Empire survive the rise of particularist nationalism in eastern Europe? Austrian statesmen had debated the question for 50 years, and the best answer seemed to be some form of federalism permitting political autonomy to the nationalities. Reforms of this nature had always been vetoed by the Hungarians, who stood to lose their own position vis-à-vis the German-Austrians and the minorities in their half of the empire. Conrad Franz, Graf von Hötzendorf, chief of the general staff, favoured preventive war against Serbia to stifle nationalist agitation for good and reinforce the old order. However, Archduke Francis Ferdinand wrote: "I live and shall die for federalism; it is the sole salvation for the monarchy, if anything can save it." Out of favour with the court for his morganatic marriage, resented by the Hungarians, and resented by conservatives, the heir-apparent was also feared by Slavic radicals as the one man who might really pacify the nationalities and so frustrate their dreams of a Greater Serbia. Hence the archduke was a marked man among the secret societies that sprang up to liberate Bosnia. Such is the logic of terrorism: Its greatest enemies are the peacemakers.

The National Defense (Narodna Odbrana) was formed in Serbia in 1908 to carry on pro-Serbian and anti-Austrian agitation across the border. Its nonviolent methods were deemed insufficient by others who formed in 1911 the secret society Union or Death (Ujedinjenje ili Smrt), also known as the Black Hand, led by the head of Serbian military intelligence, Colonel Dragutin Dimitrijević. The latter had been involved in the 1903 assassinations of the Obrenović family and favoured terrorist action over intellectual propaganda. With his support, if not on his direct orders, a band of youthful romantics conspired to assassinate Francis Ferdinand during his state visit to Sarajevo. On June 28, 1914, which happened to be the Serbian na-

The assassination of Francis Ferdinand

tional holiday, the Archduke and his wife rode in an open car through the streets of the Bosnian capital. A bomb was thrown but missed. The Archduke completed his official duties, whereupon the governor of Bosnia suggested they deviate from the planned route on the return trip for safety's sake. But the lead driver in the procession took a wrong turn, the cars stopped momentarily, and at that moment the 19-year-old Gavrilo Princip fired his revolver, killing both royal passengers.

Reaction in Vienna, and Europe generally, was surprisingly restrained. No one imagined that the outrage had more than local importance, much less that Bismarck's prophecy about "some damned fool thing in the Balkans" starting the next war was about to be fulfilled. Conrad von Hötzendorf saw the deed as pretext for his preventive war against Serbia, but the aged emperor Francis Joseph preferred to await an inquiry to determine the extent of Serbian complicity. Germany, on the other hand, pressed for a firm riposte and in the Kaiser's famous "blank check" memo promised to support whatever action Austria might take against Serbia. The Germans expected Russia to back down, since its military reforms would not be complete for several years, but even if Russia came to Serbia's aid, the German high command was confident of victory. Bethmann was less so. A move against Serbia could lead to a world war, he warned on July 7. Yet Bethmann went along in the vain hope of localizing the conflict.

Austrian Foreign Minister Leopold, Graf von Berchtold, now advocated a firm policy toward Serbia lest Austria's prestige deteriorate further and the Balkan states unite behind Russia. Gróf Tisza, the prime minister of Hungary, insisted, however, that diplomatic and legal justifications precede such a clash of arms: Austria must first present a list of demands for redress. Should Serbia accept, the empire would win a "brilliant diplomatic success"; should Serbia refuse, war could be waged with Austria-Hungary posing as the aggrieved party. In no case was Austria to annex any Serbian territory.

The Russian response to any Austrian initiative would be critical, and by chance the president and prime minister of France, Poincaré and René Viviani, were paying a state visit to St. Petersburg in July. Strangely, there is no record of the Franco-Russian conversations, but it is known that Poincaré assured the Russians that France would stand by her alliance commitments. On July 23, just after the French leaders left for home, Vienna presented its ultimatum to Belgrade, demanding dissolution of the secret societies, cessation of anti-Austrian propaganda, and Austrian participation in the investigation of the Sarajevo crime. Serbia was given 48 hours to respond.

The Austrian ultimatum

The Russian foreign minister, Sergey Dmitriyevich Sazonov, erupted at news of the ultimatum and insisted on military measures. The French ambassador, Maurice-Georges Paléologue, with or without instructions from his departed chiefs, encouraged Sazonov. For if Austria's prestige—and very future—were at stake in the Balkans, so too were tsarist Russia's, for which the Balkans was the only region left in which to demonstrate its vitality. But now Germany was competing for influence over the Young Turks, courting Bulgaria, and plotting to smash Serbia. The German slogan "From Berlin to Baghdad," referring originally only to railroads, took on ominous new political meaning. On July 25 the Russian Council of Ministers decided that if Austrian forces entered Serbia, Russia would mobilize its army. This precipitous, indeed anticipatory, decision reflected Russia's size and the inadequacy of its rail network. Sazonov seems to have considered mobilization a political threat, but given the mechanistic timetables that were integral to the planning of all the European general staffs, it could only provoke countermobilizations and an inexorable drift into war.

On July 25 Serbia accepted all the Austro-Hungarian conditions save those two that directly compromised its sovereignty. Two days later Berchtold persuaded Francis Joseph to initiate war. At the same moment the Kaiser, returning from a yachting expedition, tried belatedly to restrain Vienna. On July 28 Austria declared war and bombarded Belgrade; on the same day the Tsar approved the mobilization of the Russian army against Austria, and alarms went off all over Europe. Sir Edward Grey, Kaiser William, and the Italian government all proposed negotiations, with the Austrians to occupy Belgrade as a pledge of Serbian compliance. The German ambassador in St. Petersburg assured the Russians that Austria meant to annex no Serbian territory. But it was too little and far too late. In St. Petersburg the generals protested that partial mobilization would disrupt their contingency plans: How could Russia prepare to fight Austria-Hungary while leaving naked her border with Austria's ally Germany? The weak and vacillating Tsar Nicholas II was persuaded, and on the afternoon of July 30 he authorized general mobilization of the Russian army.

The previous day, Poincaré and Viviani had finally arrived back in Paris, where they were met with patriotic crowds and generals anxious for military precautions. In Berlin, anti-Russian demonstrations and equally anxious generals called for immediate action. On the 31st, when all the other powers had begun preparations of some sort, and even the British had put the fleet to sea (thanks to Winston Churchill's foresight), Germany delivered ultimatums to Russia, demanding an end to mobilization, and to France, demanding neutrality in case of war in the east. But Russia and France could scarcely accede without abandoning the Balkans, each other, and their own security. When the ultimatums expired, the Schlieffen Plan was put into effect. Germany declared war against Russia on August 1 and against France on August 3 and demanded safe passage for its troops through Belgium. Refused again, Germany invaded Belgium in force.

German entry into the war

On August 3, Italy took refuge in the fact that this was not a defensive war on Austria-Hungary's part and declared its neutrality. That left only Britain, faced with the choice of joining its entente partners in war or standing aloof and risking German domination of the Continent. Britain had little interest in the Serbian affair, and the kingdom was torn by the Irish question. The Cabinet was in doubt as late as August 2. But the prospect of the German fleet in the English Channel and German armies on the Belgian littoral settled the issue. On the 3rd Britain demanded that Germany evacuate Belgium, and Grey won over Parliament with appeals to British interests and international law. On August 4, Britain declared war on Germany.

THE WAR-GUILT QUESTION

The search for causes. Debate over the origins of World War I was from the start partisan and moral in tone. Each of the belligerents published documentary collections selected to shift the blame and prove that it was fighting in self-defense. Serbia was defending itself against Austrian aggression. Austria-Hungary was defending its very existence against terror plotted on foreign soil. Russia was defending Serbia and the Slavic cause against German imperialism. Germany was defending its lone reliable ally from attack and itself from entente encirclement. France, with most justification, was defending itself against unprovoked German attack. And Britain was fighting in defense of Belgium, international law, and the balance of power.

In the Treaty of Versailles (1919) the victorious coalition justified its peace terms by forcing Germany and its allies to acknowledge guilt for the war. This tactic was historically dubious and politically disastrous, but it stemmed from the liberal conviction, as old as the Enlightenment, that peace was normal and war an aberration or crime for which clear responsibility—guilt—could be established. Almost at once, revisionist historians examined the thousands of documents that governments made available after 1920 and challenged the Versailles verdict. Yes, the German government had issued the risky "blank check" and urged Vienna on an aggressive course. It had swept aside all proposals for mediation until events had gained irreversible momentum. It had, finally, surrendered its authority to a military plan that ensured the war could not be localized. Indeed, the whole course of German foreign policy since 1890 had been restless and counterproductive, calling into existence the very ring of enemies it then took extreme risks to break. But, on the other hand, Russia's hasty mobilization expanded the crisis beyond the Balkans, initiated a round of military moves, and

The "blank check"

contributed to German panic. Given the military realities of the age, Sazanov's notion of Russian mobilization as a mere "application of pressure" was either disingenuous or foolish. France could be faulted for not restraining Russia and for issuing its own "blank check." Even the British might have done more to preserve peace, either through more vigorous mediation or by making clear that they would not remain neutral in a continental war, thus deterring the Germans. Finally, what of the states at the heart of the crisis? Surely Belgrade's use of political terrorism in the name of Greater Serbia, and Austria-Hungary's determination to crush its tormentors, provoked the crisis in the first place. By the 1930s moderate historians had concluded, with Lloyd George, that no one country was to blame for the war: "We all stumbled into it."

The failure of documentary research to settle the war-guilt question led other historians to look behind the July 1914 crisis for long-range causes of the war. Surely, they reasoned, such profound events must have profound origins. As early as 1928 the American Sidney B. Fay concluded that none of the European leaders had wanted a great war and identified as its deeper causes the alliance systems, militarism, imperialism, nationalism, and the newspaper press. (Marxists, of course, from the publication of Lenin's *Imperialism, the Highest Stage of Capitalism* in 1916, held finance capitalism to be accountable for the war.) In this view the polarization of Europe into alliance systems had made "chain-reaction" escalation of a local imbroglio almost predictable. Militarism and imperialism had fed tensions and appetites among the Great Powers, while nationalism and sensational journalism had stoked popular resentments. How else could one explain the universal enthusiasm with which soldiers and civilians alike greeted the outbreak of war? Such even-handed sentiments, along with the abstraction of the terms of analysis that exculpated individuals while blaming the system, were both appealing and prescriptive. In the 1930s British statesmen in particular would strive to learn the lessons of 1914 and so prevent another war. As another generation's hindsight would reveal, the lessons did not apply to the new situation.

After World War II and the Cold War had left the issues of 1914 passé, a committee of French and German historians agreed that World War I was an unwilled disaster for which all countries shared blame. But only a few years later, in 1961, that consensus shattered. A German historian, Fritz Fischer, published a massive study of German war aims during 1914–18 and held that Germany's government, social elites, and even broad masses had consciously pursued a breakthrough to world power in the years before World War I and that the German government, fully aware of the risks of world war and of British belligerency, had deliberately provoked the 1914 crisis. Fischer's thesis sparked bitter debate and a rash of new interpretations of World War I. Leftist historians made connections between Fischer's evidence and that cited 30 years before by Eckhart Kehr, who had traced the social origins of the naval program to the cleavages in German society and the stalemate in the Reichstag. Other historians saw links to the Bismarckian technique of using foreign policy excursions to stifle domestic reform, a technique dubbed "social imperialism." Germany's rulers, it appeared, resolved before 1914 to overthrow the world order in hopes of preserving the domestic order.

The Fischer thesis

Traditionalist critics of Fischer pointed to the universality of imperialistic, social Darwinist, and militaristic behaviour on the eve of the war. The Kaiser, in his most Wagnerian moods, only spoke and acted like many others in all the Great Powers. Did not Sazonov and the Russian generals, in their unrecorded moments, yearn to erase the humiliation of 1905 and conquer the Dardanelles, or Poincaré and General J.-J.-C. Joffre wonder excitedly if the recovery of Alsace-Lorraine were finally at hand, or the Primrose and Navy leagues thrill to the prospect of a Nelsonian clash of dreadnoughts? Germans were not the only people who grew weary of peace or harboured grandiose visions of empire. To this universalist view leftist historians like the American A.J. Mayer then applied the "primacy of domestic policy" thesis and hypothesized that all the European powers had courted war as a means of cowing or distracting their working classes and national minorities.

Such "new left" interpretations triggered intense study of the connections between domestic and foreign policy, leading to the conclusion that a postulation of internal origins of the war, while obvious for Austria and plausible for Russia, failed in the cases of democratic Britain and France. If anything, internal discord made for reticence rather than assertion on the part of their foreign policy elites. The conservative historian Gerhard Ritter even challenged the Fischer thesis in the German case. The real problem, he argued, was not fear of the Social Democrats but the age-old tension between civilian and military influence in the Prussian-German government. Politicians, exemplified by Bethmann, did not share the eagerness or imprudence of the general staff but lost control of the ship of state in the atmosphere of deepening crisis leading up to 1914. Finally, a moderate German historian, Wolfgang J. Mommsen, dispensed with polemics altogether. Germany's rapid industrialization and the tardiness of modernization in Austria-Hungary and Russia, he concluded, created instabilities in central and eastern Europe that found expression in desperate self-assertion. Echoing Joseph Schumpeter, Mommsen blamed the war on the survival of pre-capitalist regimes that simply proved "no longer adequate in the face of rapid social change and the steady advance of mass politics." But this interpretation amounted to an updated and elaborated version of the unsophisticated consensus that "we all stumbled into it." Are world wars, then, beyond human control?

Backwardness as a cause of war

Thus, the search for long-range causes, while turning up a wealth of new information and insight, ran ultimately aground. After all, if "imperialism" or "capitalism" had caused the war, they had just as assuredly caused the unprecedented era of peace and growth that preceded it. Imperialist crises, though tense at times, had always been resolved, and even Germany's ambitions were on the verge of being served through a 1914 agreement with Britain on a planned partition of the Portuguese empire. Imperial politics were simply not a casus belli for anyone except Britain. Military preparedness was at a peak, but armaments are responses to tensions, not the cause of them, and they had, perhaps, served to deter war in the numerous crises preceding 1914. Capitalist activity tied the nations of Europe together as never before, and most leading businessmen were advocates of peace in 1914. The alliance systems themselves were defensive and deterrent by design and had served as such for decades. Nor were they inflexible. Italy opted out of her alliance, the Tsar was not bound to risk his dynasty on behalf of Serbia, or the Kaiser his on behalf of Austria-Hungary, while the French and British cabinets might never have persuaded their parliaments to take up arms had the Schlieffen Plan not forced the issue. Perhaps the 1914 crisis was, after all, a series of blunders, in which statesmen failed to perceive the effects their actions would have on the others.

The centrality of the Habsburg monarchy. Or perhaps a long-range view that is still serviceable is precisely the one derived from old-fashioned analysis of the balance of power system, forgotten amid the debates over national or class responsibility. This view, suggested by Paul Schroeder in 1972, asks not why war broke out in 1914 but why not before? What snapped in 1914? The answer, he argued, is that the keystone of the European balance, the element of stability that allowed the other powers to chase imperial moonbeams at will, was Austria-Hungary itself. But the heedless policies of the other powers gradually undermined the Habsburg monarchy until it was faced with a mortal choice. At that point, the most stable member of the system became the most disruptive, the girders of security—the alliances—generated destructive pressures of their own, and the European system collapsed. To be sure, Austria-Hungary was threatened with her own nationality problem, aggravated by Serbia. But it could better have met that threat if the Great Powers had worked to ameliorate pressures on it, just as they had carried the declining Ottoman Empire for a full century. Instead, the ambitions of Russia, France, and Britain, and the stifling friendship

The monarchy undermined

of Germany, only served to push Austria-Hungary to the brink. This was not their intention, but it was the effect.

The central fact of global politics from 1890 to 1914 was Britain's relative decline. This occurred naturally, as industrial power diffused, but was aggravated by the particular challenge of Germany. Overextended, the British sought partners to share the burdens of a world empire and were obliged in return to look kindly on those partners' ambitions. But the resulting Triple Entente was not the cause of Germany's frustrations in the conduct of *Weltpolitik*. Rather it was the inability of Germany to pursue an imperial policy *à outrance*. Situated in the middle of Europe, with hostile armies on two sides, and committed to the defense of Austria-Hungary, Germany was unable to make headway in the overseas world despite her strength. By contrast, relatively weak France or hopelessly ramshackle Russia could engage in adventures at will, suffer setbacks, and return to the fray in a few years. Schroeder concluded: "The contradiction between what Germany wanted to do and what she dared to do and was obliged to do accounts in turn for the erratic, uncoordinated character of German world policy, its inability to settle on clear goals and carry them through, the constant initiatives leading nowhere, the frequent changes in mid-course." All Germany could do was bluff and hope to be paid for doing nothing: for remaining neutral in the Russo-Japanese War, for not building more dreadnoughts, for letting the French into Morocco, for not penetrating Persia. Of course, Germany could have launched an imperialist war in 1905 or 1911 under more favourable circumstances. It chose not to do so, and German might was such that prior to 1914 the other powers never considered a passage of arms with Germany.

Contradictions in German policy

Instead, Triple Entente diplomacy served to undermine Austria-Hungary. Everyone recognized that it was the "sick man of Europe" and that its demise would be inconvenient at very best and would almost certainly expose the ethnic mare's nest of southeastern Europe to civil war or Russian or German domination. Yet no one did anything about it. France could scarcely afford to—its security was too tightly bound to Russia's—but France's policy of wooing Italy out of the Triple Alliance was a grave setback, not for Germany but for Austria-Hungary. Russia brazenly pushed the Slavic nationalities forward, thinking to make gains but never realizing that tsarism was as dependent on Habsburg survival as Austria-Hungary had been on Ottoman survival. Only Britain had the capacity to maneuver, to restrain the likes of Serbia and Russia and take some of the Austro-Hungarian burden off Germany's shoulders. And indeed it had done so before—in 1815–22, 1878, and 1888. But now the British chose vaguely to encourage Russia in the Balkans, letting Austria-Hungary, as it were, pay the price for distracting Russia from the frontiers of India. So by 1914 Austria was encircled and Germany was left with the choice of watching her only ally collapse or risking a war against all Europe. Having chosen the risk, and lost, it is no surprise that the Germans (as well as the other powers) gave vent to all their prewar bitterness and pursued a thorough revision of world politics in their own favour.

World War I, 1914–18

World War I has aptly been called a war of illusions that exposed in sharp relief all the follies of the prewar generation. The war plans of the generals had misfired at once, and expectations that the intensity of modern firepower would serve the offense, or that the war must be brief, proved horribly false. Germany expected to achieve hegemony in Europe as a step toward world power, and instead world powers were called into play to prevent hegemony in Europe. Socialists thought war would bring general strikes and revolution, and instead the war inspired patriotic national unity. Monarchists hoped war would bolster the old regimes, and instead it cast down the remaining dynasties of eastern Europe. Liberals hoped that war would promote the spread of freedom, and instead it forced even democratic governments to impose censorship, martial law, and command economies subordinated to the dictates of centralized bureaucracy. Each nation in its own way sacrificed one by one those values it claimed to be fighting for in the belief that final victory would make good all the terrible cost. And with terrible irony World War I also ended in various plans for peace as illusory as the plans for war had been. As the historian William McNeill wrote, "the irrationality of rational, professionalized planning could not have been made more patently manifest."

A war of illusions

World War I can be divided, without undue violence to reality, into three periods: the initial battles, struggles for new allies, and mobilization on the home fronts, occupying the period from 1914 to 1916; the onset of ideologized warfare in the Russian revolutions and American entry in 1917; and the final four-way struggle of 1918 among German imperialism, Allied war-aims diplomacy, Wilsonian liberal internationalism, and Leninist Bolshevism.

MILITARY STALEMATE AND NEW BELLIGERENTS

From grand plans to the trenches. The first months of war resounded with the collision of the war plans pored over for decades by the general staffs of Europe. The original German plan for a two-front war, drafted by Helmuth von Moltke the elder, had called for taking the offensive against Russia and standing on the defensive in the rugged Rhineland. The plan showed military prudence and complemented the stabilizing diplomacy of Bismarck. But Alfred, Graf von Schlieffen, presided over the German military in the era of Kaiser William's *Weltpolitik* and adopted a more ambitious and risky course. His plan, conceived in 1891 and completed by 1905, envisioned a massive offensive in the west to knock out the compact French forces in six weeks, whereupon the army could shift eastward to confront the plodding Russians. But a quick decision could be achieved in France only by a vast enveloping action. The powerful right wing of the German army must descend from the north and pass through the neutral Low Countries. This would virtually ensure British intervention. But Schlieffen expected British aid to be too little and too late. In sum, the Schlieffen Plan represented a pristine militarism: the belief that all factors could be accounted for in advance, that execution could be flawless, that pure force could resolve all political problems including those thrown up by the plan itself. In the event, the Germans realized all of the political costs of the Schlieffen Plan and few of the military benefits.

The German plan of war

Like the Germans, the French had discarded a more sensible plan in favour of the one implemented. French intelligence had learned of the grand lines of the Schlieffen Plan and its inclusion of reserve troops in the initial assault. General Victor Michel therefore called in 1911 for a blocking action in Belgium in addition to an offensive into Alsace-Lorraine. But this required twice the active troops currently available. France would either have to give up the Belgian screen or the offensive. The new chief of staff, J.-J.-C. Joffre, refused to believe that Germany would deploy reserve corps in immediate combat and gave up the screen.

The traditional British way of war had been maritime: destroy the enemy's fleet, impose a blockade, and use land forces only to secure key points or aid continental allies at decisive moments. In Sir John Fisher's phrase, the army "should be regarded as a projectile fired by the navy." The prewar conversations with France, however, led the War Office to consider how Britain's army might help in case of war with Germany. General Henry Wilson insisted that even Britain's six divisions of professionals could tilt the balance between France and Germany and won his case for a British Expeditionary Force. Privately, he conceded that six divisions were "fifty too few" and hoped for a mass conscript army on continental lines.

By October 1914 all the plans had unraveled. After the German defeat in the Battle of the Marne, the Western Front stabilized into an uninterrupted line for 466 miles from Nieuwpoort on the Belgian coast south to Bapaume, then southeast past Soissons, Verdun, Nancy, and so to the Swiss frontier. Both sides dug in, elaborated their trench systems over time, and condemned themselves to four years of hellish stalemate on the Western Front.

Creation of the Western Front

The situation was little better on the other front. A necessary assumption of the Schlieffen Plan was the inadequacy of the Russian rail network to support a rapid offensive. By 1914, however, railroads through Poland were much improved, and the Russian general staff agreed to take the offensive in case of war to relieve the pressure on France. Similarly, the Germans had asked the Austrian commander, Conrad von Hötzendorf, to attack Russia and ease the threat to Germany. But Austria also had a two-front war and an army too small to fight it. Due to penury and its nationality problems, the monarchy fielded fewer battalions in 1914 than it had in the war of 1866. As the saying went, Austria was always *"en retard d'une armée, d'une année et d'une idée"* ("one army, one year, and one idea behind"). Austria's solution was to send one army south against Serbia and one to Galicia against the Russians, and to deploy a third as need required. The reserves, a third of Austria's already outnumbered forces, spent the opening battles shuttling back and forth on the rails. Austria failed to penetrate Serbian defenses while the Germans smashed the Russian attack into East Prussia. In the east, too, stalemate set in.

Defense in depth

By mid-1915 the Germans had overcome supply problems and were better prepared for trench warfare than the Allies. They also pioneered the concept of "defense in depth," making a second trench line the main barrier to assault. Allied generals responded with longer and denser artillery bombardments but thereby relinquished the element of surprise. Such tactics turned western battlefields into seas of wreckage with a "storm of steel" raging above, and condemned hundreds of thousands of men for the sake of a few thousand yards of "no man's land." Allied attacks of 1915 cost the British more than 300,000 casualties and the French 1,500,000. The only German initiative, the Second Battle of Ypres, introduced poison gas to the Western Front. But no commander could see a means of breaking the deadlock, and all confessed their strategy to be one of attrition.

The war at sea and abroad. The stalemate on land was matched by stalemate at sea when the British decided to impose a distant rather than close blockade of the German coast. This reduced the danger to the Grand Fleet and, it was hoped, might entice the German navy to venture out for a decisive battle. Admiral von Tirpitz was prepared to run such a risk, believing that the technical superiority of his High Seas Fleet would balance out Britain's numerical edge. Only by risking all on a major fleet action might Germany break the blockade, but the Kaiser and civilian leadership wished to preserve their fleet as a bargaining chip in eventual peace talks, while the British dared not provoke an engagement since a major defeat would be disastrous. Admiral John Jellicoe, it was said, was "the only man who could lose the war in an afternoon."

German overseas losses

In the wide world, the Allies cleared the seas of German commerce raiders and seized the German colonial empire. In the Pacific, New Zealanders took German Samoa and Australians German New Guinea. On Aug. 23, 1914, the Japanese empire honoured its alliance with Britain by declaring war on Germany. Tokyo had no intention of aiding its ally's cause in Europe but was pleased to occupy the Marshall and Caroline archipelagos and lay siege to Germany's Chinese port of Tsingtao, which surrendered in November. Germany's African colonies were, on the outbreak of war, immediately cut off from communications and supply from home, but military operations were needed to eliminate the German presence. By early 1916, Togoland (Togo) and Kamerun (Cameroon) fell to Anglo-French colonial forces and German South West Africa (South West Africa/Namibia) to the South Africans. Only in German East Africa was a native force under Lieutenant Colonel Paul von Lettow-Vorbeck, numbering initially only 12,000 men, able to survive for the entire war, tying down 10 times that number of Allied troops.

Efforts to break the stalemate. Thus all the armies and navies of Europe faced each other across fortified front lines. The prewar plans had succumbed to the technological surprise of 1914–15: that the withering firepower of machine guns, cartridge rifles, and rapid-fire artillery favoured the defense. Infantry in deep trenches, fronted with mines and barbed wire and backed by artillery, could not be dislodged by frontal attack. Accordingly, military and political leaders spent the war groping for means of breaking the stalemate in the trenches. First, neutrals might be enticed to enter the war, perhaps throwing enough weight into the balance to provide victory. Second, new weapons, tactics, and theatres might be tried to break the deadlock or achieve strategic goals elsewhere. Third, more and more men and matériel might be squeezed out of the home economy to tip the balance of forces or wear down the enemy in a war of economic attrition. The first of these means determined much of the diplomatic history of the war. The second stimulated technological developments such as poison gas, tanks, and submarines, and the peripheral campaigns of southern Europe and the Middle East. The third determined the evolution of war economies and the character of what came to be called total war.

Turkey enters the war

The first of the European neutrals to join the fray was the Ottoman Empire. Having lost the Balkans before 1914 and fearing partition of their Arab possessions by the Triple Entente, the Young Turks under Enver Paşa looked to Germany, whose military efficiency they admired. Enver led in negotiating a secret German-Ottoman treaty, signed Aug. 2, 1914. But the Grand Vizier and others in the Sultan's court held back, even after extracting a German loan—tantamount to a bribe—of £5,000,000. The war party then resorted to more extreme measures. The Ottoman fleet, reinforced by two German cruisers, entered the Black Sea in October, bombarded Odessa and the Crimean ports, and sank two Russian ships. The commander then falsified his account to make it appear that the enemy had provoked the action. The outraged Russians declared war on November 1. The Ottoman Empire's alliance with the Central Powers was a serious blow to the Entente, for it effectively isolated Russia from its Western allies and weakened their hand in the Balkan capitals. But the Turks concluded that a Triple Entente victory in the war would lead to the partition of their empire even if they remained neutral (Allied negotiations had already begun to this effect), whereas joining forces with Germany gave them at least a fighting chance to survive and perhaps even win some spoils from Russia. Enver also declared a jihad, or holy war, inciting Muslims to rise up against British and Russian rule in India, Persia, and Central Asia.

Turkish forces deployed along the coasts of the Dardanelles and on the Caucasus frontier with Russia, where severe fighting began in the rugged mountains. Enver, with German encouragement, took the strategic offensive when he ordered 10,000 troops from Syria to attack the Suez Canal in late January 1915. After crossing the Sinai the tired soldiers found Indian and Australasian divisions in training, as well as gunboats and other equipment they could not match. The Turks fell back to Palestine and never menaced the canal again.

The Dardanelles campaign

The vulnerability and value of the Dardanelles in turn attracted the British. When Russia requested a Western assault on Turkey to relieve the pressure in the Caucasus, War Secretary Lord Herbert Kitchener and First Lord of the Admiralty Winston Churchill promoted an attack on the Dardanelles. By capturing Constantinople, the British could link up with the Russians, knock Turkey out of the war, and perhaps entice the Balkan states to rally to the Allied cause. The British War Council created an amphibious force of British, Australians, and New Zealanders to capture the heights of the Gallipoli peninsula. On April 25 the "Anzac" (Australian and New Zealand) forces went ashore, but their assaults on the heights of Sari Bair were turned back through the charismatic leadership of the young Turkish officer Mustafa Kemal. A sweltering, bloody deadlock dragged on into the summer. Five more divisions and another amphibious landing, at Suvla Bay in August, failed to take the rugged heights in the face of human wave counterattacks by the Turks. Cabinet opinion gradually turned against the campaign, and the Allied force of 83,000 was evacuated—a dangerous operation conducted with great skill—in January 1916. The Turks had lost some 300,000 men, the Allies about 250,000 to battle and disease. Gallipoli was, in Clement Attlee's

words, "the one strategic idea of the war." Its failure, through bad leadership, planning, and luck, condemned the Allies to seek a decision in bloody battles of attrition on the Western Front.

The other peripheral front that enticed Allied strategists was Austria's border with Italy. Though a member of the Triple Alliance, the Rome government maintained on Aug. 3, 1914, that it was not bound to fight since Austria had not been attacked nor had it consulted with Italy as the treaty required. Prime Minister Antonio Salandra, a nationalist dedicated to the Irredentists' goal of recovery of Trentino and Trieste from Austria, announced that Italy would be informed by *sacro egoismo.* This, he explained, was a mystical rather than cynical concept, but it set off seven months of haggling over what the Allies would offer Italy to enter the war, and what the Central Powers would offer for neutrality. Some considerations were objective: Italy's 4,160 miles of coastline made defense against the Anglo-French fleet virtually impossible; any gains extorted from the Central Powers for neutrality would hardly be secure should those powers win the war; and neutrality was incompatible with Italy's tenuous claim to be a Great Power. What was more, all the Central Powers could offer was Trentino, and even that promise had to be forced from Vienna by German pressure.

The competition for Italy

After a clumsy intervention by the Russian foreign minister, Sazonov, in which he tried to secure Italy's help and still protect Serbian interests on the Dalmatian coast, negotiations moved to London. Berlin dispatched ex-chancellor Bülow and Roman Catholic statesman Matthias Erzberger to Rome to plead for the Central Powers. On April 26, the day after the first Gallipoli landing, the Treaty of London committed Italy to enter the war against Austria-Hungary within a month. In return the Allies promised Italy Trentino, part of South Tirol, Trieste, a third of Dalmatia (at the expense of Serbian ambitions), a mandate over Albania, a portion of German East Africa, all of Libya, a part of Asia Minor, and a 1,250,000,000-lira war chest from Britain. Still, a month of crisis followed in Rome as journalists like Gabriele D'Annunzio and Benito Mussolini stoked war fever and parliamentary power-broker Giovanni Giolitti (backed by Bülow) maneuvered for peace and *parecchio*—the "much" that might be obtained from Austria without lifting a rifle. After a Cabinet crisis Salandra returned to power to declare war on Austria-Hungary on May 23, 1915 (though Italy did not declare war on Germany until August 1916).

General Luigi Cadorna's war plan called for a strategic defense in the mountainous Trentino while half the Italian army concentrated for attack along the Isonzo River to the south. In June 1915 he launched the first of 11 battles of the Isonzo, wasting some 250,000 men against the rocky parapets and spirited Austrian defenders. The southern front became another deadlock, while Italy's weak finances and industry would only make her a continuing drain on Anglo-French resources.

First battle of the Isonzo

After Turkey and Italy, attention turned to the neutral Balkan states. The entry of the Balkan states on the side of the Central Powers would doom Serbia and open direct communications between Germany and Turkey. Balkan participation on the Allied side would isolate Turkey and complete the encirclement of Austria-Hungary. The Central Powers had the upper hand in Bulgaria, still smarting from its defeat in the Second Balkan War and allied with Turkey as of Aug. 2, 1914. The Allies had little to offer Bulgaria except bribes, especially after their failure at Gallipoli. German offers proved irresistible: Macedonia (from Serbia) and parts of the Dobruja and Thrace should Romania and Greece intervene. Bulgaria joined the Central Powers on Sept. 6, 1915. In Romania the Allies had the upper hand despite a treaty, renewed in 1913, binding Bucharest and its Hohenzollern dynasty to the Triple Alliance. Romania's main ambition was to annex Transylvania, a Habsburg province populated largely by Romanians, but Prime Minister Ionel Brătianu determined to stay neutral and observe the fortunes of war.

In 1915 those fortunes appeared to favour the Central Powers on the Turkish, Italian, Serbian, and Russian fronts. The Russian front collapsed in the face of a German offensive in May, allowing the Central Powers to reoccupy Galicia, Lithuania, and Courland in the north. In July the Germans resumed the drive and threatened to pincer the entire Russian army in Poland. Warsaw fell on August 5 and Brest-Litovsk on the 26th, whereupon the German armies outran their supplies and halted the drive on a line stretching from Riga on the Baltic to Czernowitz on the Romanian border. Russian losses were apocalyptic: more than a million men captured and at least as many killed and wounded in 1915. Technical inferiority, shortage of munitions, and poor tactics led to terrible wastage of men in the attack and lack of mobility on the defense. The inadequacy of the Russian state and economy in modern war now stood revealed. Desertions increased and morale plummeted. On September 5, Tsar Nicholas himself took over supreme command, a chivalrous move but one that would identify the crown with future disasters.

German advance against Russia

In 1916 German strategists again turned west with the expressed intention of bleeding France white and breaking her army's spirit. The object of attack was to be the fortress of Verdun, and the plan called for the substitution of ordnance for manpower as much as possible, thereby using Germany's industrial might to kill Frenchmen in the most efficient way. The assault began on February 21, following an avalanche of shells and poison gas, and continued without interruption for five months. France's civilian and military leadership turned Verdun into a national symbol of resistance, symbolized by General Philippe Pétain's famous order of the day: "*Ils ne passeront pas!*" Verdun was the most intensive battle in history and cost France and Germany more than 300,000 men each.

In December 1915 an Allied conference at Chantilly had decided to coordinate simultaneous attacks on all fronts. Given Verdun, responsibility for the Western assault fell to the British. After elaborate preparation and a week of bombardment the cream of "Kitchener's New Army" went over the top on July 1, 1916, and strode in formation toward the German lines. By mid-November the Somme offensive had gained about six and a half miles across a 30-mile front at the cost of 420,000 Britons, 194,000 Frenchmen, and 440,000 Germans.

The First Battle of the Somme

On the Eastern Front in 1916 the Russian command dutifully took up the offensive to relieve the pressure on Verdun and in coordination with the push on the Somme. But failures in leadership and supply, poor intelligence and tactics again thwarted the courage of Russia's peasant-soldiers, 100,000 of whom were lost in a March attack that achieved nothing. The last gasp of the tsarist army followed in June. Russian attacks at Lutsk, Buchach, and Czernowitz beginning June 4 achieved total surprise, captured 200,000 men, and overran Bukovina by the end of the month. This apparent revival of Russia's fortunes prompted the Romanians, finally, to declare war on Austria-Hungary on Aug. 27, 1916. Half the Romanian army—12 divisions—joined the offensive and advanced into Transylvania, expecting to deal the final blow to staggering Austria-Hungary. Instead, Germany, Turkey, and Bulgaria promptly declared war on Romania. The Romanians held out for a month against a German-Austrian-Bulgarian attack at the Vulcan and Szurduk (Surduc) passes, but the Central Powers broke through and captured Bucharest on December 6. The Romanian gambit ended in disaster as the Germans acquired their oil and wheat and the Russians inherited an additional 300 miles of frontline. Meanwhile, the Russian offensive degenerated into frontal assaults and closed in August. Russia had lost 500,000 men—the last trained reserves of the tsarist army.

Fall of Bucharest

By the end of 1916 what may be called the traditional phase of the war had run its course. Despite ever greater expenditures of men and matériel and the accession of neutral powers to one side or the other, victory remained elusive. Henceforth the coalitions would rely all the more on breaking the internal cohesion of the enemy or on calling forth global forces to tip the balance. The resort to revolution, especially in Russia, and extra-European powers, especially the United States, would have profound consequences for Europe's future in the 20th century, while internal mobilization for total war had already gone far to reshape European societies.

WAR MOBILIZATION AT HOME AND ABROAD

The invention of total war. When the first campaigns failed and the belligerents steeled themselves to fight a long war of attrition, World War I became total—that is, a war fought without limitations, between entire societies and not just between armies, with total victory the only acceptable outcome. It became such a war because, for the first time, the industrial and bureaucratic resources existed to mobilize an entire nation's strength, because the stalemate required total mobilization, and because the tremendous cost and suffering of such a war seemed to preclude settling for a negotiated truce. Only victory might redeem the terrible sacrifices already made by both sides; and if final victory were the only acceptable end, then any means could be justified in pursuit of it.

The first violent battles of 1914 nearly expended prewar munitions reserves. By mid-war the artillerymen of the Western Front might fire more shells in a single day than were expended in the entire Franco-German War. Clearly the home front—the war economy—would be the most decisive of all. And yet the governments, expecting a short war, were unprepared for economic mobilization and had to adjust to emergencies and shortages as they arose. In Germany the process began in the first days of war when private manufacturers, especially Walther Rathenau, suggested a state bureau to distribute raw materials to industry. Over the years it became a model for new agencies, boards, and commissions controlling production, labour, rationing, travel, wages and prices. By late 1917, Germany came to dominate the economies of Austria-Hungary and the occupied regions by the same means. In all the belligerent nations, to a greater or lesser degree, civil and economic liberties, the free market, even national sovereignty, gave way to a kind of military socialism in the crucible of war. All the belligerents met their labour needs through employment of old men, children, and women (a fact that ensured the success of the suffragette movement in Europe after the war). The Allies also engaged in economic war through agreements with neutral countries on the Continent not to re-export goods to Germany and through preemptive purchase of everything from Chilean nitrates to Romanian wheat.

Germany's wartime command economy

An economic problem that could be postponed was the financial one. The belligerents immediately ended controvertibility of their currencies according to the gold standard and liquidated their holdings overseas. By late 1915 the British and French also began to float sizable loans on the American market, even as they themselves underwrote the war efforts of weaker economies like the Italian and Russian. British, Germans, and Americans covered a fraction of the war's expense through income and other taxes, but World War I was financed primarily through war bonds and secondarily through loans from abroad. This pattern would exacerbate the diplomatic and domestic political climates after the war, when the bills for the four years' wastage came due.

Financing the war

The weapon of morale. The mass conscripted army and labour force, the employment of women and children, and the mobilization of science, industry, and agriculture meant that virtually every citizen contributed to the war effort. Hence all governments tried to stoke morale on the home front, subvert that of the enemy, and sway the opinions of neutrals. A variety of techniques for manipulating information were used, including particularly censorship and vilification of the enemy. German propaganda depicted Russians as semi-Asiatic barbarians and the French as mere cannon fodder for the bloated, envious British Empire lusting to destroy Germany's power, prosperity, and *Kultur*. The French Maison de la Presse and British Ministry of Information took German war guilt for granted and made great play of the atrocities committed by the "Hun" in Belgium and on the high seas, where defenseless passenger ships were treacherously torpedoed. War hatred whipped up by such propaganda made it all the more difficult to justify negotiating a truce.

The Allies proved more adept than the Germans at psychological warfare. Propaganda was distributed across German lines by shells, planes, rockets, balloons, and radio. Such activities were given into the hands of an Inter-Allied Propaganda Commission in 1918. The Allies also, especially after 1917, identified themselves with such universal principles as democracy and national self-determination, while the German war effort had only a narrow national appeal. The most important target of propaganda was the United States. In the first weeks of war the British cut the German transatlantic cables and subsequently controlled the flow of news to America. German attempts to influence U.S. opinion were invariably clumsy, while the British, aided by the common language, reminded Americans of their common values for which German militarism had no respect. In political warfare, German attempts to arouse the Muslim world and incite India to rebellion were stillborn, while their exploitation of the situation in Ireland, culminating in the Easter Rising of 1916, backfired. The aristocratic and continental German officials seemed out of their element when either trying to appeal to the masses or looking beyond Europe. But their one success was nothing less than the Russian Revolution of 1917 (see below).

Allied propaganda and America

WAR AIMS AND PEACE FEELERS

War aims of the belligerents. For what were the nations of Europe making such total and mortal commitments? In public each government insisted it was fighting first in self-defense, then for victory and some hallowed national goal like naval security for Britain, Alsace-Lorraine for France, or Constantinople for Russia. But in private, now that peacetime constraints were torn off, each indulged greater ambitions. German war aims took shape at once in the September Program of Bethmann. While debate exists over how much this document reflected Bethmann's real views, it did come to represent the prevailing view of the military, which in turn came to speak increasingly for Germany as a whole. The dream of world power seemed within reach through the acquisition of Belgian and French colonies that, when joined to Germany's and perhaps Portugal's, would constitute a *Mittelafrika* of immense proportions. In Europe the Germans determined to assure that France and Russia would pose no threat in the future and to create an economic base suitable for a world power. This notion of a single economic bloc from Berlin to Baghdad, including Belgium, the Longwy-Briey mines of France, Poland, Courland, the Ukraine, and the Balkans, was popularized as *Mitteleuropa* in a 1915 best-seller by Friedrich Naumann. How committed Germany's civilian leadership was to this hegemonic plan is disputed: Bethmann favoured abandoning much of it in hopes of a negotiated peace. But a war-aims majority held the balance in the Reichstag until 1917 and in the military until the bitter end.

The idea of *Mitteleuropa*

On Sept. 5, 1914, the entente powers solemnly and severally renounced any separate peace, but throughout the war they felt constrained to bolster each other's will to fight with promises of spoils. Hence the purchase of Italy's belligerency and the shocking willingness of Britain and France to consign Constantinople to Russia in March 1915. In general, Allied ambitions added up to the partition of the German and Ottoman empires and security against Germany in Europe and on the seas. Partition of Austria-Hungary was not an initial Allied aim. In the spring of 1915 France and Russia exchanged letters promising that both could do as they wished on their borders with Germany, implying a free hand for Russia in Galicia and East Prussia and the same for France on the Rhine. French industry contemplated an advance into the Saar and Rhine regions to end France's inferiority in coal production (which would only be exacerbated by the return of Alsace-Lorraine with its rich iron deposits). For the French army and foreign ministry, however, the main motive for separating the Rhineland from Germany was security: what Poincaré called "breaking Prussian militarism" and Aristide Briand "guarantees of lasting peace." In 1917 Paris and St. Petersburg were close to a formal treaty on the German boundaries when the Russian Revolution intervened.

The Allies specified their colonial claims in an agreement of April 1916, by which Britain was granted influence in Mesopotamia and Syria; France the same in Lebanon,

Allied territorial agreements

Cilicia, and southern Kurdistan; and Russia in Armenia and northern Kurdistan; Palestine was placed under joint Anglo-French administration. The Sykes–Picot Agreement in May also divided much of the Ottoman Empire into British and French spheres. The Agreement of Saint-Jean-de-Maurienne of April 1917 promised Italy concessions on the Anatolian coast; one Allied motive in this was to persuade Rome to scale down its claims on Austria-Hungary in hopes of a separate peace with Vienna (see below). Finally, the French began in 1916 to formulate a second set of war aims directed, not at Germany, but at their own allies. British currency supports, loans, coal shipments at fixed prices, and other benefits helped sustain the French war effort, and the minister of commerce, Etienne Clémentel, lobbied for an extension of these supports beyond an armistice lest France win the military struggle only to lose the postwar economic struggle. The British agreed at the Allied Economic Conference of 1916, and the following year the French placed even greater hopes of economic solidarity in the newly associated power, the United States.

Attitude of the United States. Since 1783 the United States had acquired a number of foreign policy traditions. George Washington, in his Farewell Address, admonished his young and vulnerable country to avoid alliances that would drag it into disputes in which it had no interest. Thus was born a powerful isolationalist and exclusivist tradition. The Monroe Doctrine declared the Western Hemisphere off-limits to European adventurism, giving birth to a regionalist and paternalist tradition vis-à-vis Latin America. After the Civil War, belief in America's Manifest Destiny directed national attention to the West Coast and beyond. Then the war against Spain in 1898 yielded colonial possessions in the Caribbean and Pacific and inspired the building of a two-ocean navy and of a Panama Canal to serve it. By 1914, when the canal opened, the United States was already the greatest industrial power in the world, yet its tradition of exclusivity and its tiny standing army gave the Europeans excuse to ignore America's potential might.

Wilsonian neutrality

In August 1914 President Woodrow Wilson implored the American people to be "neutral in thought as well as deed" with respect to the European war. In so doing he was not only honouring tradition but also applying his own religious principles to foreign policy. His agenda upon entering the White House in 1913 had been domestic reform, and he had written that it would be an irony of fate should foreign policy come to dominate in his administration. Yet when fate so decreed, Wilson preferred to trust his own motives and methods rather than the advice of his secretaries of state or his other advisers. Wilson deplored the war and earnestly wished to bring about a just and lasting peace through U.S. mediation, for what greater mission could Providence assign to that "city on a hill," the United States of America?

American power began to figure in the balance of war almost from the start. Trading was suspended on the New York Stock Exchange when war broke out, but when it resumed in November 1914, Europeans sold most of the $4,000,000,000 worth of securities they held before the war. U.S. loans to belligerents were at first declared "inconsistent with the true spirit of neutrality," but the large Anglo-French orders for U.S. munitions, raw materials, and food created an economic boom, and by 1915 the Allies needed credit to continue their purchases. An initial £200,000,000 loan in September 1915 led eventually to billions being floated on the U.S. market and a complete reversal of the financial relationship between the Old World and the New. By 1917 the United States was no longer a debtor nation but the world's greatest creditor. U.S. firms also inherited many overseas markets, especially in Latin America, which the British and Germans could no longer serve.

To Americans neutrality seemed both moral and lucrative—the United States, said Wilson, was "too proud to fight." But the failure of his peace initiatives, the German assaults on neutrals' rights at sea, and the cumulative effect of Allied propaganda and German provocations conjoined to end U.S. neutrality by 1917. On Feb. 4, 1915, Germany declared the waters around the British Isles a war zone in which Allied ships would be sunk, without warning if necessary. While this procedure dispensed with traditional civilities like boarding, search and seizure, and care of civilians, effective submarine warfare required it. Underwater craft relied on stealth and surprise and exposed themselves to easy destruction once they made their presence known. Thus, even though the British blockade interfered with neutral shipping more than the German blockade, the latter appeared far more beastly. The sinking of the Cunard liner *Lusitania* on May 7, 1915, which killed over a thousand passengers, including 128 U.S. citizens, outraged U.S. public opinion despite the rightful German claim that she was carrying munitions (173 tons worth). Two more passenger ships, the *Arabic* and *Hesperia,* went down in August and September, respectively, whereupon American diplomatic protests caused civil officials in Berlin to overrule the military command and call off unrestricted submarine warfare, although the issue did not remain settled.

Sinking of the *Lusitania*

Wilson's own peace initiatives, including an offer of mediation by Secretary of State William Jennings Bryan in 1914 and a trip to Europe by Wilson's personal aide and adviser, Colonel Edward M. House, in 1915, were unsuccessful. Early in 1916 House returned to Europe and on February 22 in London agreed to a formula whereby the United States would summon a peace conference and—if Germany refused to attend or proved unreasonable—"would leave the conference as a belligerent on the side of the Allies." Wilson later drew back from the guarantee and added the word "probably" after "would." But the British themselves shied from promoting such a conference, while the other belligerents also ducked the suggestion lest they compromise the determination of their people or incur the distrust of allies.

By the end of 1916 Germany had 102 U-boats ready for service, many of the latest type, and the chief of the naval staff assured the Kaiser that unrestricted submarine warfare would sink 600,000 tons of Allied shipping per month and force Britain to make peace within five months. Bethmann fought to delay escalation of the submarine war in hopes of another Wilsonian peace move. But the President held off new initiatives during his reelection campaign. When he had still not acted by December 1916, Bethmann was forced to make a deal with his own military, which consented to tolerate a German peace offer in return for Bethmann's endorsement of unrestricted submarine warfare if the offer failed. But the army helped ensure that the German note (released December 12) would fail by insisting on implicit retention by Germany of Belgium and other battlefield conquests. Wilson followed on the 18th with an invitation to the two camps to define their war aims as a prelude to negotiation. The Allies demanded evacuation of occupied lands and guarantees against Germany in the future. The Germans stuck to their December note, and the military command decided to resume unrestricted submarine warfare on February 1.

German debate over submarine warfare

The United States broke diplomatic relations with Germany on February 3 and commenced the arming of merchant ships on March 9. Meanwhile, German Foreign Secretary Arthur Zimmermann, anticipating war with the United States over the U-boat issue, cabled an offer of alliance to Mexico on January 16, promising Mexico its own "lost provinces" of Texas, Arizona, and New Mexico in case of war with the United States. British intelligence intercepted the Zimmermann telegram and leaked it to Washington, further inflaming American opinion. When U-boats proceeded in mid-March to sink the *Algonquin, City of Memphis, Vigilancia,* and *Illinois* (the latter two without warning), Wilson went before Congress and in a lofty and moving address reviewed the reasons why America was forced to take up the sword—why, "God helping her, she can do no other." On April 6, 1917, Congress declared war on Germany, and the United States became an Associated (not an Allied) Power. Henceforth World War I hinged on whether the U-boats could force Britain to her knees and the German armies overwhelm the sagging Western Front before the men and matériel of the aroused Yankees could arrive in France.

THE CRISES OF 1917

War-weariness and diplomacy. For every belligerent, 1917 was a year of crisis at home and at the front, a year of wild swings and near disasters, and by the time it was over the very nature of the war had changed dramatically.

A French offensive in the spring soon ground to a standstill, sparking a wave of mutinies and indiscipline in the trenches that left the French army virtually useless as an offensive force. The British offensive of July–November, called variously Passchendaele or the Third Battle of Ypres, was a tactical disaster that ended in a viscous porridge of mud. That offensive action could be ordered under such conditions is a measure of how far Western Front generals had been seduced into a gothic unreality. Allied and German casualties "in Flanders Fields, where poppies grow" numbered between 500,000 and 800,000. The British Army, too, now neared the end of its offensive capacities.

The tragedy of Flanders Fields

For two years the Italian front had been left unchanged by the first nine battles of the Isonzo. But the underfinanced and underindustrialized Italian war effort gradually eroded. The Tenth Battle of the Isonzo (May–June 1917) cost Italy dearly, while the Eleventh (August–September) registered a "success" amounting to some five miles of advance at a cost of over 300,000 casualties, pushing the total for the war to more than a million. With peace propaganda, strikes, and Communist agitation spreading through Italy, and the Austrians in need of stiffening, the German high command reinforced the Austrians at Caporetto. Within days the Italian commander had to order a general retreat. The Germans broke the line of the Tagliamento as well, and not until the Italians regrouped at the Piave on November 7 did the front stabilize. Caporetto cost Italy 340,000 dead and wounded, 300,000 prisoners, and another 350,000 deserters: an incredible 1,000,000 in all, suggesting that the Italian army, like the French, was on strike against its own leadership.

Among the Central Powers also, 1917 intensified the yearning for peace. Polish, Czech, and Yugoslav leaders had formed committees in exile to agitate for the autonomy or independence of their peoples, while war-weariness among those at home grew with the food shortages, bad news from the front, and desertions among the troops. When Emperor Francis Joseph died in November 1916 after 68 years on the throne, there was a sense that the empire must die with him. Already Austro-Hungarian officials had begun to look for a way out of the war—which meant a way out of the German alliance. The new Habsburg foreign minister, the Polish Ottokar, Graf Czernin, raised the issue of war aims and peace at his first ministerial meeting with the new emperor, Charles. A negotiated peace could only be one without victors or vanquished, conquests or indemnities—so said Czernin 10 days before Wilson's own "Peace Without Victory" speech. The only means of achieving such a peace, however, was for Austria-Hungary's ally Germany to restore Belgium and, perhaps, Alsace-Lorraine.

The first Austrian demarches, made through Scandinavia, came to nothing, so Charles, Czernin, and the Empress Zita tried again in late January 1917 through the intermediary of her brother, Prince Sixtus of Bourbon-Parma, on leave from service in the Belgian army. In March, Charles drafted a letter in which he asked Sixtus to convey to the President of France his "lively sympathies" and support for the evacuation of Belgium and the lost provinces. The cautious French premier, Alexandre Ribot, shared the news in April with Lloyd George, who said simply, "That means peace." But Baron Sonnino, at the Conference of Saint Jean-de-Maurienne, refused to consider peace with Austria-Hungary (the only enemy Italy was interested in fighting) and warned Lloyd George against attempts to split their alliance. Charles second letter, in May, which inexplicably told the French and British of an "Italian peace offer" that was never made, only put the Allies on their guard.

Simultaneously the parliamentary forces of Germany rose in protest against the war, the erosion of civilian authority, and the war-aims stubbornness of the military command. A moderate annexationist deputy, Matthias Erzberger, met with Czernin and Emperor Charles in April 1917 and learned that Austria-Hungary's military strength was near its end. In May a Reichstag committee demanded that the army be placed under civilian control. The Kaiser and the military high command replied with scorn. In July Bethmann was forced to resign and the army assumed de facto control of Germany. When the Kaiser appointed a nonentity, Georg Michaelis, as chancellor, the Reichstag passed a peace resolution on July 19 by a vote of 212–126. But the resolution could have no bearing on the ruling circles, to whom compromise with the foreign enemy meant surrender to the domestic forces of reform.

Military supremacy in Germany

In mid-August, Pope Benedict XV tried to preserve momentum toward a truce by calling on all parties to evacuate occupied regions. But the German government again refused to surrender Belgium, while the American reply to the Vatican seemed to insist on the democratization of Germany. Emperor Charles and Czernin were likewise unable to make headway, for the Allies were not at this point seeking a general peace but only a separate peace with Austria-Hungary that would leave Germany stranded. That Vienna could not in honour do, nor Berlin permit. The United States declared war on Austria-Hungary on Dec. 7, 1917, and when the French government leaked news the following spring of the Austrian peace correspondence, Charles and Czernin were forced to humble themselves before the Kaiser and German high command at Spa. Austria-Hungary had become a virtual satellite of the German military empire.

The Ottoman Empire in 1917 began to give way before the relatively mild but incessant pressure on fronts the other powers considered sideshows. Baghdad fell to British forces in March. Sir Edmund Allenby, having promised Lloyd George that he would deliver Jerusalem to the British people "as a Christmas present," made good his promise on December 9. The political future of Palestine, however, was a source of confusion. In the war-aims treaties, the British had divided the Middle East into colonial spheres of influence. In their dealings with the Arabs the British spoke of independence for the region. Then on Nov. 2, 1917, the Balfour Declaration promised "the establishment in Palestine of a national home for the Jewish people," albeit without prejudice to "the civil and religious rights of existing non-Jewish communities." Foreign Secretary Arthur Balfour was persuaded that this action was in British interest by the energetic appeals of Chaim Weizmann, but in the long run it would cause no end of difficulty for British diplomacy.

The Balfour Declaration

The one flank on which Turkey had not been besieged was the Balkan, where an Allied force remained in place at Salonika pending resolution of the Greek political struggle. The Allies continued to back Prime Minister Venizélos, who, because King Constantine still favoured the Central Powers, had fled Athens in September 1916 and set up a provisional government under Allied protection at Salonika. Finally, the Anglo-French deposed Constantine in June 1917 and installed Venizélos in Athens, whereupon Greece declared war on the Central Powers. By the end of 1917, therefore, Turkey, like Austria, was exhausted, beleaguered on four fronts, and wholly dependent on German support.

The Russian Revolution. But while Britain, France, Italy, Germany, Austria-Hungary, and Turkey all survived their crises of 1917 and found the will and stamina for one last year of war, Russia succumbed to its crisis. In three years of war Russia had mobilized roughly 10 percent of its entire population and lost over half of that number in battle. The home economy was stretched to the limit, and even the arms and food it could produce were subject to vagaries of transport and corruption in the supply services. Inflation and food shortages panicked the towns, shortages of fuel isolated the countryside. Suddenly, on March 12, 1917, the parliament and Petrograd soviet (workers' and soldiers' council) joined forces to form a Provisional Government. Three days later the Tsar abdicated.

The Provisional Government

Two leading ministers in the new regime, Aleksandr Kerensky and Pavel Milyukov, hoped to streamline the state and invigorate the war effort. Political liberals, they valued Russia's ties to Britain and France and even looked

forward to capturing Constantinople as a means of legitimating the new regime. Kerensky assured the Allies on March 17 that Russia would fight "unswervingly and indefatigably" until victory. The local soviets and leftist parties, however, forced a declaration in April by which "free Russia" renounced domination over other nations and their territories. When Prince Gyorgy Lvov, the prime minister, promised to accept the revolutionary formula of "no annexations, no indemnities" on May 15, Milyukov stepped down as foreign minister. President Wilson was especially moved by the spectacle of Russia embracing democracy, and all the Allies could now truly depict their cause as moral and ideological: "to make the world safe for democracy," as Wilson said, in opposition to militarism and imperialism. But Russia's ability to fight steadily and rapidly deteriorated. The Petrograd soviet called for abolition of the officer corps, and the Provisional Government abolished courts-martial and issued a Declaration of Soldiers' Rights.

The Provisional Government's decision to continue the war was a grave disappointment to the Germans. Since 1914 they had dabbled in revolutionary intrigues in hopes of shattering Russia from within. The campaign took two forms: collaboration with nationalist agitators among the Finns, Baltic peoples, Poles, Ukrainians, and Georgians; and support for Russian social revolutionaries. Lenin, leader of the most virulent wing of Russian Marxists, the Bolsheviks, was living in Kraków when the war broke out and was promptly arrested. But an Austrian Social Democrat, Victor Adler, persuaded the Austrian minister of the interior that Lenin was an ally in the fight against Russia, whereupon he was released into Switzerland. Another Russian émigré and Socialist, Alexander Helphand, impressed the German ambassador in Constantinople with his revolutionary connections and was soon briefing the German foreign ministry in Berlin. In March 1915 the Germans set aside the first 2,000,000 of what would eventually total 41,000,000 marks spent on secret subversion in Russia.

After the first Eastern Front victories in 1915, Berlin had hoped to entice Russia into a separate peace, and efforts to that end continued up to March 1917. But behind the scenes, Helphand's organization, supported by the German foreign office, worked to spread revolutionary and pacifist ideas inside Russia. After Kerensky's declaration that Russia would stay in the war, the German command determined to facilitate Lenin's return to Russia. On April 9, 1917, he and his comrades were placed aboard a special security train in Zurich for the trip across Germany, continued by boat to Sweden and thence by rail to Petrograd.

Lenin's return

Bolshevik propaganda penetrated the army, which even the Russian high command confessed was "a huge, weary, shabby, and ill-fed mob of angry men." In an attempt to restore it to fighting trim, General Lavr Kornilov urged on Kerensky a number of reforms (August 16), but behind Kornilov were conspirators hoping for military dictatorship. Kerensky grasped the danger to himself, forbade troop movements to the capital lest they support a coup, and then had Kornilov arrested. This division between the centre and right gravely weakened the Provisional Government and strengthened the Bolsheviks, who took the lead in denouncing this "counterrevolutionary plot." The Provisional Government, bereft of authority and will, hoped to hold on until elections for a Constituent Assembly in December. Lenin, knowing he stood to lose by the fact and the result of free elections, struck in November, and the Provisional Government collapsed in the face of the Bolshevik coup d'état.

One of Lenin's first acts as revolutionary dictator of Russia was to attempt to transform the European war of nations into a war of classes. His ringing speech of November 8 appealed to workers and soldiers everywhere to force an immediate armistice, end secret diplomacy, and negotiate a peace of "no annexations, no indemnities." Lenin, Trotsky, and Karl Radek promptly organized to spread revolution abroad. The expected uprisings occurred nowhere, but peace was mandatory for Russia if the Bolshevik regime were to survive. On December 15, therefore, Lenin's regime signed an armistice with the Central Powers.

LAST BATTLES AND ARMISTICE

Russia's withdrawal from the war. The events of 1917 meant that World War I was no longer a two-sided contest. Rather, four visions of the future now competed for the allegiance of governments and peoples. Germany fought on in hope of victory and domination of the Continent. The Allies fought on to frustrate Germany and realize their own ambitious war aims. Wilson's America fought as an "associated power" for a liberal internationalist agenda opposed to German and Allied imperialism alike. Finally, Lenin's Russia raised a second challenge to the old diplomacy in the name of Socialist internationalism. German, Allied, Wilsonian, and Bolshevik images of the peace differed so radically that the war was now as much ideological as it was military.

Lloyd George and Wilson replied to Lenin's peace initiatives with speeches of their own to reassure their peoples, contrast their liberal goals with those of the Germans, and perhaps persuade Russia to remain in the field. Lloyd George insisted before the Trades Union Congress (Jan. 5, 1918) that "we are not fighting a war of aggression against the German people," and stressed autonomous development for all peoples, including those of Austria-Hungary. Wilson's Fourteen Points speech (Jan. 8, 1918) called for (1) open covenants, openly arrived at; (2) freedom of the seas; (3) lowering of economic barriers; (4) reduction of armaments; (5) colonial arrangements respecting the will of the peoples involved; (6) national self-determination for the peoples of Russia; (7) restoration of Belgium; (8) return of all invaded territory plus Alsace-Lorraine to France; (9) Italian recovery of her *irrendente;* (10) autonomy for the nationalities of Austria-Hungary; (11) restoration of the Balkan states and access to the sea for Serbia; (12) autonomy for the peoples of the Ottoman Empire and free navigation through the Dardanelles; (13) an independent Poland with access to the sea; and (14) a "general association of nations" offering "mutual guarantees of political independence and territorial integrity." In his Four Principles (February 11) and Five Particulars (September 27) speeches Wilson elaborated his views on national self-determination, a truly revolutionary idea with global, but unpredictable, implications.

Wilson's Fourteen Points

Allied assurances failed to dissuade the Bolsheviks from exiting the alliance. Lenin took power on the slogan "Peace, Bread, and Land," and he needed to be free of the war in order to consolidate Bolshevik power. A peace conference convened at Brest-Litovsk on Dec. 22, 1917, but proceeded slowly while the two sides—one imperialist, the other incipiently totalitarian—bickered about the definition of "national self-determination." On Jan. 7, 1918, Leon Trotsky asked for adjournment, still hoping for revolutionary outbreaks abroad. In fact, a mutiny in the Austrian fleet and a general strike movement in Berlin did occur but were easily suppressed. The Bolshevik leadership now faced three bad choices: to defy the Germans and risk conquest and overthrow; to relent and sign over half of European Russia to German control; or to pursue what Trotsky called "neither war nor peace" while awaiting the revolution in Germany. He also wished to avoid any sign of collusion with the German military, lest the Bolsheviks appear to be collaborationists. In the meantime the Germans and Austrians concluded the *Brotfrieden* ("bread peace") with representatives of the wheat-rich Ukraine. But when Bolshevik forces began to penetrate the Ukraine—and the German high command tired of Trotsky's rhetoric—the Germans broke off talks and ordered the army to resume its advance. The French ambassador immediately offered the Bolsheviks all aid if they would fight the Germans, but Lenin ordered an immediate capitulation. Germany now presented even harsher peace terms. On March 3 the Bolsheviks signed. The Romanians then made peace on the 5th, and newly independent Finland signed a treaty with Germany on the 7th.

Trotsky's "neither war nor peace"

In the Treaty of Brest-Litovsk the Bolshevik regime turned over to Germany 34 percent of Russia's population, 32 percent of Russia's farmland, 54 percent of Russia's industrial plant, 89 percent of Russia's coal mines, and virtually all of its cotton and oil. These economic gains in the east, plus the release of troops who could

Treaty of Brest-Litovsk

now be shifted to the Western Front, revived German hopes that victory was achievable before the Americans arrived in force.

Negative views of the Bolshevik Revolution predominated from the start in Western capitals, although some people on the left in London, Paris, and Washington sympathized with it or thought it would bring much needed "efficiency" to Russia. The French and British had talked of supporting this or that Russian faction with arms or cash and had agreed on a tentative division of southern Russia into areas of responsibility. The German advance of February then caused the Allied missions to flee Petrograd and reassemble in remote Vologda, where they waited to see what direction the Bolsheviks would take. The Treaty of Brest-Litovsk answered the question. It was an unparalleled disaster for the beleaguered Allies, who now had to consider intervention in Russia. First, if they could link up with nationalist Russians and reopen the Eastern Front, they might save their exhausted armies in France from facing the full might of the Central Powers. Second, it would be most helpful if they could save Allied war matériel that had stacked up in Russian ports (some 162,495 tons of supplies in Arkhangelsk alone) from seizure by the Germans or Bolsheviks and distribute it to Russians still willing to fight Germans.

Allied intervention in Russia

When the German onslaught on the Western Front opened in March, the French and British became desperate for a diversion in the East. In March 1918 an Anglo-French expedition docked at Murmansk, followed in June by an American cruiser and 150 marines. An Anglo-French force occupied Arkhangelsk in August, and 4,500 U.S. soldiers under British command joined them in September. These tiny contingents, totaling about 28,000 men, were never meant to overthrow the Bolshevik regime, although the British hoped they might serve as magnets for White Russian forces opposing the Bolsheviks.

The Japanese, seeking an imperial foothold on the Asian mainland, used Brest-Litovsk as pretext to occupy Vladivostok in April. Wilson then committed U.S. troops to Siberia in order to keep an eye on the Japanese and to make contact with 30,000 Czechoslovak legionnaires, mostly former prisoners of war from the Habsburg armies seeking to escape Russia to fight for an independent Czech state. The Czechoslovak Legion, released and armed by the Kerensky government, at first declared neutrality toward Russian politics, but when the Bolsheviks tried to disarm them, skirmishes ensued, and the legion became strung out along the 6,000-mile-long Trans-Siberian Railway. The Allied interventions also became entangled in the erupting Russian Civil War. Bolsheviks controlled Petrograd, Moscow, and the core regions of Russia, while White governments were established by Admiral Aleksandr Kolchak in Omsk and General Anton Denikin in Odessa.

The eastern minorities. The saga of the Czechoslovak Legion was symbolic of the growing vigour of the national movements inside the Habsburg Empire. Early in the war the subject peoples had remained loyal to beloved old Francis Joseph. But martial law, which fell especially hard on minorities, war weariness, hunger, and the example of the Russian Revolution converted moderates among the Czechs, Galician Poles, and South Slavs to the cause of independence. The Czechs and Slovaks were brilliantly served by Tomáš Masaryk and Edvard Beneš, who lobbied for Allied recognition of a Czech national council. The Polish movement, led by Józef Piłsudski, sought to establish similar national institutions and cooperated with the Central Powers after their Two Emperors' Manifesto (Nov. 5, 1916) promised autonomy to the Poles. The Polish National Committee in France, and famed pianist Ignacy Paderewski in the United States, also pleaded the Polish cause. Yugoslav (or South Slav) agitation was complicated by rivalries between the Serbs (Orthodox, Cyrillic alphabet, and politically stronger) and the Croats and Slovenes (Roman Catholic, Latin alphabet, politically disinherited), as well as Serbia's and Italy's conflicting claims to the Dalmatian coast. In July 1917 the factions united in the Corfu Declaration that envisioned a Kingdom of Serbs, Croats, and Slovenes. All the committees then gathered in Rome for a Congress of Oppressed Nationalities in April 1918.

The Corfu Declaration

The Allies stood aloof from the nationalities while hope persisted of detaching Austria-Hungary from Germany. But in 1918 the Allies took up the revolutionary weapon. In April 1918 Masaryk sailed to the United States, won personal recognition from Wilson and Secretary of State Robert Lansing, and concluded the Pittsburgh Convention by which Slovak-Americans, on behalf of their countrymen, agreed to join the Czechs in a united state. The Czechoslovak National Council won official recognition as a co-belligerent and de facto government-in-exile from France in June, Britain in August, and the United States in September. Only their quarrel with Italy kept the Yugoslavs from achieving the same. Thus, de facto governments were prepared to assume control of successor states as soon as Habsburg authority should collapse, internally or on the military fronts.

Germany's final battles. Ironically the Germans did not take maximum advantage of Brest-Litovsk after all, leaving about a million men—60 divisions—in the East in order to coerce the Ukrainians into relinquishing foodstuffs, to pursue political goals in the Baltic, and to ensure Bolshevik compliance. Facing virtual starvation as economic exhaustion deepened and the Allied blockade grew more effective, the German high command decided on a series of all-out attacks on the Western Front, beginning in March 1918. But tactical errors, together with the Allies' creation at last of a unified command and the arrival in strength of eager U.S. divisions, blunted and then turned back the offensives. By late July it was clear that Germany had lost the war. The 1918 offensives cost 1,100,000 men and drained the Reich of reserves. Morale plummeted on the Western Front and at home. Then on Aug. 8, 1918, British, Australian, and Canadian divisions struck on the Somme and overwhelmed German forces not adequately dug in. The 20,000 casualties, and an equal number of prisoners taken in one day, testified to the broken spirit of the German troops. Further Allied successes followed, and on Sept. 29, 1918, General Erich Ludendorff, the chief of staff, informed the Kaiser that the army was finished. The next day the new chancellor, the moderate Maximilian, Prince of Baden, was authorized to seek an armistice. On the night of October 3–4 he requested an armistice from President Wilson on the basis of the Fourteen Points.

Germany requests an armistice

While negotiations began for an armistice in the West, Germany's allies elsewhere collapsed. The collapse of the Bulgarian front before the Franco-Serbian offensive ended with the French cavalry capture of Skopje on September 29, whereupon the Allies accepted Bulgaria's petition for peace in the Armistice of Salonika. This opened Constantinople to attack and prompted the Turks as well to sue for peace. It also left Austria-Hungary, stymied on the Italian front, with little recourse. On October 4 Vienna appealed to President Wilson for an armistice on the basis of the Fourteen Points. But the U.S. note of the 18th indicated that autonomy for the nationalities no longer sufficed and thus amounted to the writ of execution for the Habsburg Empire. On October 28, in Prague and Kraków, Czech and Polish committees declared independence from Vienna. The Croats in Zagreb did the same on the 29th pending their union with the Serbs, and Germans in the Reichsrat proclaimed rump Austria an independent state on the 30th. The Armistice of Villa Giusti (November 4) required Austria-Hungary to evacuate all occupied territory, the South Tirol, Tarvisio, Gorizia, Trieste, Istria, western Carniola, and Dalmatia, and to surrender its navy. Emperor Charles, his empire gone, pledged to withdraw from Austria's politics on November 11 and from Hungary's on the 13th.

The first U.S. note responding to the German request for an armistice was sent on October 8 and called for evacuation by Germany of all occupied territory. The German reply sought to ensure that all the Allies would respect the Fourteen Points. The second U.S. note reflected high dudgeon about Germany's seeking assurances, given her own war policies. In any case, the British, French, and Italians (fearing Wilsonian leniency and angry about not being consulted after the first note) insisted that their military commands be consulted on the armistice terms. This in turn gave the Allies a chance to ensure that Germany

Allied solidarity on war aims

be rendered unable to take up resistance again in the future, whatever the eventual peace terms, and that their own war aims might be advanced through the armistice terms; *e.g.*, surrender of the German navy for the British, occupation of Alsace-Lorraine and the Rhineland for the French. Wilson's second note, therefore, shattered German illusions about using the armistice as a way of sowing discord among the Allies or winning a breathing space for themselves. The third German note (October 20) agreed to the Allies setting the terms and indicated, by way of appeasing Wilson, that Maximilian's civilian Cabinet had replaced any "arbitrary power" (Wilson's phrase) in Berlin. The third U.S. note (October 23) specified that the armistice would render Germany incapable of resuming hostilities. Ludendorff now wanted further resistance, but the Kaiser instead asked for his resignation on the 26th. The next day Germany acknowledged Wilson's note.

Some Allied leaders, most notably Poincaré and General John Pershing, bitterly disputed the wisdom of offering Germany an armistice when her armies were still on foreign soil. But Foch drafted military terms harsh enough for the skeptics, and Georges Clemenceau could not in good conscience permit the killing to go on if Germany were rendered defenseless. Meanwhile House, sent by Wilson to Paris to consult with the Allies, threatened a separate U.S.-German peace to win Allied approval of the Fourteen Points on November 4 (excepting a British reservation about "freedom of the seas," a French one about "removal of economic barriers and equality of trade conditions," and a clause enjoining Germany to repair war damage). House and Wilson jubilantly concluded that the foundations of a liberal peace were in place: substitution of the Fourteen Points for the Allies' "imperialist" war aims and the transition of Germany to democracy. The fourth U.S. note (November 5) informed the Germans of Allied agreement and the procedures for dealing with Marshal Foch.

Germany, however, seemed to be moving less toward democracy than anarchy. On October 29 the naval command ordered the High Seas Fleet to leave port for a last-ditch battle, prompting a mutiny, then full insurrection on November 3. Workers' and soldiers' councils formed in ports and industrial cities, and a socialist Republic of Bavaria was declared on the 8th. Two days later Maximilian announced the abdication of Kaiser William II and his own resignation, and Social Democratic leader Friedrich Ebert formed a provisional government. On the 10th the Kaiser went into Dutch exile. The armistice delegation led by Erzberger, meanwhile, met with Marshal Foch in a railway carriage at Rethondes on the 8th. Erzberger begged for amelioration of the Allies' terms and especially for the lifting of the blockade so that Germany might be fed. He raised the spectre of Bolshevism. Receiving only minor concessions, the Germans relented and signed the Armistice on Nov. 11, 1918. It called on Germany to evacuate and turn over to Allied armies all occupied regions, Alsace-Lorraine, the left (west) bank of the Rhine, and the bridgeheads of Mainz and Coblenz. A neutral zone of 10 kilometres on the right bank of the Rhine must also be evacuated, the entire German navy surrendered, and the treaties of Brest-Litovsk and Bucharest renounced. Germany was also to turn over a large number of locomotives, munitions, trucks, and other matériel—and to promise reparation for damage done.

Abdication and armistice

THE COLLAPSE OF THE OLD ORDER

The four years' carnage of World War I was the most intense physical, economic, and psychological assault on European society in its history. The war took directly some 8,500,000 lives and wounded another 21,000,000. The demographic damage done by the shortage of young, virile men over the next 20 years is incalculable. The cost of the war has been estimated at more than 200,000,000,000 1914 dollars, with some $36,800,000,000 more in damage. Much of northern France, Belgium, and Poland lay in ruin, while millions of tons of Allied shipping rested at the bottom of the sea. The foundation stone of prewar financial life, the gold standard, was shattered, and prewar trade patterns were hopelessly disrupted.

Economic recovery, vital to social stability and the containment of revolution, depended on political stability. But how could political stability be restored when four great empires—the Hohenzollern, Habsburg, Romanov, and Ottoman—had fallen, the boundaries of old and new states alike were yet to be fixed, vengeful passions ran high, and conflicting national aims and ideologies competed for the allegiance of the victors? In World War I, Europe lost its unity as a culture and polity, its sense of common destiny and inexorable progress. It lost much of its automatic reverence for the old values of country, church, family, duty, honour, discipline, glory, and tradition. The old was bankrupt. It remained only to decide which newness would take its place.

Spiritual cost of the war

The damage wrought by war would live on through the erosion of faith in 19th-century liberalism, international law, and Judeo-Christian values. For whatever the isolated acts of charity and chivalry by soldiers struggling in the trenches to remain human, governments and armies had thrown away, one by one, the standards of decency and fair play that had governed European warfare, more or less, in past centuries. Total war meant starving of civilians through naval blockade, torpedoing of civilian craft, bombing of open cities, poison gas in the trenches, and tactics of assault that took from the private soldier any dignity, control over his fate, or hope of survival. World War I subordinated the civilian to the military and the human to the machine. It remained only for such imperious cynicism to impose itself in peacetime as well, in totalitarian states modeled on war government, until the very distinction between war and peace broke down in the 1930s.

Peacemaking, 1919–22

The bells, flags, crowds, and tears of Armistice Day 1918 testified to the relief of exhausted Europeans that the killing had stopped and underscored their hopes that a just and lasting peace might repair the damage, right the wrongs, and revive prosperity in a broken world. Woodrow Wilson's call for a new and democratic diplomacy, backed by the suddenly commanding prestige and power of the United States, suggested that the dream of a new Jerusalem in world politics was not merely Armistice euphoria. A century before, Europe's aristocratic rulers had convened in the capital of dynasties, Vienna, to fashion a peace repudiating the nationalist and democratic principles of the French Revolution. Now, democratic statesmen would convene in the capital of liberty, Paris, to remake a Europe that had overthrown monarchical imperialism once and for all in this "war to end war."

In fact, the immense destruction done to the political and economic landmarks of the prewar world would have made the task of peacemaking daunting even if the victors had shared a united vision, which they did not. Central and Eastern Europe were in a turmoil in the wake of the German, Habsburg, Russian, and Ottoman collapses. Revolution sputtered in Berlin and elsewhere, and civil war in Russia. Trench warfare had left large swaths of northern France, Belgium, and Poland in ruin. The war had cost millions of dead and wounded and more than $200,000,000,000, not to mention the $30,000,000,000 more in damage to property and sunken ships. Ethnic hatreds and rivalries could not be expunged at a stroke, and their persistence hindered the effort to draw or redraw dozens of boundaries, including those of the successor states emerging from the Habsburg Empire. In the colonial world the war among the imperial powers gave a strong impetus to nationalist movements. India alone provided 943,000 soldiers and workers to the British war effort, and the French empire provided the home country with 928,000. These men brought home a familiarity with European life and the new anti-imperialist ideas of Wilson or Lenin. The war also weakened the European powers vis-à-vis the United States and Japan, destroyed the prewar monetary stability, and disrupted trade and manufactures. In sum, a return to 1914 "normalcy" was impossible. But what could, or should, replace it? As French foreign minister Stéphen Pichon observed, the war's end only meant that "the era of difficulties begins."

The legacy of war

The Paris Peace Conference ultimately produced five treaties, each named after the suburban locale in which it was signed: the Treaty of Versailles with Germany (June 28, 1919); the Treaty of Saint-Germain with Austria (Sept. 10, 1919); the Treaty of Neuilly with Bulgaria (Nov. 27, 1919); the Treaty of Trianon with Hungary (June 4, 1920); and the Treaty of Sèvres with Ottoman Turkey (Aug. 10, 1920). In addition, the Washington Conference treaties on naval armaments, China, and the Pacific (1921–22) established a postwar regime in those areas.

COMPETING VISIONS OF STABILITY

The idealist vision. According to the armistice agreement the peace was to be based on Wilson's Fourteen Points. But the French and British had already expressed reservations about them, and, in many cases, the vague Wilsonian principles lent themselves to varying interpretations when applied to complex realities. Nevertheless, Wilson anticipated the peace conference with high hopes that his principles would prevail, either because of their popularity with common people everywhere, or because U.S. financial leverage would oblige European statesmen to follow his lead. "Tell me what is right," he instructed his delegation on the *George Washington* en route to Paris, "and I will fight for it." Unique among the victor powers, the United States would not ask any territorial gains or reparations and would thereby be free to stand proudly as the conference's conscience and honest broker.

America as honest broker

Wilsonianism, as it came to be called, derived from the liberal internationalism that had captured large segments of the Anglo-American intellectual elite before and during the war. It interpreted war as essentially an atavism associated with authoritarian monarchy, aristocracy, imperialism, and economic nationalism. Such governments still practiced an old diplomacy of secret alliances, militarism, and balance of power politics that bred distrust, suspicion, and conflict. The antidotes were democratic control of diplomacy, self-determination for all nations, open negotiations, disarmament, free trade, and especially a system of international law and collective security to replace raw power as the arbiter of disputes among states. This last idea, developed by the American League to Enforce Peace (founded in 1915), found expression in the Fourteen Points as "a general association of nations" and was to be the cornerstone of Wilson's edifice. He expected a functioning League of Nations to correct whatever errors and injustices might creep in to the treaties themselves.

Liberal internationalism set the tone for the Paris Peace Conference. European statesmen learned quickly to couch their own demands in Wilsonian rhetoric and to argue their cases on grounds of "justice" rather than power politics. Yet Wilson's principles proved, one by one, to be inapplicable, irrelevant, or insufficient in the eyes of European governments, while the idealistic gloss they placed on the treaties undermined their legitimacy for anyone claiming that "justice" had not been served. Wilson's personality must bear some of the blame for this disillusionment. He was a proud man, confident of his objectivity and prestige, and he insisted on being the first U.S. president to sail to Europe and to conduct negotiations himself. He had visited Europe only twice before, as a tourist, and now delayed the peace conference in order to make a triumphant tour of European capitals. Moreover, the Democrats lost their Senate majority in the elections of November 1918, yet Wilson refused to include prominent Republicans in his delegation. This allowed Theodore Roosevelt to declare that Wilson had "absolutely no authority to speak for the American people." Wilson's flaws exacerbated the difficulty of promoting his ideals in Paris and at home. Still, he was a prophet in world politics, both as lawgiver and as seer. Only a peace between equals, he said, can last.

Woodrow Wilson's triumph

The realist vision. Georges Clemenceau also approached peacemaking as a personal quest, stacking the French delegation with loyal supporters and minimizing the influence of the foreign ministry, the army, and parliament. Even political enemies hailed Clemenceau (known as "the tiger") as "père la victoire," and he determined not to betray the soldiers' victory in the peace negotiations to come. But the French vision of a just peace contrasted sharply with Wilson's. France alone in 1914 had not chosen war, but had been summarily attacked. France had provided the major battleground, suffered the most physical damage, and sacrificed a generation of manhood. France faced the most massive task of reconstruction, the most direct threat of German revenge, and the most immediate responsibility for executing the armistice and peace treaties by dint of its contiguity with Germany. Clemenceau, therefore, sought material advantage from the peace according to a traditional balance-of-power viewpoint and did so with almost universal support in the government. The 77-year-old Clemenceau, who had begun his political career during the German siege of Paris in 1870–71, placed little faith in Germany's sudden conversion to democracy, nor in Wilson's lofty idealism, which he characterized with irony as "noble candour." The French government judged early on that Wilson's dream of a prosperous German republic taking its place in the council of nations was the primary obstacle to a peace serving France's real needs. Indeed, his decision to accept the armistice may have been influenced by the fact that a more thorough victory over Germany would also have meant another million American soldiers at the front and proportionally greater U.S. influence over the peace.

Postwar France faced a severe triple crisis. The first involved future security against German attack: Germany remained far more populous and industrial than France, and now France's erstwhile eastern ally, Russia, was hors de combat. The French would try to revive an anti-German alliance system with the new states in eastern Europe, but the only sure way to restore a balance of power in Europe was to weaken Germany permanently. The second crisis was financial. France had paid for the war largely by domestic and foreign borrowing and inflation. To ask the nation to sacrifice further to cover these costs was politically impossible. Indeed, any new taxes would spark bitter social conflict over which groups would bear the heaviest burdens. Yet France also faced the cost of rebuilding the devastated regions and supporting an army capable of forcing German respect for the eventual treaty. The French, therefore, hoped for inflows of capital from abroad to restore their national solvency. Third, France faced a crisis in her heavy industry. The "storm of steel" on the Western Front made obvious the strategic importance of metallurgy in modern war. Recovery of Alsace-Lorraine lessened France's inferiority to Germany in iron but by the same token worsened her shortage of coal, especially metallurgical coke. European coal production was down 30 percent from prewar figures by 1919, creating acute shortages everywhere. But France's position was especially desperate after the flooding of French mines by retreating German soldiers. To realize the industrial expansion made possible by the recovery of Alsace-Lorraine, France needed access to German coal and markets and preferably a cartel arrangement allowing French industry to survive German competition in the peacetime to come.

France's triple crisis

Wilson's program was not without promise for France if collective security and Allied solidarity meant permanent British and American help to deter future German attacks and restore the French economy. In particular, the French hoped that the wealthy United States would forgive the French war debts. On the other hand, if Britain and the United States pursued their own interests without regard to French needs, then France would be forced to find solutions to its triple crisis through harsher treatment of Germany.

In some respects, Britain stood between France and the United States. It would be more accurate, however, to view Britain as the third point of a triangle, attached to the interests of France in some cases, to the principles of the United States in others. Hence, Prime Minister David Lloyd George, second only to Wilson in liberal rhetoric, was accused by Americans of conspiring with Clemenceau to promote old-fashioned imperialism, and, second only to the French in pursuing balance of power, was accused by Clemenceau of favouring the Germans. But that was Britain's traditional policy: to prop up the defeated power in a European war and constrain the ambitions of the victor. To be sure, in the election campaign held after the

Lloyd George's third way

Armistice, Lloyd George's supporters brandished slogans like "Hang the Kaiser" and "Squeeze the German lemon til the pips squeak," but at the peace conference to come, Lloyd George equivocated. Britain would take the toughest stand of all on German reparations in hopes of ameliorating its own financial situation vis-à-vis the United States, but otherwise promoted a united, healthy Germany that would contribute to European recovery and balance the now ascendant power of France. Of course, Lloyd George also demanded a ban on German naval armaments and partition of Germany's colonies.

Exhausted Italy was even less able than France to absorb the costs of war. Labour unrest compounded the usual ministerial instability and enhanced the public appeal of anti-Communist nationalists like Benito Mussolini. But the hope that the war would prove somehow worthwhile put peace aims at the centre of Italian politics. In April 1918 the terms of the Treaty of London were proclaimed on the floor of Parliament, sparking months of debate between nationalists and Wilsonians over their propriety. By January 1919, however, Prime Minister Vittorio Emanuele Orlando and Foreign Minister Sidney Sonnino had won a mandate for a firm position at the peace conference in favour of all Italy's claims with the exception of that to the entire Dalmatian coast.

Japan's war gains

The other victorious Great Power, Japan, suffered the least human and material loss in the war and registered astounding growth. Between 1913 and 1918 Japanese production exploded, foreign trade rose from $315,000,000 to $831,000,000, and population grew 30 percent until 65,000,000 people were crowded into a mountainous archipelago smaller than California. Clearly Japan had the potential and the opportunity for rapid expansion in the Pacific and East Asia.

Finally, the defeated Germans also looked with hopes to the peace conference. Throughout the first half of 1919 the new Weimar Republic (so called after the site of its constitutional convention) was in gestation, and the Germans hoped that their embrace of democracy might win them a mild peace. At the very least they hoped to exploit differences among the victors to regain diplomatic equality, as Talleyrand had done for France at the Congress of Vienna. Instead, the Allies found compromise among themselves so arduous that they could brook no further negotiation with Germany. German delegates were not invited to Paris until May, and the "preliminaries of peace" became, with few exceptions, the final treaty. To Germans, Wilson's promise of "open covenants, openly arrived at" proved a sham, and the final treaty a *Diktat.*

THE VERSAILLES "DIKTAT"

Hammering out the treaty. The Paris Peace Conference opened on Jan. 18, 1919, in a politically charged atmosphere. The delegations of 27 nations harassed the Great Powers with their various and conflicting complaints and demands. The Great Powers, in turn, sent five delegates each, supported by sprawling staffs of geographers, historians, and economists. Clearly, peace could not be made in such a global assembly; hence the five leading victors created a Council of Ten—the heads of government and their foreign ministers. But even this proved unwieldy, and since Italy and Japan tended to focus on questions of local interest, major decisions were hammered out in private by an informally constituted Big Three: Wilson, Lloyd George, and Clemenceau. The French had tried to impose a schedule of priorities for the conference, but Wilson insisted on tackling the League of Nations first in order to prevent the others from rejecting the League or using it as a bargaining chip in later disputes. The French were skeptical of the idealistic basis of the League but hoped that it might be turned into an instrument of security committing the British and Americans to the defense of the new European order. In this they were disillusioned, for the British viewed the League less as a means for mobilizing force against an aggressor than as a means of preventing future conflicts in the first place.

The Covenant of the League

The Covenant of the proposed League provided for a plenary assembly of all members and a council of the Great Powers and outlined a system of sanctions against aggressor states. But the British chose to focus on moral sanctions (not unlike Wilson's belief in the "court of world opinion"), or at most economic sanctions, and participation in military sanctions was made voluntary. The Covenant also contained machinery for declaring boundary changes, implying that the League's primary function was to secure peace, not to secure the status quo. Upon final rejection in April of a Franco-Italian plan for tougher collective security and an international force adequate to enforce peace, French newspapers scorned the League as a toothless debating society. And since Clemenceau had succeeded in having Germany barred from the League pending good behaviour, the German press denounced it as a "League of Victors."

In mid-February Wilson returned to the United States to attend to presidential duties, and in his absence committees went to work on the details of the German treaty. Foremost in the minds of the French was security against future German attack. As early as November 1918 Marshal Ferdinand Foch drafted a memo identifying the Rhine as "the frontier of democracy" and arguing for the separation of the Rhineland from Germany and its occupation in perpetuity by Allied troops. This plan echoed earlier French war aims: The victory of 1871 had created a unified Germany; the defeat of 1918 should undo it. Foch's occupation forces tried also to locate and encourage the Rhenish autonomist tendencies that grew up for a brief time in 1919 out of the desire to escape the burden of defeat and fear of the Communist agitation in Berlin. But the primary French argument was strategic: Four times in a century German armies had invaded France from the Rhineland (1814, 1815, 1870, 1914), and a united Germany would remain potentially overwhelming. As General Fayolle put it, "One speaks of the League, but what can this hypothetical society do without a means of action? One promises alliances, but alliances are fragile, like all human things. There will always come a time when Germany will have a free hand. Take all the alliances you want, but the greatest need for France and Belgium is a material barrier."

The Rhineland question

André Tardieu, Clemenceau's chief aide, sought to give the Rhineland scheme a Wilsonian gloss in a lengthy memo distributed on February 25. The Rhenish people, he claimed, were largely Celtic, Catholic, and liberal and resented the rule of Germanic, Protestant, and authoritarian Prussia. They had been loyal citizens of the French Republic and Empire from 1792 to 1815. Thus an autonomous Rhineland would serve both self-determination and the defense of democracy. The British and Americans rejected Tardieu's brief in the strongest terms and warned that dismemberment of Germany would only create "a new Alsace-Lorraine" and the seeds of a new war. In April, after Wilson returned to Paris, he and Lloyd George countered with an unprecedented offer: an Anglo-American guarantee to fight on the side of France in case of future German aggression. The French were again skeptical. In a future war the United States and Britain would need months or years to raise and transport armies, by which time France might be lost. On the other hand, how could Clemenceau refuse an unlimited extension of the wartime coalition? On March 17 he proposed a mixed solution—the guarantee treaties, plus material safeguards including German disarmament, demilitarization, and Allied occupation of the Rhine.

This acrimonious debate over security overlapped with the negotiations over reparations. The latter was perhaps an even more emotional issue, since the financial settlement would affect every taxpayer in every country. The moral issues also seemed clearer: Surely Germany, and not her victims, should pay for reconstruction; surely the wealthy British and Americans should forgive France's war debt, a small sacrifice beside those made by France in the joint effort. The French government had borrowed 26,000,000,000 francs from its own people during the war and owed another $3,600,000,000 to Britain and the United States. The franc had lost 70 percent of its value. Yet French hopes for Allied economic unity were dashed when the U.S. Treasury refused to discuss abrogation of war debts, rejected French and Italian proposals for a

"financial League of Nations," and opposed economic favouritism of all kinds in accord with the Fourteen Points. The British, in turn, repudiated the resolutions of the 1916 Allied Economic Conference and refused to forgive France her debt so long as the United States insisted on repayment from London.

"If it is France or Germany that must be ruined," wrote a conservative French journal about the reparations debate, "let us be sure that it is Germany!" The French chamber refused to vote a tax on capital and relied on German payments to cover the cost of repairing the devastated regions. Wilson accepted German responsibility for war damage, but the British vastly inflated reparations by insisting on repayment for "invisible damage" like sunken ships and cargo, lost markets and production, and veterans' pensions. On the other hand, the British favoured setting a fixed indemnity in the treaty, while the French claimed that Germany should agree to pay whatever reparation ended up costing. When negotiations failed to fix either a total sum or the percentage shares to flow to France, Britain, Belgium, and the others, the U.S. delegation recommended on March 24 that the whole problem be postponed. On April 5 it was agreed that a Reparations Commission would determine, by May 1, 1921, the amount and timing of German payments and be empowered to declare defaults and sanctions in case of noncompliance. But in the meantime Germany would make immediate transfers totaling 20,000,000,000 gold marks. Thus the peace conference obliged the Germans to sign an open account and adjourned without plans to stabilize currencies or settle war debts.

The Reparations Commission

In economic matters the French delegation laboured to improve the imbalance in heavy industry between Germany and France. At first Clemenceau fought hard for annexation of the Saar—the French "frontier of 1814"—and then settled for French control of the Saar coal mines and a League of Nations administration for 15 years, at which time the Saarlanders would hold a plebiscite to decide their permanent status. Germany was also obliged to deliver 20,000,000 tons of coal per year to France and Belgium and to allow the products of Alsace-Lorraine into Germany duty-free for five years.

Such punitive clauses ensured German feebleness for some time to come. France, on the other hand, now possessed both the largest army in Europe and a set of natural allies among the new states in eastern Europe. Not surprisingly, many British observers came to consider France the primary threat to dominate the Continent. In late March Lloyd George's eloquent Fontainebleau Memorandum warned that vindictiveness in the hour of victory would serve not justice and reconciliation but German revanchism and Bolshevik propaganda. Nevertheless Clemenceau, under attack from President Poincaré, Marshal Foch, and the parliament for "giving up the Rhine," dared not compromise further. On April 22, Wilson and Lloyd George accepted his material guarantees of security in addition to the Anglo-American pacts. These included the limitation of the German army to 100,000 men with no offensive weapons; demilitarization of a zone extending 50 kilometers east of the Rhine; and an Allied occupation of the left bank of the Rhine, with bridgeheads at Cologne, Koblenz, Mainz, and Kehl. The occupation would be divided into three zones, to be evacuated serially at five-year intervals.

Reaction to the treaty. On May 7 the German delegation was finally summoned to receive the draft treaty. Additional important clauses called for the abolition of the German high seas fleet, the general staff, and conscription; partition of Germany's African colonies; cession of the Eupen-et-Malmédy district to Belgium, Alsace-Lorraine to France, most of Upper Silesia and West Prussia to Poland, including a corridor to the Baltic that cut Germany in two; plebiscites to determine whether Allenstein and Marienwerder should go to Poland and Schleswig to Denmark; a League of Nations administration for the free city of Danzig (to provide Poland a coastal port); prohibition of *Anschluss* (union) between Germany and Austria; and abrogation of the Treaty of Brest-Litovsk. Finally, Article 231 enjoined Germany to accept full responsibility for the war caused "by the aggression of Germany and her allies."

German territorial cessions

The draft treaty caused acute consternation in Germany (though it left Germany intact and was mild compared to Germany's terms to Russia at Brest-Litovsk), and the German delegation argued without success for substantial revisions. The Germans could not reject the treaty, however, without inviting a continuation of the Allied blockade, revolutionary outbreaks, an Allied military advance, or French intrigues against German unity. (On June 1, Foch's generals in the occupation implicated themselves in an abortive separatist putsch aimed at creating a "Rhineland Republic" and thereby magnified German—and British—suspicions.) Hence, the German delegation—frock-coated professionals bearing little resemblance to the spike-helmeted militarists the Allies meant to punish—affixed their signatures to the treaty in the Hall of Mirrors at Versailles on the fifth anniversary of the Sarajevo assassination (June 28, 1919). The Weimar coalition of Democrats, Social Democrats, and the Catholic Centre party ratified the treaty on July 9. German nationalists, however, denounced acceptance of the treaty as treason and immediately began propounding the myth that the German army had been "stabbed in the back" by Socialists and defeatists, the "November criminals" who signed the Armistice, and the liberal parties who signed the Versailles *Diktat*. The war-guilt clause was particularly damaging, since any historical evidence suggesting that Germany did not bear sole guilt for the war would tend to undermine the treaty's legitimacy.

Allied delegates and populations were scarcely happier with the treaty than the Germans. British diplomat Harold Nicolson echoed the views of disillusioned Wilsonians when he left the signing ceremony in disgust, "and thence to bed, sick of life." Economist John Maynard Keynes quit the peace conference in protest and returned to Britain to write a scathing critique of Wilson and the treaty, whose economic clauses, he said, stymied European recovery. Nor were the French satisfied. Marshal Foch despaired of containing the power of a united Germany and prophesied: "This is not peace, but a truce for 20 years." Poincaré predicted willful German default and Allied disputes over execution. Clemenceau had to exploit all his prestige to win parliamentary ratification, and still he lost the presidential election that followed.

As for Wilson, the treaty he had personally helped to fashion, and the global obligations it imposed on the United States, proved unpopular with various factions in American politics, including nationalists, isolationists, "Monroe Doctrine" regionalists, xenophobes, and tariff protectionists. The immediate postwar years also gave rise to the "red scare," the first legislation limiting immigration to the United States on an ethnic basis, and the belief that Wilson had been duped by the clever Europeans so that the war redounded only to the benefit of Anglo-French imperialism. But it is not true that the United States retreated at once into isolationism. The debate over Versailles was essentially a debate over the terms on which the United States would continue to play a role in world affairs. Most important was fear that Article 10 of the League Covenant might embroil the United States in foreign quarrels and even violate the Constitution. The Senate Committee on Foreign Relations, led by Henry Cabot Lodge, eventually proposed ratification of the Treaty of Versailles subject to 14 reservations, but Wilson insisted on an all-or-nothing strategy and embarked on a hectic national tour to mobilize public support. In October 1919 he suffered a debilitating stroke, and on November 19 the Senate voted down the treaty. Further compromise led to a final vote on March 19, 1920, but Wilson instructed his own loyalists to reject any reservations. The 49–35 vote fell short of the necessary two-thirds majority. By failing to ratify the Treaty of Versailles, the United States also rejected the League of Nations (which its own president had forced on the Europeans), the security guarantee by which Clemenceau had been persuaded to give up the Rhineland, and U.S. commitment to the economic and political reconstruction of Europe. All this gave those who clung to the belief that the French cause had been betrayed the opportunity to deal even more harshly with Germany.

U.S. reaction to the treaty

THE WEST AND THE RUSSIAN CIVIL WAR

Bolshevik diplomacy. France's deep fears about a future German threat sprang in large part from the elimination of Russia as a factor in the European balance. Indeed, the Russian question was at least as important as the German one and absorbed as much time and worry at the peace conference. After Brest-Litovsk, Anglo-French policy turned sharply anti-Bolshevik, and Clemenceau and Foch worked to build a *cordon sanitaire* in eastern Europe against German and Bolshevik expansion alike. The Lenin regime also repudiated the tsarist debts to Britain and France (the latter being more delicate since most of it dated from before the war and was owed to private bondholders). But Wilson still believed in the innate desire of the Russian people for democracy and searched desperately for ways to end the civil war and liberalize the Reds, the Whites, or both. As early as July 1918 he wrote Colonel Edward House: "I have been sweating blood over what is right and feasible to do in Russia. It goes to pieces like quicksilver under my touch."

France's *cordon sanitaire*

After Brest-Litovsk the Bolsheviks came quickly to a two-track policy toward the West. Their rhetoric still condemned Allied and German imperialists in vitriolic terms, but their deeds aimed at securing their own survival at all costs. These included attempts to open negotiations with Allied governments, to exploit differences among them, to persuade them to withdraw support for the Whites, and to encourage the opposition to intervention in Russia that already existed among French and British workers and soldiers. On the other hand, the Red Terror launched by the Bolsheviks in 1918, including the murder of the royal family, convinced many in the West that this new breed was beyond the pale. U.S. Secretary of State Robert Lansing called Bolshevism "the most hideous and monstrous thing that the human mind has ever conceived." When, in August 1918, the Cheka (secret police) arrested 200 British and French residents of Moscow, invaded their consulates, and murdered the British naval attaché, opinion spread in Paris and London that the Bolsheviks were thugs and bandits, if not German agents. In the autumn the Allies imposed a blockade on the Moscow regime and broke the last contacts (diplomatic missions and the Red Cross) that still existed.

The Bolsheviks' paramount need was a breathing spell in which to consolidate their power, mobilize the economy in the lands under their control, and subdue the White armies. By the end of 1918 these forces included the Cossacks of General Anton Denikin in the south, supported by the French from Odessa; the Ukrainian separatists; General Nikolay Yudenich's army of the Baltic; a puppet government in the north supported by the Anglo-French from Arkhangelsk; and the government of Admiral Aleksandr Kolchak at Omsk in Siberia. American and Japanese troops occupied Vladivostok on the Pacific. The Bolsheviks had also invaded Estonia only to be met by local troops, a British naval squadron, Yudenich's Russian nationalists, and even General Rüdiger von der Goltz's German veterans seeking to maintain German authority on the Baltic. Against these disparate and uncoordinated forces the Bolsheviks deployed the Red Army under the command of Leon Trotsky. In the opening stages of the Revolution they experimented with a "people's army" in which ranks were abolished and officers were elected by the troops. This quickly gave way to traditional military practice and even recruitment of ex-tsarist officers and technicians. By the turn of 1919 the Red Army numbered in the millions.

The opponents of the Bolsheviks

Lenin instructed the new commissar for foreign affairs, Georgy Chicherin, to try to separate the United States from the Allies. In October and November 1918 he addressed long notes to Wilson protesting Allied intervention and proposing a cease-fire in return for Allied evacuation. Then in December, Maksim Litvinov appealed to Wilson in terms drawn from the Fourteen Points, ending with the plea *auditur et altera pars* ("let the other side be heard"). Some historians have judged these demarches as a genuine opportunity for early reconciliation between the Bolsheviks and the West. Others consider them the equivalent of the Brest-Litovsk negotiations with the Germans, a "peace offensive" designed to serve the internal security of the regime. The Western powers, however, were confused about how to influence events in Russia. In January 1919, Lloyd George showed Wilson an intelligence report indicating that the Allied interventions, if not increased massively, would only strengthen the appeal of the Bolsheviks. He favoured negotiation; Clemenceau favoured a stronger intervention.

Given the Bolsheviks' single-minded dedication to power and ideology (which was, after all, their sole source of legitimacy), it is difficult to imagine how Allied–Soviet friendship, or a compromise settlement among the Russian factions, could have emerged. Nevertheless, the snarled diplomacy of the two sides during the peace conference widened the gap between them. Lenin had postponed his summons to European Socialists to form the Third (or Communist) International (Comintern) until January lest it spoil his efforts to open negotiations with the West. He finally issued the call on Jan. 25, 1919, just as the Paris Peace Conference finally decided to make an initiative. It appeared, therefore, as if Lenin was intent on remaining an international outlaw seeking to destroy the very governments with which he claimed to want normal relations. The Comintern was founded on March 2, and at its second congress (July 1920) Lenin insisted that member parties accede to 21 conditions imposing rigorous Communist discipline and subordinating local parties to the will of Moscow. It divided European Socialists, most of whom rejected the Communists' violent tactics, Lenin's dictatorship, or both. From its inception, therefore, the Comintern was an arm of Soviet foreign policy more than a vehicle of Socialist internationalism.

Founding of the Comintern

Allied approaches to the Bolsheviks. Meanwhile, Wilson and Lloyd George agreed on an appeal directed to the White forces (and radioed to the Bolsheviks) to declare a cease-fire and send representatives to the island of Prinkipo (Büyükada), in the Sea of Marmara. This was a fruitless gesture, since neither the Red nor the White regime could survive except by the other's total destruction. The Bolsheviks ignored the call for a truce but accepted the invitation; the Whites, with French encouragement, candidly declined both. The Big Three were informed of the failure on February 12, two days before Wilson's return to the United States. Winston Churchill then hurried to Paris to urge on Wilson a vigorous Allied military campaign on behalf of the Whites. But even if the Big Three had agreed to launch an anti-Bolshevik crusade, their war-weary populations, depleted treasuries, and aroused labour unions would not have permitted it.

Five days later Colonel House, who was given charge of Russian matters by Wilson, asked a young American liberal, William Bullitt, to journey to Russia for direct talks with Lenin. Bullitt reached Petrograd on March 8, spoke with Chicherin and Litvinov, then went on to Moscow. Lenin offered an immediate cease-fire and negotiations in return for the cessation of Allied occupation, aid to the Whites, and the blockade. The Bolsheviks, in turn, promised amnesty to all Russians who had collaborated with the Allies. Bullitt returned to Paris in great excitement at the end of March, only to be denied an audience with Wilson and to find the conference near collapse over the Rhineland question. Lloyd George was under pressure from parliamentary Tories to avoid conciliating Lenin, while the general level of Allied anxiety had been raised by declaration of a Soviet republic in Bavaria and Béla Kun's Communist coup d'état in Hungary on March 21. Kun immediately invaded Czechoslovakia and appealed to Lenin for help (which the Bolsheviks were in no condition to provide). On April 10 a Romanian army attacked Hungary, and successive Red and White terrors ensued. The episodes ended on May 1, when German federal troops deposed the Bavarian Communists, and August 1, when Kun fled the approaching Romanian army.

Communist revolts

Historians debate whether the Bullitt mission was a missed opportunity. Considering the Bolsheviks' final victory, the Allies would have done well to extricate themselves on Lenin's March 1919 terms. On the other hand, the document held out little hope for a Russia in line with Western principles or interests. Allied acceptance would

have obliged them to pull out their own forces, cut off aid to the Whites, and resume trade with the Bolsheviks. If hostilities had then resumed—on any pretext—the Reds would have been able to crush the divided Whites and solidify their control. On the other hand, Lenin was hard pressed in the spring of 1919—Kolchak was launching a major offensive—and was probably sincere in seeking relief. Bullitt himself was consumed with bitterness over his reception in Paris and rebuked Wilson for having "so little faith in the millions of men, like myself, in every nation who had faith in you." (Bullitt testified before the Senate against the Versailles treaty and retired to France until, in 1933, he was appointed the first U.S. ambassador to the Soviet Union. Disillusioned with Stalin, he soon resigned.)

The fourth approach by the peace conference to Russia grew out of letters from the director of European food relief, Herbert Hoover (March 28), and the Norwegian explorer and philanthropist Fridtjof Nansen (April 3) urging massive deliveries of food to Russia. The way to fight Communism, they argued, was with bread, not guns. Colonel House procured Allied consent to offer relief to Russia, but only if Russian transportation facilities were placed at the disposal of an Allied commission. The Bolsheviks replied in derisory terms on May 13, since the conditions would have meant de facto Allied control of Russia. (In 1921 the American relief commission nonetheless began distribution of food that saved countless Russians from starvation.)

Consolidation of the Revolution. The peace conference's inability to frame a common policy toward the Lenin regime meant that Russia's future was now solely a military matter. By May, Kolchak's offensive reached its greatest extent, approaching Moscow from the east, and the French and British resolved to recognize the Whites. Wilson also gave up on the Reds and began cajoling White leaders to pledge democratization of Russia in the event of their victory. But the Red Army turned back Kolchak in the summer, and the Allies gave up in the north, evacuating Arkhangelsk, after a number of clashes with Red forces, on Sept. 30, 1919, and Murmansk on October 12.

War Communism

The Russian Civil War was a vast, protean struggle fought out in five major theatres with rapid thrusts over hundreds of miles made possible by railroads and cavalry. The Reds took good advantage of their interior lines, while their control of Russia's industrial heartland and trunk rail lines and their ruthless requisitioning (known as "War Communism") procured enough food and supplies for them to outlast their enemies. The outcome was not inevitable, but the inability of the far-flung White forces to coordinate their actions exposed them to defeat in detail. Denikin took Kiev in September 1919, but a Soviet counteroffensive forced him steadily back until his last base fell in March 1920. Command in the south fell to General Pyotr Wrangel. Meanwhile, the Red Army drove out Kolchak and recaptured Omsk in November 1919. On April 25, 1920, war broke out between the Soviets and Poland as the Polish leader, Marshal Józef Piłsudski, pursued his ambition of a grand Polish-Lithuanian-Ukrainian empire. On May 7 the Poles captured Kiev, but a Soviet counterstroke drove them out (June 11), captured Vilnius (July 15), and soon threatened Warsaw itself. Alarms arose in western Europe over the possible sovietization of Poland and even a German–Bolshevik alliance to overthrow the Treaty of Versailles. But Piłsudski, with advice from French attaché General Maxime Weygand, hurled back the overextended Reds, took 66,000 prisoners, and recaptured extensive Belorussian territories. Distressed by the resistance of the Poles to the Revolution, Lenin counseled peace, as at Brest-Litovsk, even on humiliating terms. A preliminary treaty (October 12) and final Treaty of Riga (March 18, 1921) fixed the Soviet–Polish border just to the west of Minsk and far to the east of the Curzon Line proposed at Paris.

Peace with Poland freed the Red Army to turn south and eliminate the last resistance from Wrangel, who evacuated the Crimea on Nov. 14, 1921. Soviet forces invested the Caucasus as well, setting up an "autonomous" federation of Communist regimes in Georgia, Armenia, and Azerbaijan. The original anti-imperialism of the Bolsheviks thus gave way to a policy of domination of all the subject nationalities of the Russian Empire that the Bolsheviks could subdue. On Oct. 25, 1922, the Japanese withdrew from Vladivostok under U.S. pressure, bringing all foreign interventions in Russia to a close.

Formation of the U.S.S.R.

The Union of Soviet Socialist Republics came into existence on Dec. 30, 1922. In the World War and Civil War, Russia had lost Poland, Finland, the Baltic states, and Bessarabia. The Communist government had survived, but the Revolution had failed to spread. Hence, the Bolshevik leaders were left to construct a permanent relationship to an outer world which they defined as implacably hostile. The Western powers, in turn, faced the challenge of living with a Great Power that repudiated, at least publicly, all norms of international behaviour.

CENTRAL EUROPE AND THE MIDDLE EAST

The reorganization of central Europe. Although the Habsburg Empire had ceased to exist, the peace conference dealt with the new republics of Austria and Hungary as defeated powers and systematically favoured the interests of the successor states that had arisen from the ruins of the empire in the last weeks of the war. It was Wilson's hope that peace and self-rule might finally bless the troubled regions between Germany and Russia through strict application of the principle of nationality. But east-central Europe comprised a jumble of peoples with conflicting claims based on language, ethnicity, economics, geography, military considerations, and historic ties. What was more, the new states themselves were in no case homogeneous. The name Yugoslavia could not hide the rivalries within that kingdom of Serbs, Croats, and Slovenes. Czechoslovakia was born of an alliance of convenience among Czechs, Slovaks, and Ruthenes. Historic Poland embraced Ukrainians, Germans, Lithuanians, and Yiddish-speaking Jews. Romania, enlarged by the accession of Transylvania and Bessarabia, now numbered millions of Ukrainians, Hungarians, Jews, and other minorities. In short, the Balkanization of central Europe raised more political disputes than it solved and created many little multinational states in place of a few empires.

The Polish patchwork

Poland was a favourite of the Americans and the French by dint of historic sympathies, the votes of Polish-Americans, and Clemenceau's hope for a strong Polish ally in Germany's rear. The Fourteen Points promised Poland an outlet to the sea, but the resulting corridor on the mouth of the Vistula contained 1,500,000 Kashubians and Germans. In the north, the Baltic states of Lithuania, Latvia, and Estonia won their independence from Moscow and were sheltered by the British fleet. But in a perfect example of the vagaries of national self-determination, Lithuania and Poland quarreled over the disposition of Vilnius. The town of Vilnius (according to 1897 Russian statistics) was 40 percent Jewish, 31 percent Polish, 24 percent Russian, and 2 percent Lithuanian. Vilnius Province, however, was 61 percent Russian, 17 percent Lithuanian, 12 percent Jewish, and 8 percent Polish. In December 1919 the Supreme Allied Council provisionally awarded Vilnius to Lithuania. Poland and Czechoslovakia similarly quarreled over the coal-rich Teschen district. Poles predominated in the district, but historic claims lay with Bohemia. In the end the Great Powers merely ratified the de facto partition effected by occupying Polish and Czech troops—a solution that favoured Czechoslovakia and left a bitterness the two states could ill afford and never overcame. Finally, the Polish–German conflict over Upper Silesia, another coal-rich region of mixed nationality, proved that even the League of Nations could not make an objective judgment. The March 1921 plebiscite called for in the Treaty of Versailles (one of the few concessions awarded the German delegation) showed German preponderance in the region as a whole but Polish majorities in the vital mining districts. The British delegation in the League argued that Germany could hardly be expected to pay reparations if it lost yet another rich source of coal, while the French sought to weaken Germany further and bolster the Polish economy. Finally, in October 1922, Poland was granted the greater portion of the mines.

The Treaty of Saint-Germain disposed of the Austrian

half of the former Habsburg monarchy. Tomáš Masaryk and Edvard Beneš, sincere Wilsonians, exploited their personal goodwill to win two major concessions that otherwise violated the principle of national self-determination. First, they retained for Czechoslovakia the entire historic province of Bohemia. This afforded the vulnerable new state the military protection from Germany of the Sudeten mountains, but it also brought 3,500,000 Sudeten Germans under the rule of Prague. Second, Czechoslovakia received territory stretching south to Bratislava on the Danube, providing it with a riverine outlet but creating a minority of a million Magyars. The Austrian boundary with Yugoslavia at Klagenfurt was fixed by plebiscite in Austria's favour in October 1920, as was the division of the Burgenland district between Austria and Hungary in December 1921.

The Sudeten Germans

Italy's boundaries with Austria and Yugoslavia became one of the most volatile issues of the peace conference owing to Italian truculence and Wilsonian sanctimoniousness. Orlando clung to the Allied promises that had enticed Italy into the war in the first place. But Wilson, offended by the secret war-aims treaties, vented his frustration on Italy. He went so far as to plead his case publicly in the French press on April 24, 1919, a violation of diplomatic etiquette that provoked the Italians to bolt the conference. Upon their return, a compromise of sorts was achieved: Italy received Trieste, parts of Istria and Dalmatia, and the Upper Adige as far as the Brenner Pass with its 2,000,000 German-speaking Austrians. But Wilson refused to budge on Fiume, a province whose hinterland was Yugoslav but whose port city was Venetian. On June 19 Orlando's government fell over the issue. In August, Fiume was declared a free city, and in September a band of Italian freebooters led by the nationalist poet Gabriele D'Annunzio, declared Fiume a free state. Such passions among Italians over their "mutilated victory" helped prepare the way for the triumph in 1922 of Mussolini's Fascists.

The Treaty of Trianon, delayed until 1920 by the Communist coup in Hungary, partitioned that ancient kingdom among its neighbours. Transylvania, including its minority of 1,300,000 Magyars, passed to Romania. The Banat of Temesvár (Timişoara) was divided between Romania and Yugoslavia, Sub-Carpathian Ruthenia passed to Czechoslovakia, and Croatia to Yugoslavia. All told, Hungary's territory shrank from 109,000 to 36,000 square miles. The armies of rump Austria and Hungary were limited to 35,000 men.

Partition of Hungary

The Treaty of Neuilly with Bulgaria marked yet another stage in the old struggles over Macedonia dating back to the Balkan wars and beyond. Bulgaria lost its western territories back to the kingdom of Serbs, Croats, and Slovenes (now Yugoslavia) and nearly all of Western Thrace to Greece, cutting the Bulgarians off from the Aegean. Their armed forces were likewise limited to 20,000 men. Austria, Hungary, and Bulgaria also accepted war guilt and reparations obligations, but these were later remitted in light of their economic weakness.

The settlement in east-central Europe was a generally well-meaning attempt to apply the principle of nationality under the worst imaginable circumstances. The new governments all faced aggrieved minorities, not to mention the onerous tasks of state-building—drafting constitutions, supporting currencies, raising armies and police—with no democratic tradition or financial resources beyond what they could borrow from the already strapped British and French. Austria in particular was a head without a body—over a quarter of its population lived in Vienna—yet was forbidden union with Germany. Hungary suffered violations of self-determination to an even greater degree and was bound to become a centre of aggressive revanche. Disputed borders, ethnic tensions, and local ambitions hampered economic and diplomatic cooperation among the successor states and would make them easy prey to a resurgent Germany, or Russia, or both.

The reorganization of the Middle East. The Treaty of Sèvres likewise dismembered the Ottoman Empire. Here again secret war-aims treaties reflected Allied ambitions in the Middle East, but Wilson was less willing to challenge them given his belief that the Arab peoples were not ready for self-rule. To avoid the tinge of imperialism the victors took control of the former Ottoman (and German) territories under "mandates" from the League: Class A mandates for those lands to be prepared for independence (Iraq, Transjordan, and Palestine, entrusted to Britain, and Syria and Lebanon to France); Class B mandates for those judged not ready for self-rule in the foreseeable future (Tanganyika to Britain, Cameroons and Togoland divided between Britain and France, and Rwanda-Urundi to Belgium); and Class C trust territories (German South West Africa to South Africa, Kaiser Wilhelms Land [New Guinea] to Australia, German Samoa to New Zealand, and the Mariana, Marshall, and Caroline islands to Japan).

The League mandate system

The victors also agreed, informally, that southern Anatolia was to be a French sphere of influence, while Italy received the Dodecanese Islands and a sphere of influence in eastern Anatolia. The Greek government of Venizélos, still a British client, occupied Smyrna (İzmir) and its hinterland, to the consternation of the Italians, who considered this poaching on their zone. Armenia was a special consideration because of its Christian population and the wartime deaths of hundreds of thousands (some claimed millions) of Armenians—through battle, mass murder, or forced deportation—at the hands of the Young Turks, who considered them a seditious element. Talk of an American mandate for Armenia gave way to independence. The collapse of the tsarist regime spared the Allies from having to award Constantinople and the Straits to Russia. The British proposed a League of Nations regime under U.S. administration for these areas, but Wilson refused this responsibility, while Indian Muslims protested any weakening of the Islāmic caliphate. So the status of Constantinople remained in abeyance, although the Straits were demilitarized and an Anglo-French-Italian commission regulated free passage. In August 1920 the helpless sultan's delegation signed the Treaty of Sèvres.

It was a dead letter. For Mustafa Kemal, the Turkish war hero, rallied his army in the interior and rebelled against the foreign influence in Anatolia and Constantinople. Unwilling to dispatch British armies, Lloyd George encouraged the Greeks to enforce the treaty instead. Indeed, Venizélos harboured a dream, the *megali idea,* of conquering the entire Turkish littoral and making the Aegean Sea a "Greek lake" as in ancient times. The Treaty of Sèvres, therefore, was the signal for the start of a Greco-Turkish War. By the end of 1920 the Greeks had fanned out from İzmir, occupied the western third of Anatolia, and were threatening the Turkish Nationalists' capital of Ankara. In March 1921 the British and French proposed a compromise that was rejected by the Turks, who nonetheless kept open diplomatic links in an effort to split the Allies. But as Kemal, later called Atatürk, put it: "We could not flatter ourselves that there was any hope of diplomatic success until we had driven the enemy out of our territory by force of arms." The tide of battle turned in August 1921, and the Greeks were forced to retreat precipitously through a hostile countryside. The French then made a separate peace with Ankara, settled their Syrian boundary, and withdrew support for the Anglo-Greek adventure. In March 1921 Turkey also signed a treaty of friendship with the new U.S.S.R. regulating the border between them and dooming the briefly independent Armenian and Transcaucasian republics.

The Greco-Turkish War

Another Allied offer (March 1922) could not tempt Kemal, who now had the upper hand. His summer attack routed the Greeks, who engaged in a panicky naval evacuation from İzmir which the Turks reentered on September 9. Kemal then turned north toward the Allied zone of occupation at Çanak (now Çanakkale) on the Dardanelles Strait. The French and Italians pulled out, and the British commissioner was authorized to open hostilities. At the last moment the Turks relented, and the Armistice of Mudanya (October 11) ended the fighting. Eight days later Lloyd George's Cabinet was forced to resign. A new peace conference produced the Treaty of Lausanne (July 24, 1923), which returned eastern Thrace to Turkey and recognized the Nationalist government in return for demilitarization of the Straits. The Treaty of Lausanne was to prove a durable solution to the old "Eastern question."

The Young Turk and Kemalist rebellions were models for other Islāmic revolts against Western imperialism. Persian nationalists had challenged the shah and Anglo-Russian influence before 1914 and flirted with the Young Turks (hence with Germany) during the war. By August 1919, however, British forces had contained both domestic protest and an ephemeral Bolshevik incursion and won a treaty from Tehrān providing for British administration of the Persian army, treasury, and railroads in return for evacuation of British troops. The Anglo-Persian Oil Company already controlled the oil-rich Persian Gulf. In June 1920, however, nationalist agitation resumed, forcing the shah to suspend the treaty. In Egypt, under British occupation since 1882 and a protectorate since 1914, the nationalist Wafd Party under Sa'd Zaghlūl Pasha, agitated for full independence on Wilsonian principles. Their three weeks' revolt of March 1919, suppressed by Anglo-Indian troops, gave way to passive resistance and bitter negotiations between Zaghlūl and the British high commissioner, Edmund Allenby. On Feb. 28, 1922, the British ended the protectorate and granted legislative power to an Egyptian assembly, though they retained military control of the Suez Canal.

Britain in Egypt

In India, where Britain controlled the fate of some 320,000,000 people with a mere 60,000 soldiers, 25,000 civil servants, and 50,000 residents, the war also sparked the first mass movement for independence. Out of hostility to Britain's Turkish policies, Islāmic leaders joined forces with Hindus in protest against the British raj. Edwin Montagu promised constitutional reform in July 1918, but the Indian National Congress deemed it insufficient. In 1919 famine, the return of Indian war veterans, and the inspiration of Mohandas Gandhi provoked a series of ever larger demonstrations until, on April 13, a nervous British general at Amritsar ordered his troops to open fire, and 379 Indians were killed. The amīr of Afghanistan, Amānollāh Khān, then sought to exploit the unrest in India to throw off the informal protectorate Britain enjoyed over his country. Parliament hastily approved the Montagu reforms, vetoed a campaign through the Khyber Pass, and so staved off a general uprising. But the Indian independence movement became a British preoccupation.

Other challenges to the empire arose from white minorities. After the Armistice, Lloyd George finally bowed to Irish demands for independence. After much negotiation and a threatened revolt in the northern counties, the compromise of December 1921 established the Irish Free State as a British dominion in the south while predominantly Protestant Northern Ireland remained in the United Kingdom. (The Sinn Féin nationalists continued to protest the treaty until, in 1937, Éire achieved complete independence, Ulster remaining British.) In South Africa the war propelled General Jan Smuts to international prominence and an influential role at the peace conference. South African expansionists clung to their own version of manifest destiny and dreamed of absorbing German South West Africa, Bechuanaland, and Rhodesia to forge a vast empire on the southern third of the continent. The British Colonial Office sternly resisted such ambitions. Yet the white minority of 1,500,000, dwarfed by a population of 5,000,000 blacks, 200,000 Indians, and 600,000 Chinese labourers, was itself split among Boer nationalists, "reconciled Boers," and British. The nationalists cited Wilsonian principles in a symbolic claim to restore the independent Transvaal and Orange republics in 1919 and remained a disaffected nationality within the Union of South Africa.

Ireland and South Africa

The non-European revolts, however—in Turkey, Persia, Egypt, India, and China—were the first expressions of what would become a major theme of the 20th century. Native elites, often educated in Europe and citing the anti-imperialist ideas of Wilson or Lenin, formed the first cadre of mass movements for decolonization. Often alienated from Europeans by their colour and customs, but no longer able to fit comfortably into their pre-modern societies, they became deracinated agitators for independence and modernization. Their growing numbers demonstrated that European imperialism, even as it reached its greatest extent through the 1919 treaties, must inevitably be a passing phenomenon.

THE NEW BALANCE IN EAST ASIA

The three Pacific powers. World War I also overthrew the power structure in East Asia and the Pacific. Before 1914 six imperial rivals had struggled for concessions on the East Asian coast. But the war eliminated Germany and Russia from colonial competition and weakened Britain and France, leaving the United States, Japan, and China in an uncomfortable triangular relationship that would persist until 1941.

Americans, largely ignorant of Asian realities, harboured a mix of attitudes before 1914. Contemptuous of what seemed to some of them, at least, as a barbaric and frozen Chinese culture, they nevertheless saw China as an unequalled opportunity for both Christian proselytizing and commercial exploitation. American investment in China in 1914 was only a quarter that of Japan and a 10th that of Britain, but moralism and manifest destiny both seemed to endow the United States with a special mission in China. On the other hand, Americans admired Japan for its mastery of modern technology but by the same token feared it as the primary obstacle to U.S. hopes for China. In 1899, a year after American acquisition of the Philippines and a year before the Boxer Rebellion, Secretary of State John Hay circulated his two "Open Door" notes imploring the Great Powers to eschew the dismemberment of China and to preserve free commercial access for all. The growing Japanese fleet worried American naval planners, who drafted at the time of the Russo-Japanese War the "Plan Orange" contingency for war with Japan. (They also conceded the impossibility of defending the Philippines against Japanese attack.)

The "Open Door"

The Chinese Revolution of 1911–12, inspired by the democratic principles of Sun Yat-sen (educated in Hawaii and British Hong Kong), expelled the Manchu dynasty and elevated the Nationalist Party, or Kuomintang (KMT), to power. But Sun quickly gave way in 1913 to General Yüan Shih-k'ai, whose failure to unify the giant land of 400,000,000 condemned China to a struggle among rival warlords that kept it in turmoil until at least 1928. Even as the Chinese revolted against foreign influence and exploitation, they remained nonetheless vulnerable to imperial predations or, conversely, dependent on foreign protection. In 1913 the Wilson administration entered office with a decidedly pro-Chinese leaning, and at the same time many Americans on the West Coast had become alarmed about the growing presence and success of enterprising Japanese immigrants and had begun to seek, in Washington and California, to legalize various forms of discrimination against them.

Japanese expansion during World War I only magnified American concern. After seizing Germany's Pacific islands and Chiao-chou Bay on the strategic Shantung Peninsula, Japan imposed on China the "Twenty-one Demands" (January 1915), claiming greatly expanded economic privileges and rights in Manchuria and Inner Mongolia (Sept. 3, 1916). After U.S. entry into the war, the Peking regime (but not the Nationalists in Canton) declared war on the Central Powers (Aug. 14, 1917) in hopes of defending its interests at the peace conference. The United States moved to end the embarrassment stemming from its cobelligerency with both China and Japan through the Lansing–Ishii Agreement of Nov. 2, 1917, in which Japan paid lip service to the Open Door while the United States recognized Japan's "special interests" in China. Wilson also sent troops to Vladivostok to monitor the Japanese intervention in Siberia.

The Paris Peace Conference exposed the two branches of Japanese expansionism, rooted in a bursting population and a booming industry in need of raw materials and markets. Delegate Saionji Kimmochi demanded the inclusion of a clause in the League of Nations Covenant proscribing racial discrimination, a principle that would have obliged the United States, Canada, and Australia to admit immigrants from Japan on equal terms with those of other nations. This was politically impossible for Wilson and Lloyd George to accept. The Japanese also demanded the rights formerly held by Germany at Chiao-chou, which Peking resisted vehemently. Finally Saionji agreed to drop the racial-equality plank in return for the granting of Japan's

Japan and the treaty

Chinese demands and threatened to reject the League of Nations if they were denied. Against Lansing's advice, Wilson acquiesced. Announcement of the terms provoked the anti-Western May Fourth Movement in China and caused it to be the only state that refused even to sign the Treaty of Versailles. Japan's triumph was an inauspicious precedent for diplomatic extortion by imperialist states from liberal states at the expense of helpless third parties.

The organization of power in the Pacific. In the United States, liberal internationalists, balance-of-power realists, Protestant churches with Chinese missions, and xenophobes all decried the cynical expansionism of Japan and what they took to be Wilson's capitulation. The Republican administration of Warren G. Harding in 1921 therefore determined to continue an ambitious naval construction plan dating from before the war and to pressure London to terminate the Anglo-Japanese Alliance dating from 1902. War debts gave the United States financial leverage over the British, as did American influence (based in a large Irish-American segment of the electorate) in the Irish question then reaching its climax. In June 1921 the British Commonwealth Conference bowed to this pressure and decided not to renew the alliance. This in turn confronted the Japanese with the prospect of a Britain aligned with Washington, not Tokyo, as well as a costly arms race against the world's two leading naval powers. A postwar business slump and worker unrest also suggested to Tokyo the wisdom of a tactical retreat.

Secretary of State Charles Evans Hughes invited the Great Powers to Washington, D.C., to forge a new order for East Asia and the Pacific. A Four-Power Pact negotiated at the conference (November 1921–February 1922) enjoined the United States, Japan, Britain, and France to respect each other's Pacific island dependencies for 10 years. A Nine-Power Pact obliged all parties to respect "the sovereignty, the independence, and the territorial and administrative integrity of the state of China" and the commercial Open Door. A separate Sino-Japanese agreement provided for Japanese evacuation of Shantung. In a Five-Power Treaty on naval armaments, Britain, the United States, Japan, France, and Italy agreed severally to maintain the naval balance of capital ships in the ratios 5:5:3:1.67:1.67 and agreed not to fortify their Pacific possessions. The latter three powers protested, but the United States frankly threatened to use its superior resources to dwarf the Japanese fleet, while France and Italy could not afford to compete with the British. France was also hoping for British support at this time in the struggle over German reparations (see below). Still, domestic displeasure with the treaties forced both the French and Japanese cabinets to resign.

The Five-Power naval treaty

Hughes's balance-of-power diplomacy for the Pacific seemed to reflect a realist turn in American statecraft in reaction to Wilson's idealism insofar as the United States flexed its muscle to compel the British and Japanese to keep hands off China and limit armaments. But in so doing the United States assumed responsibility as the balancer and container of Japanese power, for the naval agreement still left the Japanese fleet dominant in Asian waters. Moreover, the Japanese had clearly bowed to *force majeure* and, while resigned for the time being, would shrug off these constraints as soon as the Great Depression began to sap American resolve. In the long run, East Asian stability could come only through a strong and united China, for a weak and divided China represented constant temptation to a Japan bursting with strength, anxious for outlets, and resentful of Anglo-American containment.

THE POSTWAR GUILT QUESTION

Looking back on 1919–21 from the perspective of World War II, historians easily concluded that the Paris peacemakers had failed. In fact, debate over a "postwar guilt question" began even before the Big Three had completed their work. Anglo-American liberals felt betrayed by Wilson's failure to fashion a new diplomacy, while exponents of traditional diplomacy ridiculed Wilson's self-righteous intrusions. As Harold Nicolson put it: "We had hoped to call a new world into existence; we ended only by fouling the old." In other words, the peace amounted to a self-defeating mixture of contradictory ends or of tough ends and gentle means. Many Britons said the Treaty of Versailles was too harsh, would destroy Germany's economy and fragile new democracy, and would drive the bitter Germans to embrace militaristic revanche or Bolshevism. Many Frenchmen replied that the treaty was too mild, that a united Germany would resume its drive for hegemony, and that German democracy was sheeps' clothing put on for Wilson's benefit. Historians persuaded by the former argument often cast the peace conference as a morality play, with the messianic Wilson frustrated in his lofty mission by the atavistic Clemenceau. Those persuaded by the second argument speculate that the French plan for a permanent weakening of Germany might have made for a stabler Europe but for Wilson's and Lloyd George's moralizing, which, incidentally, served American and British interests at every turn. Clemenceau said: "Wilson speaks like Jesus Christ, but he operates like Lloyd George." And Lloyd George, when asked how he had done at Paris, said, "Not badly, considering I was seated between Jesus Christ and Napoleon."

The peace conference as morality play

Such caricatures skirt the facts that the war was won by the greatest coalition in history, that the peace could only take the form of a grand compromise, and that ideas are weapons. Once taking them up to great effect in the war on Germany, the Big Three could not cynically shrug them off any more than they could their constituents' interests, hopes, and fears. A purely Wilsonian peace, therefore, was never a possibility, nor was a purely power-political one on the order of the Congress of Vienna. Perhaps the new diplomacy was revealed as a sham or a disaster, as many professional diplomats claimed. Perhaps Wilson's moral insinuations only gave all parties grounds to depict the peace as illegitimate, one man's justice being always another's abomination. But it was still the old diplomacy that had spawned the hideous war in the first place. The pursuit of power without regard to justice, and the pursuit of justice without regard to power, were both doomed and dangerous occupations—such seemed to be the lesson of Versailles. The democratic states would spend the next 20 years searching in vain for a synthesis.

In the 1960s this portrait of the peace conference as a Manichaean duel gave way to new interpretations. New left historians depicted peacemaking after World War I as a conflict between social classes and ideologies, hence as the first episode in the Cold War. Arno J. Mayer wrote of 1919 as an "international civil war" between the "forces of movement" (Bolsheviks, Socialists, labour, and left-Wilsonians) and the "forces of order" (the Russian Whites, Allied governments, capitalists, and conservative power-politicians). While this thesis attracted overdue attention to the domestic political concerns of the Big Three, it imposed an equally dualistic set of categories, derived from the "primacy of domestic policy" paradigm, on the convoluted events of 1919. Perhaps it is most accurate to describe the Paris Peace Conference as the birthplace of all the major tactics, confrontational and conciliatory, for dealing with the Bolshevik phenomenon that have reappeared time and again to the present day. Prinkipo was the first attempt to get Communists and their opponents to substitute negotiations for force. Bullitt made the first stab at détente: direct negotiation of a modus vivendi. Churchill was the first "hawk," declaring that the only thing Communists understand is force. And Hoover and Nansen first acted on the theory that Communism is a social disease for which aid, trade, and higher standards of living were the cure.

The new left view

Thus, to say that the democratic, free-market statesmen at Paris were anti-Bolshevik is to state the obvious; to make this the wheel around which all else turned is to ignore the subtle. As Marshal Foch observed in counseling against exaggeration of the Bolshevik threat: "Revolution never crossed the frontiers of victory." That is, Communism was a product not just of privation, but of defeat, as in Russia, Germany, and Hungary. Perhaps, as Churchill thought, the Western democracies were not obsessed enough with the Bolshevik threat. They also understood it poorly, differed as to tactics, and were continually absorbed in other issues. Yet the failure to reintegrate Russia into the Eu-

ropean order was as poisonous to future stability as the German peace.

Whatever one's interpretation and assessment of the personalities and policies that collided at Paris, the overall settlement was surely doomed, not only because it sowed seeds of discord in almost every clause, but because all the Great Powers scurried from it at once. Germans denounced Versailles as a hypocritical *Diktat* and determined to resist it as much they were able. Italians fulminated against the "mutilated victory" given them by Wilson and then succumbed to Fascism in 1922. The Russian Communists, not privy to the settlements, denounced them as the workings of rapacious rival imperialisms. From the start, the Japanese ignored the League in favour of their imperial designs, and they soon held the Washington treaties to be unfair, confining, and dangerous to their economic health. The United States, of course, rejected Versailles and the League. Only Britain and France remained to make a success of Versailles, the League, and the chronically unstable successor states. But by 1920 British opinion was already turning against the treaty, and even the French, bitter over their "betrayal" at the hands of the United States and Britain, began to lose faith in the 1919 system. It was a new order that many yearned to overthrow and few were willing to defend.

The Versailles discords

A fragile stability, 1922–29

The 1920s are usually depicted as a bridge between the turmoil of the war and the turmoil of the 1930s, a brief truce in the "Thirty Years' War" of the 20th century. The disputes over execution of the Treaty of Versailles suggest a continuation of the Great War by other means, while the economic and security arrangements of mid-decade, and the era of good feeling they engendered, were flawed from their inception and collapsed with the onset of the Great Depression. Still, the postwar decade was Shakespeare's "time for frighted peace to pant." The conflicts of the early 1920s notwithstanding, weary populations had no stomach for war and demanded, in President Harding's words, a "return to normalcy," however fragile it might prove.

A BROKEN WORLD

The failure of democratic consensus. But what was normal in a world broken by total war? The pillars of the antebellum system—the balance of power, the non-interventionist state, the gold standard, and the free-market economy—lay in ruins and in any case reflected a faith in the natural play of political and economic forces that many Europeans had ceased to share. Wilsonians and Leninists blamed balance-of-power diplomacy for the war and fled from such normalcy. Technocrats, impressed by the productivity of regulated war economies, hoped to extend them into peacetime to promote recovery and dampen competition. Some economists and politicians even applauded the demise of the gold standard ("a barbarous relic," said Keynes) since inflation seemed the only means of financing jobs and veterans' pensions, thus stabilizing domestic societies. Finally, the free-market economy that had made high growth rates and technological dynamism seem normal from 1896 to 1914 was itself challenged by Socialists on the left and corporate interest groups on the right. In every case governments found it easier to try to shift the burden of reconstruction on to foreign powers, through reparations, loans, or inflation, than to impose taxes and austerity on quarreling social groups at home. It soon became clear that the effects of the war would continue to politicize economic relations within and between countries; that the needs of internal stability conflicted with the needs of international stability; that old dreams clashed with new realities, and new dreams with old realities.

The politicization of economic life

The search for a new stability. The lack of consensus on democracy itself also hampered the quest for a new stability. Wilson expected victory to mean a heyday of democracy in which the will of the people would oblige states to value peace and compromise. Instead, Communists and Fascists alike challenged democratic assumptions and elevated social class, race, and the state to the role Wilson reserved for the individual. In terms of the distribution of world power, the 1920s gave rise to a false normalcy, an Indian summer of European Great Power politics thanks to the peripheral roles played by the United States and the Soviet Union. In diplomacy, affairs of state came to be conducted increasingly by politicians meeting in grand conferences or at the League of Nations rather than by experts communicating with precision through written notes. Inevitably, style replaced substance at such meetings as prime ministers worried as much about their political image at home as about the actual issues at hand. The prime ministers of France and Britain held no less than 23 meetings from 1919 to 1923. As French Ambassador Camille Barrère complained, "Politicians have replaced diplomats at these conferences and seem to believe that nations conduct business like deputies in the Palais-Bourbon." But the trend was irreversible, for the crises of war and peace impressed on voters how much foreign policy affected their pocketbooks and daily lives, and they were sure to hold their elected officials responsible. Technological developments—the telephone, the wireless, and soon the airplane—also tended to reduce the role of professional ambassadors to that of messengers.

Behind the contradictory mixture of old and new in politics lay a profound cultural confusion. For the cultural shock of the Great War had turned modernist iconoclasm from the conceit of bohemian cliques into a new conventional wisdom. Respect for elders, for established authority, for "bourgeois" decency and restraint, died in the trenches. Faith in God and faith in reason, the two abiding fonts of Western culture, withered under the war's barbarizing bombardment, as did the belief in human progress born of the Enlightenment and the Industrial Revolution. Science and technology, those engines of progress, had only perfected an economy of death, and turned soldiers and civilians into mere cogs in the war machine. In the 1920s Einsteinian relativity, or a debased and popularized notion of it, replaced the comfortable order of the Newtonian universe, offering skeptics a pseudo-scientific justification for their rejection of absolute moral values. Popular Freudianism, depicting man as the victim of irrational, subconscious drives, seemed to describe the behaviour of 1914–18 better than the old Aristotelian psychology of man as a rational, moral creature. Nietzsche's transvaluation of values, implying that in a social Darwinist world compassion and charity were suicidal and force and mastery progressive, became a fad. To vulgar minds on the right and the left, Nietzsche's critique of modern mass civilization was an anthem for a politics of the violent deed. And while some artists despaired of man's fate in the crucible of the machine age, there were others, like the German Bauhaus school, who extolled steely power or, like the Italian Futurists, even modern war.

The rise of relativism

Oswald Spengler's 1918–22 best-seller *The Decline of the West* mourned the engulfing of *Kultur* by the cosmopolitan anthill of *Zivilisation* and argued that only a dictatorship could arrest the decline. Sociologist Max Weber hoped for charismatic leadership to overcome bureaucracy. Much painting, music, and film of the 1920s illustrated the theme of decline: Paul Klee's Cubist depiction of literally broken people and societies; George Grosz's looks beneath the veneer of respectable society to the rot underneath; the broken musical scales of Arnold Schoenberg; and the political drama of Bertolt Brecht. The intelligentsia of the 1920s leveled a comprehensive assault on bourgeois values, forms, and traditions. Tradition won scarcely more respect in the salons of Paris and London. The decade that was to have spawned a democratic diplomacy prepared the way instead for the totalitarian diplomacy of the 1930s.

To be sure, these were the years when European statesmen, in historian Charles Maier's words, set themselves the task of "recasting bourgeois Europe" and pioneered corporatist compromise among organized interest groups and bureaucracies when the increasingly polarized parliaments were unable to distribute the costs and benefits of reconstruction. By 1925 they had made a good show of it, as currencies and world trade stabilized and food, coal, and industrial production again reached 1913 levels. But the American economy alone boomed following the

postwar slump of 1920–21. Between 1922 and 1929, U.S. steel production climbed 70 percent, oil 156 percent, and automobiles 255 percent. Overall, national income soared 54 percent in those years; by 1929 the U.S. economy accounted for 44.8 percent of global industrial output, compared to 11.6 percent for Germany, 9.3 for Britain, 7.0 for France, and 4.6 for the Soviet Union. Yet the demobilization of American armed forces and United States refusal to make political-military engagements abroad meant that this mighty power existed in semi-isolation from the rest of the world. France and Britain, though engaged, lacked the resources and the will to run the risks inherent in trying to reintegrate Germany and Russia into the European order. A world with such disparities in the distribution of power and responsibility could not be returned to normal. It could only be given the appearance of normalcy by pasting paper constitutions, paper money, and paper treaties over the absence of common values, common interests, or a true balance of power.

The rise of the U.S. economy

REPARATIONS, SECURITY, AND THE GERMAN QUESTION

The continuing problem of Germany. The Great War failed to solve the German question. To be sure, Germany was exhausted and in the shackles of Versailles, but its strategic position actually improved in the war. Britain and France were at least as exhausted, Russia was in chaos and her boundary driven far to the east, and Italy was disaffected from her former allies, so that Germany's eastern and southern approaches now consisted of a broad ring of weak states. If and when Germany escaped Versailles, therefore, it might pose a greater threat to Europe than in 1914.

This danger obsessed postwar French leaders, but they quarreled among themselves over the proper response: strict execution of the Versailles treaty and perhaps even the breaking of German unity, or a Wilsonian policy of "moral disarmament" and reconciliation? In late 1919 the French electorate returned a staunchly conservative decision. The peace conference had not solved France's triple crisis of security, finance, and industrial reconstruction. Postwar French governments undertook to replace the abortive Anglo-American guarantee with an alliance system of Germany's neighbours. Belgium shrugged off neutrality, which had failed spectacularly to shelter it in 1914, and concluded a military alliance with France in September 1920. The Franco-Polish alliance (February 1921) and a Franco-Czechoslovak entente (January 1924) created an eastern counterweight to Germany. But these states, while wedded to the Versailles system, needed more protection than they offered. France could come to their aid only by a vigorous offensive against Germany from the west, which in turn required access to the bridgeheads over the Rhine. Thus, not only French security but that of east-central Europe as well depended on German disarmament and Allied occupation of the Rhineland.

France's new alliances

French finances were strained by the costs of rebuilding the devastated regions, the army, imperial obligations, and the refusal of the French chamber to accept sizable new taxes until Germany had paid reparations or France's war debts were annulled. To the extent that Germany reneged, France would face deficits imperiling its currency. As to industrial reconstruction, France depended on Germany for the coal needed to revive iron and steel production and at the same time was forced to countenance a cartel arrangement to escape Germany's economic competition.

Far from sympathizing with France's plight, the United States and Britain quickly withdrew from the Versailles treaty. Britain found itself in the midst of a postwar economic slump magnified by its wartime losses in ships and markets. Lloyd George had promised the veterans a land "fit for heroes," yet unemployment reached 17 percent in 1921. The war had accelerated the decline of the aging British industrial plant and the economy more generally. Unemployment never dipped much below 10 percent during the decade before the onset of the Great Depression, and in the early 1920s the pressure was on the British government to boost employment by reviving trade. Keynes argued persuasively that while Europe could never recover until the German economy took its natural place at the centre, virtually every clause of the treaty seemed designed to prevent that particular return to normalcy. To be sure, the British needed the reparations debt from Germany on the books to balance against their own war debts to the United States. But soon after the war Lloyd George came to favour German recovery in the interest of trade. The entente with France became strained as early as 1920 over the issues of reparations, Turkey, and the coal shortage of that year, from which Britain garnered windfall profits at the expense of the French.

Britain's economic woes

German politics and reparations. Germany, meanwhile, weathered both the leftist agitation of 1919 and the right-wing Kapp Putsch of March 1920. But elections showed a swing to the centre-right in German politics away from the parties that had voted to ratify Versailles. The insecure coalition cabinets of the early 1920s, therefore, found themselves with little room to maneuver on the foreign stage. They dared not rebel openly against Versailles, but dared not endorse fulfillment too eagerly in the face of domestic opinion. Nor could the weak Berlin government take forceful measures to end inflation, impose taxes, or regulate big business. The industrial magnates of the Ruhr thus acquired a virtual veto power over national policy by dint of their importance to the economy, a fact the embittered French did not fail to notice. German leaders themselves differed over how to win relief from the treaty. Army chief Hans von Seeckt and the eastern division of the foreign office thought in Bismarckian terms and favoured close ties with Russia, despite its obnoxious regime. But other economic and foreign policymakers preferred to rely on Britain and the United States to restrain France and revise the treaty. German diplomats soon synthesized these approaches, threatening closer ties with Moscow in order to win concessions from the West.

The Reparations Commission bickered throughout 1920 over the total sum to be demanded of Germany and its distribution among the Allies. At the Spa Conference (July 1920), France won 52 percent of German payments, Britain 22 percent, Italy 10, and Belgium 8. At the conferences of Hythe, Boulogne, and Brussels, France presented a total bill of 230,000,000,000 gold marks, although the British warned that this was far beyond Germany's capacity to pay. But when German foreign minister Walter Simons offered a mere 30,000,000,000 (Paris Conference, February 1921), French Premier Aristide Briand and Lloyd George made a show of force, seizing in March the Ruhr river ports of Düsseldorf, Duisburg, and Ruhrort, taking over the Rhenish customs offices, and declaring a 50 percent levy on German exports. Finally, on May 5, 1921, the London conference presented Berlin with a bill for 132,000,000,000 gold marks, to be paid in annuities of 2,000,000,000 plus 26 percent ad valorem of German exports. The Germans protested adamantly that this was "an injustice without equal." Historians have differed sharply as to whether the obligations were within the capacity of the German economy. But the May 1921 schedule was less harsh than it seemed, for the bill was divided into three series—A bonds totaling 12,000,000,000 marks, B bonds for 38,000,000,000, and the unlikely C bonds in the amount of 82,000,000,000. The latter would not even be issued until the first two series were paid and existed as much to balance against the Allies' debts to the United States as actually to be paid by Germany. Nevertheless, Chancellor Konstantin Fehrenbach resigned rather than accept this new *Diktat,* and his successor, Joseph Wirth, acquiesced only under threat of occupation of the Ruhr.

The German reparations bill

The "fulfillment" tactic adopted by Wirth and his foreign minister, Walther Rathenau, was to make a show of good faith to demonstrate that the reparations bill was truly beyond Germany's capacity. They were aided in this by the continuing deterioration of the paper mark. The prewar value of the mark was about 4.2 to the dollar. By the end of 1919 it reached 63, and after the first payment of 1,000,000,000 marks under the London plan, the mark fell to 262 to the dollar. The French argued that the inflation was purposeful, designed to feign bankruptcy while allowing Berlin to liquidate its internal debt and German industrialists like Hugo Stinnes and Fritz Thyssen to borrow, expand, and dump exports on the world market.

Recent research suggests, however, that the government did not fully understand the causes of the inflation even though it recognized its social utility in stimulating employment and permitting social expenditures. Of course, the reparations bill, while not the cause of inflation, was a strong disincentive to stabilization for Berlin could hardly plead bankruptcy if it boasted a strong currency, a balanced budget, and a healthy balance of payments. And insofar as the German government was dependent on those who benefited most from inflation—the industrialists—it was incapable of implementing austerity measures. This financial tangle might have been avoided by a program of reparations-in-kind whereby German firms delivered raw and finished goods directly to the Allies. The Seydoux Plan of 1920 and the Wiesbaden Accords of 1921 embraced such a mechanism, but the Ruhr magnates, delighted that the French might "choke on their iron" in the absence of German coal, and the British, fearful of any continental cartel, together torpedoed reparations-in-kind. By December 1921, Berlin was granted a moratorium.

Allied politics and reparations. At the Cannes Conference (January 1922) the Allies searched for common ground on reparations, a security pact, and Lloyd George's scheme for a grand economic conference including Soviet Russia. But the French chamber rebelled, and Briand was replaced as prime minister by the wartime president, Poincaré. A hard-headed lawyer from Lorraine, Poincaré was determined to relieve France's triple crisis without sacrificing its treaty rights. He approached London for a security pact, only to learn that the British were not willing to guarantee the Rhenish demilitarized zone and demanded French concessions on reparations in return. In June a conference of international bankers in Paris recommended loans to stabilize the German mark, but only if Germany were granted a long moratorium on reparations. (Meanwhile, the U.S. Congress created the World War Foreign Debt Commission to pressure the Allies to fund their war debts.) The grand economic conference promoted by Lloyd George was held at Genoa in April and May 1922 and was the first to bring German and Russian delegations together with the Allies on a status of equality. But the Soviets refused to recognize the tsarist regime's prewar debts and then shocked the Allies by signing the Treaty of Rapallo (April 16) with Germany, an innocuous document (providing for annulment of past claims and restoration of diplomatic relations) that nonetheless appeared to signal an unholy alliance between the two European outcasts. (Innocuous or not, Rathenau was assassinated by German rightists on June 24; Erzberger, signer of the Armistice, had also been murdered in 1921.) French representatives also bargained directly with the Ruhr magnates late in 1922, hoping for a coal-for-iron exchange and market-sharing, but the German price was evacuation of the Rhineland and substantial revision of the Treaty of Versailles. Meanwhile, the German mark tumbled to 7,500 to the dollar in December. Poincaré concluded that only force would break the deadlock. As he told the Belgians in July, "I will propose a short moratorium subject to guarantees. If England refuses I will act alone. The German industrialists conspire to destroy the mark. They hope to ruin France."

The Treaty of Rapallo

The new German Cabinet of Wilhelm Cuno made a desperate appeal to the United States. Secretary of State Hughes responded on December 29 with an offer to convene a committee of experts to study means of stabilizing the mark, but he held out no hope that the United States might relent on war debts. When the Reparations Commission declared that Germany had defaulted on its 1922 timber deliveries (Britain dissenting), Poincaré had his mandate to take sanctions. On Jan. 11, 1923, French and Belgian troops began to occupy the Ruhr. If the Germans submitted peacefully, the Ruhr would constitute a "productive guarantee," generating coal and receipts for France and giving her a valuable bargaining chip. If the Germans resisted, the French might take whatever measures seemed fit, up to and including political change in the Rhineland.

Occupation of the Ruhr

German workers protested the occupation of the Ruhr with an immense sitdown strike that proprietors and the government quickly joined. Berlin supported this passive resistance with unemployment relief that, in seeking to prove that the hated French could not "mine coal with bayonets," completed the destruction of the German currency. The railroads, mines, factories, and public services in the Ruhr and Rhineland ground to a halt. Poincaré steeled his will and dispatched French engineers and workers to revive the Rhine-Ruhr complex through the Inter-Allied Control Commission for Factories and Mines (MICUM) and a Franco-Belgian directorate for the railroads. The Allied Rhineland Commission (Britain dissenting) seized all executive, legislative, and judicial power in the occupied territories, expelled 16,000 uncooperative German officials (and more than 100,000 persons in all), and sequestered all German government property, energy resources, and transportation. France began covertly subsidizing separatist agitation. The Ruhr adventure thus became an economic war of attrition with stakes potentially as high as in a shooting war. If France retreated, the Treaty of Versailles was as good as dead; if Germany collapsed, the Rhineland might be lost.

The paper mark reached 4,000,000 to the dollar in August, and the Reich treasury was at the end of its tether. Business in non-occupied Germany was choking, and social unrest was spreading. Bavarian rightists called for war or separatism, while the Communist Party made gains in the cities. Gustav Stresemann, the conservative, business-oriented politician who replaced Cuno, finally ended passive resistance in September 1923 "to preserve the life of the nation and the state." But Poincaré, instead of naming his terms to Germany, apparently threw away the victory and accepted, after nine months' delay, Hughes's invitation to form a committee of experts. Poincaré's inaction baffled contemporaries, but in fact he had little to gain from dealing with Berlin. Only Britain and the United States could cancel France's war debts, stabilize the mark with loans to fund reparations, and offer security pacts or legitimize an autonomous Rhenish state, while only the Ruhr magnates could satisfy French industrial needs. So Poincaré ordered his Ruhr army commander to negotiate directly with Thyssen, Stinnes, Krupp, and their colleagues for the MICUM Accords (November 23) under which German industry went back to work, while he himself saw to the mandate of the international committee of experts.

Poincaré's plans misfired, however, for by the time the committee of experts began its deliberations at the turn of 1924, France's dearly purchased leverage had eroded and Germany had begun to recover. Troops expelled Communists from the governments of Saxony and Thuringia, a Communist putsch in Hamburg misfired, and Bavarian police quashed the Nazi putsch led by Adolf Hitler and Ludendorff. Hjalmar Schacht, recently appointed president of the Reichsbank, halted the inflation with a temporary currency called the Rentenmark, and on New Year's Day 1924 the president of the Bank of England, Montagu Norman, extended a 500,000,000 gold mark credit to back a new German mark. In October 1923, meanwhile, rowdy bands supported by the French occupation began to seize public buildings from Aachen to Speyer and to proclaim a Rhineland Republic. These separatists had no support from the population or from genuine Rhenish notables like the mayor of Cologne, Konrad Adenauer, and their actions only further discredited French policy in the eyes of Britain. By January the separatists had been driven out or murdered by fellow Germans. Finally, the French franc also succumbed to the pressure it had been under since the war. Poincaré tried austerity measures, but a new collapse in March forced him to borrow $89,000,000 from J.P. Morgan, Jr., of New York to stabilize the exchange rate. All these blows to France's position told in the report of the committee of experts under American Charles G. Dawes, released in April 1924. It called for a grand loan to Germany and the resumption of reparations payments, but made the latter contingent on French withdrawal from the Ruhr and restoration of German economic unity. Jacques Seydoux, an economist in France's foreign ministry, had predicted this outcome as early as November 1923: "There is no use hiding the fact that we have entered on the path of the 'financial reconstruction of Europe.' We will not deal with Germany as conqueror to vanquished; rather

German fiscal measures

the Germans and Frenchmen will sit on the same bench before the United States and other lending countries." On May 11, 1924, the French electorate defeated Poincaré in favour of the Cartel des Gauches (leftist coalition) under Édouard Herriot, who favoured a policy of accommodation with Germany.

THE AGREEMENTS OF MID-DECADE

Reparations agreements. Out of the exhaustion of France and Germany after the Ruhr struggle and the desire of American bankers and British diplomats to promote their reconciliation, the period 1924–26 finally produced agreements on reparations, security, and industrial cooperation. An interim reparations plan, the Dawes Plan, emerged from the London conference of July–August 1924. Expecting to join Ramsay MacDonald, Britain's first Labour prime minister, in Socialist brotherhood, Herriot instead found himself a supplicant whose bargaining points were few and feeble. France was obliged to evacuate the Ruhr (by August 1925), to end sanctions on the Rhine, and to promise never again to impose sanctions on Germany without the unanimous agreement of the Reparations Commission. The United States would lend $200,000,000 to Germany to "prime the pump," and Germany would pay from 1,000,000,000 to 2,500,000,000 marks in reparations for five years. The French government, by contrast, issued bonds worth 44,000,000,000 francs from 1919 to 1925 to finance reconstruction of its devastated regions. In the end, Germany received more money in loans than it ever paid in reparations, so the cost of repairing war damage was borne ultimately by the taxpayers, investors, and consumers of the Allied nations and the United States.

Loans to Germany

The influx of American capital through the Dawes Plan nevertheless broke the postwar spiral of inflation, default, and hostility and made possible a return to the gold standard. Germany stabilized its currency in 1924, Britain followed in 1925, and France did so in 1926 (officially in 1928). The smaller countries of Europe and Latin America, in turn, pegged their currencies against either the dollar, the pound, or the franc. Finally, the French government agreed in the Mellon–Berenger Accords (April 20, 1926) to fund its war debts at the favourable rates offered by the United States. The new gold standard and the cycle of international transfers, however, depended on a continuous flow of American capital. Should that flow ever cease, the normalcy so painfully achieved would quickly be imperiled.

Security and the League of Nations. In the realm of security, France had still achieved nothing. Of course, the Versailles restrictions on German armaments were still in force, as well as France's rear alliance system, but in striving for collective security the French suffered a series of disappointments. League of Nations Assembly Resolution XIV of September 1922 endorsed the disarmament commission's recommendation for a treaty on collective security. The Czech delegation, led by Edvard Beneš, quickly rose to a position of leadership in security matters, with the support of French and British proponents of the League such as Lord Robert Cecil, whose Draft Treaty of Mutual Assistance came under discussion in 1923. Beneš rightly criticized the Draft Treaty for requiring unanimity on the League Council to declare sanctions against an aggressor. For only in rare cases was the accused party's guilt obvious to all, as the 1914 case itself illustrated. Beneš also wanted some mechanism for pacific settlement of disputes before resort to arms. But more telling was opposition to the concept of collective security in British opinion. Canada, Australia, and other dominions especially opposed an instrument that might involve them in war over some obscure conflict in eastern Europe. In July 1924 London rejected the Draft Treaty.

The Draft Treaty of Mutual Assistance

Beneš submitted an improved Geneva Protocol (or Protocol for the Pacific Settlement of International Disputes) in October. Under the protocol, states would agree to submit all disputes to the Permanent Court of International Justice, any state refusing arbitration was ipso facto the aggressor, and the League Council could impose binding sanctions by a two-thirds majority. France enthusiastically supported the Geneva Protocol, but British Foreign Secretary Austen Chamberlain rejected it in March 1925.

Herriot had made it known that France would not proceed with the first partial evacuation of the Rhineland, scheduled for January 1925, unless he could show the French people something in the way of security. Chamberlain suggested to Stresemann in February 1925 that the Germans themselves reassure France through a regional security pact. Stresemann took up the idea, seeing in it a way to head off a bilateral Anglo-French alliance. Herriot's government fell in April, but Aristide Briand stayed on as foreign minister to carry through negotiations. Stresemann and Briand met and embraced at Locarno, swore to put the war behind them once and for all, and signed five treaties (Oct. 16, 1925) designed to pacify postwar Europe. Locarno seemed truly a second peace conference and was greeted with cheers and relief in world capitals. The main treaty, the Rhineland Pact, enjoined France, Belgium, and Germany to recognize the boundaries established by the Treaty of Versailles as inviolate and never again to resort to force in an attempt to change them. Moreover, the pact was guaranteed by Britain and Italy, who pledged to resist whatever country violated the demilitarized Rhineland. Germany also signed arbitration agreements with France, Belgium, Poland, and Czechoslovakia, agreeing to submit future disputes to international authority.

The Locarno Pact

Locarno seemed a giant step forward. Rather than a *Diktat,* it was a voluntary recognition by Germany of the 1919 borders in the West. Britain had been brought in to guarantee, not only France, but demilitarization of the Rhineland. Italy's adherence was a bonus. Germany had negotiated as an equal and looked forward to further abridgement of Versailles restrictions. Above all, Briand hoped, Locarno was the start of the "moral disarmament" of Germany. But some contemporaries, and many historians, criticized Locarno for being an incomplete system, as dangerous as it was seductive. By way of granting German equality, Britain had guaranteed Germany against French attack as much as France against Germany. "England," said Poincaré, "becomes the arbiter of Franco-German relations." To be sure, France still promised to help Poland and Czechoslovakia in case of German attack, but after Locarno, Prague and Warsaw discounted the French commitment. What was more, Locarno all but invited German revisionism in the east by explicitly providing, not for recognition, but arbitration on Germany's eastern borders. Changes in French military policy also bode ill for eastern Europe. Since 1919, Foch and Pétain had quarreled over whether to adopt an offensive or defensive contingency plan for the French army. In the wake of Locarno the Pétain faction won, and France began to design an imposing system of concrete fortresses along the border with Germany. This Maginot Line (after Minister of War André Maginot) was not meant to preclude offensive action by the French army but was in effect (in Foch's words) a "Great Wall of China" that would breed a false sense of security and weaken France's will to take the offensive on behalf of her eastern allies.

Finally, the aftermath of the Ruhr episode provided French and German industry with a chance to normalize their relations. The evacuation of the Ruhr restored Germany's coal leverage, and Berlin recovered tariff sovereignty in 1925 under the Treaty of Versailles, but the French inflation of 1924–26 shifted the export price advantage from Germany to France. Long and complicated four-way negotiations (French and German public and private sectors) produced a Franco-German steel syndicate in 1926 providing for coal-for-iron exchanges and an international committee to fix production quotas quarterly. The latter awarded France a 31 percent share compared to 43 percent for Germany, a marked improvement over the 1 to 4 ratio France had suffered before 1914. Franco-German commercial treaties followed in 1926–27.

Franco-German coal and iron agreements

The agreements of mid-decade ended the bickering and uncertainty of the immediate postwar years and made Germany a partner in the new Europe. But in every case they replaced French rights under Versailles with voluntary agreements dependent both on Anglo-American support and German goodwill.

ITALY AND EAST-CENTRAL EUROPE

Fascism and Italian reality. The peoples of east-central Europe enjoyed a degree of freedom in the 1920s unique in their history. But the power vacuum in the region resulting from the temporary impotence of Germany and Russia pulled in other Great Powers—chiefly Mussolini's Italy and France—seeking respectively to revise or uphold the 1919 order.

Fascism was the most striking political novelty of the interwar years. Fascism defied precise definition. In practice it was an anti-Marxist, antiliberal, and antidemocratic mass movement that aped Communist methods, extolled the leadership principle and a "corporatist" organization of society, and showed both modern and anti-modern tendencies. But the three states universally acknowledged to be Fascist in the 1930s—Italy, Germany, and Japan—were most similar in their foreign, rather than their domestic, ideology and policy. All embraced extreme nationalism and a theory of competition among nations and races that justified their revolts—as "proletarian nations"—against the international order of 1919. In this sense, Fascism can be understood as the antithesis of Wilsonianism rather than of Leninism.

In the first decade of Mussolini's rule, changes in Italian diplomacy were more stylistic than substantive. But recent historiography argues that this decade of relatively good behaviour was a function of the continuing constraints on Italian ambitions rather than moderation in Fascist goals. Mussolini proclaimed upon taking power that "treaties are not eternal, are not irremediable," and declared loudly and often his determination to restore Italian grandeur. This would be accomplished by revision of the "mutilated victory," by the transformation of the Mediterranean into an Italian *mare nostrum,* and by the creation of "a new Roman Empire" through expansion and conquest in Africa and the Balkans. Such reveries reflected not only Mussolini's native grandiloquence but also Italy's relative poverty and surplus rural population and need for markets and raw materials secure from the competition of more developed powers. In this sense, Italy was a sort of weak Japan. And like the Japanese, Italians bristled at the tendency of the Great Powers to treat them, in Mussolini's words, "as another Portugal." Still, Fascist bluster seemed safely unmatched in actions, and London in particular was pleased with the tendency of the Fascist foreign minister Dino Grandi to "take refuge on rainy days under the ample and capacious mantle of England" in traditional Italian fashion. More than once Grandi dissuaded Il Duce from provocative actions, taking care not to offend his vanity. The Italian navy's inferiority to the British and French, and the army's need for reorganization, also suggested prudence.

The "new Roman Empire"

Fascist diplomacy. Italian diplomacy in the 1920s, therefore, was a mix of bombast and caution. At the Lausanne Conference, Mussolini dramatically stopped his train to oblige Poincaré and Curzon to come to him. He made Italy the first Western power to offer a trade agreement and recognition to the Bolsheviks and was proud of Italy's role in the League (though he considered it "an academic organization") and as a guarantor of the Locarno Pact. In the Mediterranean, Mussolini protested French rule in Tunis and asserted for Italy a moral claim to the province. But he satisfied his thirst for action against weaker opponents. He broke the Regina Agreement with the Sanūsī tribesmen of Libya, which had limited Italian occupation to the coast, and by 1928 completed Italy's conquest of that poor and weak country.

Italy's main sphere of activity was the Balkans. When an Italian general surveying the border of a Greek-speaking district of Albania was killed in August 1923, Mussolini ordered a naval squadron to bombard the Greek isle of Corfu. The League of Nations awarded Italy an indemnity, but not the island. In January 1924, Wilson's Free State of Fiume disappeared when Yugoslav Premier Nikola Pašić granted Italian annexation in the Treaty of Rome. Diplomatic attempts to regularize relations between Belgrade and Rome, however, could not overcome Yugoslavia's suspicion of Italian ambitions in Albania. In 1924 a coup d'état, ostensibly backed by Belgrade, elevated the Muslim Ahmed Bey Zogu in Tiranë. Once in power, however, Ahmed Zogu looked to Italy. The Tiranë Pact (Nov. 27, 1926) provided Italian economic aid and was followed by a military alliance in 1927 and finally a convention (July 1, 1928) declaring Albania a virtual protectorate of Italy. Ahmed Zogu then assumed the title of King Zog I.

Italy and Albania

To the north, Italian diplomacy aimed at countering French influence among the successor states. In 1920 the French even courted Hungary and toyed with the idea of resurrecting a Danubian Confederation, but when the deposed Habsburg King Charles appeared in Hungary in March 1921, Allied protests and a Czech ultimatum forced him back into exile. Hungarian revisionism, however, motivated Beneš to unite those states that owed their existence to the Treaty of Trianon. A Czech–Yugoslav alliance (Aug. 14, 1920), Czech–Romanian alliance (April 23, 1921), and Romanian–Yugoslav alliance (June 7, 1921) together formed what was known as the Little Entente. When Charles tried again in October to claim his throne in Budapest, the Little Entente threatened invasion. While France had not midwived the combination, it associated strongly with the successor states through Franco–Czech (Oct. 16, 1925), Franco–Romanian (June 10, 1926), and Franco–Yugoslav (Nov. 11, 1927) military alliances. The latter implied that France would side with Belgrade against Rome in case of war and exacerbated the strained relations between France and Italy.

Mussolini had more luck in the defeated states of central Europe, Austria and Hungary. But in the former case, Italy was not siding with the revisionists. In return for financial aid to end its own hyperinflation, Austria had promised the League of Nations in 1922 that it would not seek *Anschluss* with Germany. Mussolini proclaimed in May 1925 that he, too, would never tolerate the *Anschluss* but set out to curry favour with the Austrian government. An Italo-Hungarian commercial treaty (Sept. 5, 1925), a friendship treaty (April 5, 1927) moving Hungary "into the sphere of Italian interests," and a rapprochement with Bulgaria in 1930 completed Italy's alignments with the states defeated in the war. Hungary in particular attracted Mussolini's sympathy. But as long as the combined will of the Little Entente, backed by France, opposed revisionism, Italy alone could force no alterations. On the other hand, military or economic cooperation among the congeries of states in east-central Europe also proved impossible. Czech–Polish rivalry continued, however illogical, and after Piłsudski's coup d'état in Poland in 1926 even the internationalist Beneš sought to steer German revisionism against Poland rather than Austria and the Danubian basin. The Little Entente and French alliances, therefore, amounted to a fair-weather system that would collapse in the first storm.

THE INVENTION OF SOVIET FOREIGN POLICY

Lenin's diplomacy. In November 1920 Lenin surprised Western observers and his fellow Bolsheviks alike by declaring that "we have entered a new period in which we have . . . won the right to our international existence in the network of capitalist states." By 1921, the generally accepted turning point in Soviet policy, Bolshevism had made the transition from a revolutionary movement to a functioning state. The Civil War was won, the New Economic Policy ended the brutal "War Communism" and restored a measure of free market activity to peasants, and the Soviet government was organized along traditional ministerial lines (though subject to the dictates of the Communist Party). Russia was ready—needed—to pursue traditional relations with foreign powers in search of capital, trade, and technology for reconstruction. The emergence of what Stalin called "Socialism in one country" therefore obliged the Soviets to invent out of whole cloth a "Communist" foreign policy.

The New Economic Policy

That invention took shape as a two-track approach whereby Russia (from 1922 the U.S.S.R.) would on the one hand continue to operate as the centre of world revolution, dedicated to the overthrow of the capitalist powers, and yet conduct an apparently regular existence as a nation-state courting recognition and assistance from those same powers. The first track was the responsibility of the

Comintern (Third International) under Grigory Zinovyev and Karl Radek; the second, of the Narkomindel (foreign commissariat) directed from 1920 to 1930 by the timid and cultured prewar nobleman, Georgy Chicherin. The Comintern enjoyed direct access to the Politburo, whereas the Narkomindel had no voice even in the Central Committee until 1925. In practice, however, the foreign policy interests of the U.S.S.R. dominated even the Comintern to such an extent that other Communist parties were not factions in their own country's politics so much as Soviet fifth columns operating abroad. When subversive activity flagged, diplomacy came to the fore; when diplomacy was unfruitful, revolution was emphasized. The goal was not to encourage "peace" or "progressive reform" in the West, but solely to enhance Soviet power. Thus Lenin instructed Comintern parties "to unmask not only open social patriotism but also the falseness and hypocrisy of social pacifism"; in other words, to do all that was possible to undermine Moscow's rivals on the left as well as on the right through the infiltration and subversion of Western labour unions, armed forces, newspapers, and schools. Yet Moscow readily ignored or confounded the efforts of local Communists when diplomatic opportunities with foreign countries seemed promising. The scent of betrayal this caused made mandatory the secrecy, discipline, and purges demanded of Communist parties abroad.

At the third congress of the Comintern in 1921 even Trotsky, the impassioned advocate of world revolution, admitted that the struggle of the proletariat in other countries was slackening. At that time the mutiny of Russian sailors at Kronshtadt and widespread famine in Russia impelled the party to concentrate on consolidating its power at home and reviving the economy. The Soviets, therefore, turned to the capitalists who, Lenin jeered, would "sell the rope to their own hangmen" in search of profits. Indeed, Western leaders, especially Lloyd George, viewed the vast Russian market as a kind of panacea for Western industrial stagnation and unemployment. But he and others misunderstood the nature of the Soviet state. Private property, commercial law, and hard currency no longer existed in Russia; one did business, not in a market, but on terms laid down by a state monopoly. What was more, by 1928 the whole point of trade was to allow the Soviet economy to catch up to the West in the shortest possible time and thus achieve complete self-sufficiency. It was, in George Kennan's words, a "trade to end all trade."

Domestic imperatives

The Anglo-Russian commercial pact of March 1921 and secret contacts with German military and civilian agents were the first Soviet openings to the Great Powers. Both culminated the following year in the Genoa Conference, where the Soviet representatives appeared, to the relief of their counterparts, in striped pants and on good behaviour. Indeed, having seized power as the minority faction of a minority party, the Bolsheviks sought legitimacy abroad as the most adamant sticklers for etiquette and legalism. But the Western powers insisted on an end to Communist propaganda and recognition of the tsarist debts as prerequisites to trade. Chicherin countered with a fanciful claim for reparations stemming from the Allied interventions, at the same time denying that Moscow bore any responsibility for the doings of the Comintern. As Theodore von Laue has written, "To ask the Soviet regime . . . to refrain from making use of its revolutionary tools was as futile as to ask the British Empire to scrap its fleet." Instead, a German-Russian knot was tied in the Treaty of Rapallo, whereby the U.S.S.R. was able to take advantage of Germany's bitterness over Versailles to split the capitalist powers. Trade and recognition were not the only consequences of Rapallo; in its wake began a decade of clandestine German military research on Russian soil.

Upon the occupation of the Ruhr the Soviets declared solidarity with the Berlin government. By August 1923, however, with Stresemann seeking negotiations with France and German society disintegrating, revolutionary opportunism again took precedence. The Politburo went so far as to designate personnel for a German Communist government, and Zinovyev gave German Communists the signal to stage a putsch in Hamburg. When it proved a fiasco, the Soviets returned to their Rapallo diplomacy with Berlin. The political victories of the leftists MacDonald in Britain and Herriot in France then prompted recognition of the Soviet government by Britain (Feb. 1, 1924), Italy (February 7), France (October 28), and most other European states. Later in 1924, however, publication during the British electoral campaign of the infamous (and probably forged) "Zinovyev letter" ordering Communists to disrupt the British army created a sensation. British police also suspected Communists of subversive activities during the bitter General Strike of 1926 and launched the "Arcos raid" on the Soviet trade delegation in London in May 1927. Anglo-Soviet relations did not resume until 1930.

Recognition of the Soviet Union

Stalin's diplomacy. Lenin's incapacity and death (Jan. 21, 1924) triggered a protracted struggle for power between Trotsky and Joseph Stalin. In foreign policy their conflict seemed one of an emphasis on aiding the European peoples "in the struggle against their oppressors" (Trotsky) versus an emphasis on "building Socialism in one country" (Stalin). But that was largely a caricature meant to discredit Trotsky as an "adventurer." During the intraparty struggle, however, Soviet foreign policy drifted. The "partial stabilization of capitalism in the West" through the Dawes Plan and the Locarno treaties was a rude setback for Moscow. When Germany later joined the League of Nations, the Soviet press warned Germany against this "false step" into "this wasp's nest of international intrigue, where political sharpers and thieving diplomatists play with marked cards, strangle weak nations, and organize war against the U.S.S.R." But the Germans were not about to throw away their Russian card. Negotiations to expand the Rapallo accord produced the Treaty of Berlin (April 24, 1926) by which Germany pledged neutrality in any conflict between the U.S.S.R. and a third power, including the League of Nations. Germany also provided a 300,000,000-mark credit and in the late 1920s accounted for 29 percent of Soviet foreign trade.

From 1921 on, the Politburo judged Asia to be the region that offered the best hope for Socialist expansion, although this required collaboration with "bourgeois nationalists." The Bolsheviks suppressed their own subject nationalities at the first opportunity, yet declared their solidarity with all peoples resisting Western imperialism. In 1920 they paid homage to the "great and famous Amīr Amānollāh" in cementing relations with the new Afghan leader, and they were the first to sign treaties with Nationalist Turkey. In September 1920 the Comintern sponsored a conference of "the peoples of the East" at Baku. Zinovyev and Radek presided over a contentious lot of Central Asian delegates, whose own quarrels, of which the Armenian-Turkish was the most vitriolic, made a mockery of any notion of regional or political solidarity. Thereafter, Soviet Asian activity went underground, alternately aiding Communists against nationalists like Reza Khan and Mustafa Kemal, and aiding nationalists against the European powers.

Central Asian overtures

The centrepiece of Soviet designs in Asia could only be China, whose liberation Lenin viewed in 1923 as "an essential stage in the victory of socialism in the world." In 1919 and 1920 the Narkomindel made much of its revolutionary sympathy for China by renouncing the rights acquired by tsarist Russia in its concessionary treaties. But soon the Soviets were sending troops into Outer Mongolia, allegedly at the request of local Communists, and concluding their own treaty with Peking (May 31, 1924) that granted the U.S.S.R. a virtual protectorate over Outer Mongolia—its first satellite—and continued ownership of the Chinese Eastern Railway in Manchuria.

The political disintegration of China, and their own devious tactics, inevitably complicated Soviet policy. While pursuing superficially correct relations with Peking, the Politburo placed its future hopes on the Canton-based Nationalists (KMT), whose members were impressed by the Bolsheviks' example of how to seize and master a vast undeveloped country. In 1922 the Comintern directed Chinese Communists to enroll in the KMT even as Adolf Yoffe renounced all Soviet intentions of importing Marxism into China. The Communist presence in the KMT grew rapidly until, after Sun Yat-sen's death in March 1925, Comintern agent Mikhail Borodin became the main strategist for the KMT. Still, the Soviets

were uncertain how to proceed. In March 1926, Trotsky counseled caution lest precipitate attacks on foreign interests in China impel the imperialists—including Japan—into anti-Soviet action. Indeed, Stalin did his best to woo Tokyo, noting that Japanese nationalism had great anti-Western potential.

On March 20, 1926, Chiang Kai-shek turned the tables with a coup that elevated him within the KMT and landed many Communists in prison. Ignoring the outrage of the Chinese Communists, Borodin remained in Chiang's good graces, whereupon Chiang staged the northern expedition in which he greatly expanded KMT power with the help of Communist organizations in the countryside. But Borodin also advised leftist KMT members to leave the south for a new base in the Wu-han cities to escape Chiang's immediate control. This "Left KMT" or "Wu-han Body" was to steer the KMT in a Communist direction and eventually seize control. The Soviet Party Congress in January 1927 even declared China the "second home" of world revolution, and Stalin confided to a Moscow audience that Chiang's forces were "to be utilized to the end, squeezed out like a lemon, and then thrown away." But Chiang preempted again by ordering a bloody purge of Shanghai Communists on April 12–13, 1927. Trotsky blamed Stalin's lack of faith in revolutionary zeal for the debacle, declaring that he should have unleashed the Communists sooner. Instead, the Left KMT eroded, many of its former adherents going over to Chiang. With the party thus fractured, Stalin changed his mind and ordered an armed revolt by Communists against the KMT. This, too, ended in carnage, and by mid-1928 only scattered bands (one under Mao Zedong) remained to take to the hills.

Stalin and China's revolution

Stalin's triumph at home and failure in China ended the formative era of Soviet foreign policy. The Politburo had expelled Zinovyev, Radek, and Trotsky by October 1926; the Party Congress condemned all deviation from the Stalinist line in December 1927; and Trotsky went into exile in January 1929. Thenceforth Soviet foreign policy and the Comintern line reflected the will of one man. Communist parties abroad likewise purged all but Stalinists and reorganized in rigid imitation of the U.S.S.R.'s ruthless dictatorship. The Sixth Party Congress (summer 1928) anathematized social democracy in the strongest terms ever and strengthened its call for subversive activities against democratic institutions. Above all, Stalin declared after an ephemeral war scare of 1926 that the era of peaceful coexistence with capitalism was coming to an end and ordered vigorous measures to prepare the U.S.S.R. for war.

First Five-Year Plan

The New Economic Policy gave way to the First Five-Year Plan (Oct. 1, 1928) for collectivization of agriculture and rapid industrialization, which condemned millions of peasants to expropriation, starvation, or exile to Siberia, but enabled the regime to sell wheat abroad to pay for industrial goods. Stalin imported entire factories from the United States, France, Italy, and Germany as the basis for the Soviet steel, automotive, aviation, tire, oil, and gas industries. In 1927 he launched the first of the show trials of industrial "wreckers" who had allegedly conspired with reactionaries and foreign agents, and in 1929 he purged all those—the "Right Opposition"—who questioned the Five-Year Plan.

The Bolsheviks interpreted their survival and consolidation in the 1920s as confirmation of their reading of the objective forces of history. In fact, Soviet foreign policy could boast of few successes. It was the Allied defeat of Germany in 1918 and the Red Army's military prowess that permitted the revolution to survive; the Versailles restraints on Germany and *cordon sanitaire* in eastern Europe that sheltered Russia from the West as much as it sheltered Europe from Bolshevism; American pressure on Japan that restored Vladivostok to the U.S.S.R.; Anglo-French recognition that opened much of the world to Soviet trade; and Western technology that enabled Stalin to hope for rapid economic modernization. The link with Germany was a Soviet achievement, but even it had a double edge, for it helped Germany to prepare for its own remilitarization. Of course, Stalin was ultimately right that a crisis of capitalism and new round of imperialism and war were just around the corner, but in part it was Comintern assaults on Western liberals and Socialists that helped to undermine the fragile stability of the 1920s.

THE UNITED STATES, BRITAIN, AND WORLD MARKETS

U.S. leverage in world markets. The economic dislocations and technological advances of the war, the relative rise of American power, and territorial changes in the colonial world all made stabilization of world markets a pressing issue in the 1920s. The resolution of this issue was chiefly the responsibility of the two economies that bestrode the world: the United States and the British Empire. Their interests diverged in many regions. At the Allied Economic Conference of 1916 the British and French had projected a postwar Allied cartel to control raw materials, while in 1918 the British drafted plans for excluding American capital from the British Empire. At the peace conference Wilson and Lloyd George engaged in backstage debate over the allocation of United States and Allied shipping with an eye to expanding their respective countries' share of world trade. On the heels of the merchant shipping rivalry came naval competition that culminated in the breaking of the Anglo-Japanese Alliance and the Washington Treaty limitations. Finally, the war debts raised the issue of whether Britain would seek a "debtors' cartel" with the French to defy Wall Street, or join the United States in a "creditors' cartel." At stake in the U.S.–British disputes was their relative global power in coming decades.

Traditional American protectionism triumphed after the electoral victory of the Republicans. The Fordney–McCumber Tariff (September 1922) was the highest in U.S. history and angered the Europeans, whose efforts to acquire dollars through exports were hampered even as the United States demanded payment of war debts. In raw materials policy, however, the United States upheld the Open Door. Secretary of Commerce Herbert Hoover rejected both statist economic competition that bred war and laissez-faire competition that bred cycles of boom and bust. Instead, he advocated formal cooperation among firms of various nations to stabilize the price and supply of commodities, raise living standards, and yet avoid the waste and oppression of regulatory bureaucracies. This "third alternative" would create "a new economic system, based neither on the capitalism of Adam Smith nor upon the Socialism of Karl Marx." By dint of leverage and persuasion, the United States gradually brought Britain around to this model of informal entente. By late 1922 London bankers also took the American position on war debts, and the two nations also cooperated in such new areas as transoceanic cables and radio. Of surpassing importance for national power in the mechanized 20th century, however, was oil.

The Fordney–McCumber Tariff

After the Great War, known oil reserves outside the industrial powers themselves were concentrated in the British mandates of the Middle East, Persia, the Dutch East Indies, and Venezuela. The Royal Dutch/Shell Group and Anglo-Persian Oil Company dominated oil exploration and production in Asia, but increasingly they confronted revolutionary nationalism, Bolshevik agitation (in Persia), and U.S. opposition to imperialism. When the British and French agreed at San Remo (1920) to coordinate their oil policies in the Middle East, the American Petroleum Institute and the U.S. State Department protested any exclusion of U.S. firms. What was more, the United States invoked the Mineral Lands Leasing Act of 1920 against the Dutch, denying them access to American reserves in retaliation for Shell's monopoly in the East Indies. In 1921, Hoover and Secretary of State Hughes encouraged seven private firms to form an American Group, led by Standard Oil of New Jersey, to seek a share of Mesopotamian oil reserves, while State Department expert Arthur Millspaugh outlined a plan for worldwide Anglo-American reciprocity. The British, fearing American retaliation and anxious to have help against native rebellions, granted the American Group a 20 percent share of the rich Mesopotamian fields. In 1922 a similar arrangement spawned the Perso-American Petroleum Company. In 1925 the Iranian nationalist Reza Khan, inspired in part by the Kemalist revolt in Turkey, seized power and had

Sharing Middle Eastern oil

himself proclaimed Reza Shah Pahlavi, but he was unable to play the British and Americans off against each other. Oil politics and nationalism in the Middle East, therefore, presaged events of the post-1945 era. (Another anticipation occurred in Palestine, where the Balfour Declaration encouraged thousands of Jewish Zionists to immigrate, leading to bloody clashes with Palestinian Arabs in 1921 and 1929.) Reciprocity also triumphed in U.S.–Dutch oil diplomacy, and Standard Oil of New Jersey acquired a 28 percent share in the East Indies by 1939.

U.S. leverage in Latin-American affairs. In Venezuela and Central America the situation was the reverse. During the war the State Department endorsed all-American oil concessions, but, in accordance with the principle of reciprocity, Hughes instructed his Latin-American ambassadors in 1921 to respect foreign interests. Latin America in general became far more of an American sphere of influence during the war than ever before owing to the growth of American commerce at Britain's expense. Central American governments now relied on New York banks to manage their public finance rather than those of London and Paris, while the U.S. share of Latin-American trade totaled 32 percent, double Britain's share, though British capital still predominated in the economics of Argentina, Brazil, and Chile.

Ever since the 17 republics of mainland Latin America emerged from the wreck of the Spanish Empire in the early 19th century, North Americans had viewed them with a mixture of condescension and contempt that focused on their alien culture, racial mix, unstable politics, and moribund economies. The Western Hemisphere seemed a natural sphere of U.S. influence, and this view had been institutionalized in the Monroe Doctrine of 1823 warning European states that any attempt to "extend their system" to the Americas would be viewed as evidence of an unfriendly disposition toward the United States itself. On the one hand, the doctrine seemed to underscore republican familiarity, as suggested by references to "our sister republics," "our good neighbors," our "southern brethren." On the other hand, the United States later used the doctrine to justify paternalism and intervention. This posed a quandary for the Latin Americans, since a United States strong enough to protect them from Europe was also strong enough to pose a threat itself. When Secretary of State James G. Blaine hosted the first Pan-American Conference in 1889, Argentina proposed the Calvo Doctrine asking all parties to renounce special privileges in other states. The United States refused.

U.S. paternalism in Latin America

After the Spanish–American War in 1898 the United States strengthened its power in the Caribbean by annexing Puerto Rico, declaring Cuba a virtual protectorate in the Platt Amendment (1901), and manipulating Colombia into granting independence to Panama (1904), which in turn invited the United States to build and control the Panama Canal. In the Roosevelt Corollary (1904) to the Monroe Doctrine the United States assumed "an international police power" in cases where Latin-American insolvency might lead to European intervention. Such "dollar diplomacy" was used to justify—and probably made inevitable—the later "gunboat diplomacy" of U.S. military intervention in Santo Domingo, Nicaragua, and Haiti. In his first term President Wilson also became embroiled in the Mexican Revolution. An affront to U.S. sailors led to his bombardment of Veracruz (1914), and border raids by Pancho Villa prompted a U.S. expedition into northern Mexico (1916). The Mexican Constitution of 1917 then granted to the state all subsoil resources to prevent their exploitation by U.S. firms. Such revolutionary efforts to nationalize resources, however, only meant that they went undeveloped or were exploited at home by corrupt officials, while the United States retaliated by cutting off loans and trade. The Latin-American dilemma of weakness and disunity in proximity to a mighty and united power was thus insoluble through unilateral efforts or a Pan-American movement dominated by Washington.

Wilson's proposed League of Nations seemed to offer Latin America a means of circumventing U.S. influence. But the United States inserted Article 21 to the effect that "Nothing in this Covenant shall be deemed to affect the validity of international engagements, such as treaties of arbitration or regional understandings like the Monroe Doctrine." Secretary of State Hughes later defended U.S. behaviour by candidly questioning the ability of some Latin-American states to maintain public order, sound finance, and the rule of law. When the Chaco dispute between Bolivia and Paraguay erupted into war, League of Nations President Briand offered his personal good offices, but he refused to assert League authority for fear of irritating the United States. In the end, the Pan-American Commission of Inquiry assumed jurisdiction.

Latin America and the League of Nations

Latin-American protests grew in volume, especially in 1926, when a Mexican-supported leftist rebellion in Nicaragua prompted U.S. Secretary of State Frank B. Kellogg to report to the Senate Foreign Relations Committee on "Bolshevist Aims and Policies in Mexico and Latin America." But intervention by United States marines in Nicaragua only paved the way for the dictatorial regime of the Somozas. At the Pan-American Conference of 1928, rivalry between Argentina and Brazil and the Chaco contestants, and the caution of other states, precluded their presenting a united Latin-American front. But the U.S. administrations of the decade did labour to improve the American image. The Clark Amendment of 1928 repudiated the Roosevelt Corollary, while Hoover toured 10 Latin-American nations after his election as president and repudiated the "big brother" role. In the 1920s, therefore, the United States continued to squeeze out European influence in Latin America but was itself moving slowly toward the "Good Neighbor" policy of the 1930s.

THE LOCARNO ERA AND THE DREAM OF DISARMAMENT

The Locarno treaties promised a new era of reconciliation that seemed fulfilled in the mid-to-late 1920s as the European and world economies recovered and the German electorate turned its back on extremists of the right and left. Locarno had also anticipated Germany's entry into the League. But the prospect of expanding the League Council kicked off an indelicate scramble for Council seats as Britain supported Spain, France supported Poland, and Brazil insisted that it represent Latin America (angering the Argentines). Sweden and Czechoslovakia helped to break the deadlock by magnanimously sacrificing their seats, although Brazil in the end quit the League. Finally, on Sept. 8, 1927, Stresemann led a German delegation into the halls of Geneva, pledging that Germany's steadfast will was to labour for freedom, peace, and unity. Briand, by now the statesman most associated with "the spirit of Geneva," replied in like terms: "No more blood, no more cannon, no more machine-guns! . . . Let our countries sacrifice their amour-propre for the sake of the peace of the world." The same month, Stresemann tried to capitalize on the goodwill during an interview with Briand at Thoiry. He suggested a 1,500,000,000-mark advance on German reparations payments (to ease the French fiscal crisis then nearing its climax) in return for immediate evacuation of the last two Rhineland zones. The French chamber would likely have rejected such a concession, and in any case Poincaré, again in power, stabilized the franc soon after.

The very goodwill expressed at Geneva—and removal of the Interallied Military Control Commission from Germany in January 1927—prompted London and Washington to ask why the French (despite their pleas of penury when war debts were discussed) still maintained the largest army in Europe. France clung firm to its belief in military deterrence of Germany, even when isolated in the League of Nations Disarmament Preparatory Commission, but the German demand for equality of treatment under the League Charter impressed the Anglo-Americans. To avert U.S. suspicions, Briand enlisted Secretary Kellogg's participation in promoting a treaty by which all nations might "renounce the resort to war as an instrument of national policy." This Kellogg–Briand Pact, signed on Aug. 27, 1928, and eventually subscribed to by virtually the entire world, marked the high point of postwar faith in paper treaties and irenic promises.

The Kellogg–Briand Pact

On July 3, 1928, Chancellor Hermann Müller (a Social Democrat) and Stresemann decided to force the pace of Versailles revisionism by claiming Germany's moral right

to early evacuation of the Rhineland. In return they offered a definitive reparations settlement to replace the temporary Dawes Plan. The French were obliged to consider the offer—a revival of Thoiry—because the French chamber had refused to ratify the 1926 agreement with the United States on war debts on the ground that it did not yet know what could be expected of Germany in reparations. So another committee of experts under another American, Owen D. Young, drafted a plan that was approved at the Hague Conference of August 1929. The Young Plan projected German annuities lasting until 1989. In return, the Allies abolished the Reparations Commission, restored German financial independence, and promised evacuation of the Rhineland by 1930, five years ahead of the Versailles schedule.

The Young Plan

Why did Briand and even Poincaré make so many concessions between 1925 and 1929? Briand, of course, had sincerely hoped for Germany's "moral disarmament," and both concluded that France's treaty rights had become a wasting asset. Better to sacrifice them now in return for concessions and goodwill, since they would expire sooner or later anyway. But Stresemann was far from accepting the status quo. His policy of accommodation was designed to achieve the gradual abolition of the Versailles strictures until Germany recovered its prewar freedom of action, at which time he could set out to restore its prewar boundaries as well. For instance, he showed no interest in an "Eastern Locarno" ensuring the boundaries of the successor states. That is not to say, however, that Stresemann anticipated the use of force or the revival of Germany's extreme war aims.

As the decade of the 1920s came to a close, most Europeans expected prosperity and harmony to continue. Briand even went so far as to propose in 1929 that France and Germany explore virtual political integration in a European union, asking only that Germany confirm her 1919 boundaries as immutable. But Stresemann died suddenly on Oct. 3, 1929, and three weeks later the New York stock market crashed. In the storms to come, the need for firm, material guarantees of security would be greater than ever. But on June 30, 1930, in accordance with the Young Plan, the last Allied troops departed the German Rhineland for home.

The origins of World War II, 1929–39

The 1930s were a decade of unmitigated crisis culminating in the outbreak of a second total war. The treaties and settlements of the first postwar era collapsed with shocking suddenness under the impact of the Great Depression and the aggressive revisionism of Japan, Italy, and Germany. By 1933 hardly one stone stood on another of the economic structures raised in the 1920s. By 1935 Adolf Hitler's Nazi regime had torn up the Treaty of Versailles and by 1936 the Locarno treaties as well. Armed conflict began in Manchuria in 1931 and spread to Abyssinia in 1935, Spain in 1936, China in 1937, Europe in 1939, and the United States and U.S.S.R. in 1941.

The context in which this collapse occurred was an "economic blizzard" that enervated the democracies and energized the dictatorial regimes. Western intellectuals and many common citizens lost faith in democracy and free-market economics, while widespread pacifism, isolationism, and the earnest desire to avoid the mistakes of 1914 left Western leaders without the will or the means to defend the 1919 order. This combination of demoralized publics, stricken institutions, and uninspired leadership led historian Pierre Renouvin to describe the 1930s simply as "*la décadence.*"

The militant authoritarian states on the other hand—Italy, Japan, and (after 1933) Germany—seemed only to wax stronger and more dynamic. The Depression did not cause the rise of the Third Reich or the bellicose ideologies of the German, Italian, and Japanese governments (all of which pre-dated the 1930s), but it did create the conditions for the Nazi seizure of power and provide the opportunity and excuse for Fascist empire-building. Hitler and Mussolini aspired to total control of their domestic societies, in part for the purpose of girding their nations for wars of conquest which they saw, in turn, as necessary for revolutionary transformation at home. This ideological meshing of foreign and domestic policy rendered the Fascist leaders wholly enigmatic to the democratic statesmen of Britain and France, whose attempts to accommodate rather than resist the Fascist states only made inevitable the war they longed to avoid.

The authoritarian states

THE ECONOMIC BLIZZARD

Political consequences of the Depression. The debate over the origins of the Great Depression and the reasons for its severity and length is highly political, given the implications for the validity of theories of free market, regulated, and planned economies, and of monetary and fiscal policy. It is usually dated from the New York stock-market crash of October 1929, which choked the domestic and international flow of credit and severely damaged global trade and production. Wall Street prices fell from an index of 216 to 145 in a month, stabilized in early 1930, then continued downward to a bottom of 34 in 1932. Industrial production fell nearly 20 percent in 1930. Unlike previous swings in the business cycle, this financial panic did not eventuate in the expected period of readjustment, but rather defied all governmental and private efforts to restore prosperity for years until it seemed to a great many that the system itself was breaking down.

Mutual recriminations flew across the Atlantic. Americans blamed the Europeans for the reparations tangle, for pegging their currencies too high upon the return to gold, and for misuse of the American loans of the 1920s. Europeans blamed the United States for its insistence on repayment of war debts, high tariffs, and the unfettered speculation leading to the stock-market crash. Certainly all of these factors contributed. More tangibly, however, a sudden contraction of international credit in June 1928 made an international emergency likely. Since the Dawes Plan of 1924, Europe had depended for capital and liquidity on the availability of American loans, but increasingly American investors were flocking to the stock market with their savings, and new capital issues for foreign account in the United States dropped 78 percent, from $530,000,000 to $119,000,000. Loans to Germany collapsed from $200,000,000 in the first half of 1928 to $77,000,000 in the second half and to $29,500,000 for the entire year of 1929. A world crisis was also brewing in basic commodities, a market in which prices had been depressed throughout the decade. Mechanization of agriculture stimulated overproduction, and Soviet dumping of wheat on the world market to earn foreign exchange for the First Five-Year Plan compounded the problem.

The Smoot–Hawley Tariff, the highest in U.S. history, became law on June 17, 1930. Conceived and passed by the House of Representatives in 1929, it may well have contributed to the loss of confidence on Wall Street and signaled American unwillingness to play the role of leader in the world economy. Other countries retaliated with similarly protective tariffs, with the result that the total volume of world trade spiraled downward from a monthly average of $2,900,000,000 in 1929 to less than $1,000,000,000 by 1933. The credit squeeze, bank failures, deflation, and loss of exports forced production down and unemployment up in all industrial nations. In January 1930 the United States had 3,000,000 idle workers, and by 1932 there were more than 13,000,000. In Britain 22 percent of the adult male work force lacked jobs, while in Germany unemployment peaked in 1932 at 6,000,000. All told, some 30,000,000 people were out of work in the industrial countries in 1932.

The Smoot–Hawley Tariff

The Depression naturally magnified European bitterness over the continuing international obligations, but the weakest link in the financial chain was Austria, whose central bank, the Creditanstalt, was on the verge of bankruptcy. In March 1931, Stresemann's successor as German foreign minister, Julius Curtius, signed an agreement with Vienna for a German–Austrian customs union, but French objections to what they saw as a first step toward the dreaded *Anschluss* provoked a run on the Creditanstalt and forced Berlin and Vienna to renounce the union on September 3. The panic then spread to Germany, rendering the Reichs-

he
loover
lorato-
um

bank unable to meet its obligations under the Young Plan. President Hoover responded on June 20, 1931, with a proposal for a one-year moratorium on all intergovernmental debts. Short of a general recovery or global agreement on the restoration of trade, however, the moratorium could only be a stopgap. Instead, every country fled toward policies of protection, self-sufficiency, and the creation of regional economic blocs in hopes of isolating itself from the world collapse. On Sept. 21, 1931, the Bank of England left the gold standard, and the pound sterling promptly lost 28 percent of its value, undermining the solvency of countries in eastern Europe and South America. In October a national coalition government formed to take emergency measures. The Ottawa Imperial Economic Conference of 1932 gave birth to the British Commonwealth of Nations and a system of imperial preferences, signaling the end of Britain's 86-year-old policy of free trade.

The Lausanne Conference of June–July 1932 took up the question of what should be done after the Hoover Moratorium. Even the French granted the impossibility of further German payments and agreed to make an end of reparations in return for a final German transfer of 3,000,000,000 marks (which was never made). The United States, however, still insisted that the war debts be honoured, whereupon the French parliament willfully defaulted, damaging Franco-American relations.

Failures of the League. Panicky retrenchment and disunity also rendered the Western powers incapable of responding to the first violation of the postwar territorial settlements. On Sept. 10, 1931, Viscount Cecil assured the League of Nations that "there has scarcely ever been a period in the world's history when war seemed less likely than it does at the present." Just eight days later officers of Japan's Kwantung Army staged an explosion on the South Manchurian Railway to serve as pretext for military adventure. Since 1928, China had seemed to be achieving an elusive unity under Chiang Kai-shek's Nationalists (KMT), now based in Nanking. While the KMT's consolidation of power seemed likely to keep Soviet and Japanese ambitions in check, resurgent Chinese nationalism also posed a threat to British and other foreign interests on the mainland. By the end of 1928, Chiang was demanding the return of leased territories and an end to extraterritoriality in the foreign concessions. On the other hand, the KMT was still split by factions, banditry continued widespread, the Communists were increasingly well-organized in remote Kiangsi, and in the spring of 1931 a rival government sprang up in Canton. To these problems were added economic depression and disastrous floods that took hundreds of thousands of lives.

Japan, meanwhile, suffered rudely from the Depression because of her dependence on trade, her ill-timed return to the gold standard in 1930, and a Chinese boycott of Japanese goods. But social turmoil only increased the appeal of those who saw in foreign expansion a solution to Japan's economic problems. This interweaving of foreign and domestic policy, propelled by a rabid nationalism, a powerful military-industrial complex, hatred of the prevailing distribution of world power, and the raising of a racialist banner (in this case, anti-white) to justify expansion, all bear comparison to European Fascism. When the parliamentary government in Tokyo divided as to how to confront this complex of crises, the Kwantung Army acted on its own. Manchuria, rich in raw materials, was a prospective sponge for Japanese emigration (250,000 Japanese already resided there) and the gateway to China proper. The Japanese public greeted the conquest with wild enthusiasm.

Japan's
nvasion of
Manchuria

China appealed at once to the League of Nations, which called for Japanese withdrawal in a resolution of October 24. But neither the British nor U.S. Asiatic fleets (the latter comprising no battleships and just one cruiser) afforded their governments (obsessed in any case with domestic economic problems) the option of intervention. The tide of Japanese nationalism would have prevented Tokyo from bowing to Western pressure in any case. In December the League Council appointed an investigatory commission under Lord Lytton, while the United States contented itself with propounding the Stimson Doctrine, by which Washington merely refused to recognize changes born of aggression. Unperturbed, the Japanese prompted local collaborationists to proclaim, on Feb. 18, 1932, an independent state of Manchukuo, in effect a Japanese protectorate. The Lytton Commission reported in October, scolding the Chinese for provocations but condemning Japan for using excessive force. Lytton recommended evacuation of Manchuria but privately believed that Japan had "bitten off more than she can chew" and would ultimately withdraw of its own accord. In March 1933, Japan announced its withdrawal instead from the League of Nations, which had been tested and found impotent, at least in East Asia.

The League also failed to advance the cause of disarmament in the first years of the Depression. The London Naval Conference of 1930 proposed an extension of the 1922 Washington ratios for naval tonnage, but this time France and Italy refused to accept the inferior status assigned to them. In land armaments, the policies of the powers were by now fixed and predictable. Britain and the United States deplored "wasteful" military spending, especially by France, while reparations and war debts went unpaid. But even Herriot and Briand refused to disband the French army without additional security guarantees that the British were unwilling to tender. Fascist Italy, despite its financial distress, was unlikely to take disarmament seriously, while Germany, looking for foreign-policy triumphs to bolster the struggling Republic, demanded equality of treatment: Either France must disarm, or Germany must be allowed to expand its army. The League Council nonetheless summoned delegates from 60 nations to a grand Disarmament Conference at Geneva beginning in February 1932. When Germany failed to achieve satisfaction by the July adjournment it withdrew from the negotiations. France, Britain, and the United States devised various formulas to break the deadlock, including a No Force Declaration (Dec. 11, 1932), abjuring the use of force to resolve disputes, and a five-power (including Italy) promise to grant German equality "in a system providing security for all nations." On the strength of these the Disarmament Conference resumed in February 1933. By then, however, Adolf Hitler was chancellor of the German Reich.

Geneva
Disar-
mament
Conference
of 1932

A common impression of Herbert Hoover is that he was passive in the face of the Depression and isolationist in foreign policy. The truth was almost the reverse, and in the 1932 campaign his Democratic opponent, Franklin Roosevelt, was the more traditional in economic policy and isolationist in foreign policy. Indeed, Hoover bequeathed to his successor two bold initiatives meant to restore international cooperation in matters of trade, currency, and security: the London Economic Conference and the Geneva Disarmament Conference. The former convened in June 1933 in hopes of restoring the gold standard but was undermined by President Roosevelt's suspension of the gold convertibility of the dollar and his acerbic message rejecting the conference's labours on July 3. At home, Roosevelt proposed the series of government actions known as the New Deal in an effort to restore U.S. productivity, in isolation, if need be, from the rest of the world. The Disarmament Conference came to a similar end. In March, Ramsay MacDonald proposed the gradual reduction of the French army from half a million to 200,000 men and the doubling of Germany's Versailles army to the same figure, accompanied by international verification. But a secret German decree of April 4 created a National Defense Council to coordinate rearmament on a massive scale. Clearly the German demand for equality was a ploy to wreck the conference and serve as pretext for unilateral rearmament.

Germany
moves to
rearm

Negotiations were delayed by a sudden initiative from Mussolini in March calling for a pact among Germany, Italy, France, and Britain to grant Germany equality, revise the peace treaties, and establish a four-power directorate to resolve international disputes. Mussolini appears to have wanted to downgrade the League in favour of a Concert of Europe, enhancing Italian prestige and perhaps gaining colonial concessions in return for reassuring the Western powers. The French watered down the plan until

the Four-Power Pact signed in Rome on June 7 was a mass of anodyne generalities. Any prospect that the new Nazi regime might be drawn to collective security disappeared on Oct. 14, 1933, when Hitler denounced the unfair treatment accorded Germany at Geneva and announced its withdrawal from the League of Nations.

THE RISE OF HITLER AND FALL OF VERSAILLES

Failure of the German Republic. The origins of the Nazi Third Reich must be sought not only in the appeal of Hitler and his party but also in the weakness of the Weimar Republic. Under the republic, Germany boasted the most democratic constitution in the world, yet the fragmentation of German politics made government by majority a difficult proposition. Many Germans identified the republic with the despised Treaty of Versailles and, like the Japanese, concluded that the 1920s policy of peaceful cooperation with the West had failed. What was more, the republic seemed incapable of curing the Depression or dampening the appeal of the Communists. In the end, it self-destructed. The first Depression-era elections, in September 1930, reflected the electorate's flight from the moderate centrist parties: Communists won 77 seats in the Reichstag, while the Nazi delegation rose from 12 to 107. Chancellor Heinrich Brüning, unable to command a majority, governed by emergency decree of the aged president, Paul von Hindenburg.

The National Socialist German Workers' Party (Nazis) exploited the resentment and fear stemming from Versailles and the Depression. Its platform was a clever, if contradictory, mixture of socialism, corporatism, and virulent assertion in foreign policy. The Nazis outdid the Communists in forming paramilitary street gangs to intimidate opponents and create an image of irresistible strength, but unlike the Communists, who implied that war veterans had been dupes of capitalist imperialism, the Nazis honoured the Great War as a time when the German *Volk* had been united as never before. The army had been "stabbed in the back" by defeatists, they claimed, and those who signed the Armistice and Versailles had been criminals; worse, international capitalists, Socialists, and Jews continued to conspire against the German people. Under Nazism alone, they insisted, could Germans again unify under *ein Reich, ein Volk, ein Führer* and get on with the task of combating Germany's real enemies. This amalgam of fervent nationalism and rhetorical socialism, not to mention the charismatic spell of Hitler's oratory and the hypnotic pomp of Nazi rallies, was psychologically more appealing than flaccid liberalism or divisive class struggle. In any case, the Communists (on orders from Moscow) turned to help the Nazis paralyze democratic procedure in Germany in the expectation of seizing power themselves.

The "stab in the back"

Brüning resigned in May 1932, and the July elections returned 230 Nazi delegates. After two short-lived rightist cabinets foundered, Hindenburg appointed Hitler chancellor on Jan. 30, 1933. The president, parliamentary conservatives, and the army all apparently expected that the inexperienced, lower-class demagogue would submit to their guidance. Instead, Hitler secured dictatorial powers from the Reichstag and proceeded to establish, by marginally legal means, a totalitarian state. Within two years the regime had outlawed all other political parties and coopted or intimidated all institutions that competed with it for popular loyalty, including the German states, labour unions, press and radio, universities, bureaucracies, courts, and churches. Only the army and foreign office remained in the hands of traditional elites. But this fact, and Hitler's own caution at the start, allowed Western observers fatally to misperceive Nazi foreign policy as simply a continuation of Weimar revisionism.

Adolf Hitler recounted in *Mein Kampf,* the autobiographical harangue written in prison after his abortive putsch of 1923, that he saw himself as that rare individual, the "programmatic thinker and the politician become one." Hitler distilled his *Weltanschauung* from the social Darwinism, anti-Semitism, and racialist anthropology current in prewar Vienna. Where Marx had reduced all of history to struggles among social classes, in which revolution was the engine of progress and the dictatorship of the proletariat the culmination, Hitler reduced history to struggle among biologic races, in which war was the engine of progress and Aryan hegemony the culmination. The enemies of the Germans, indeed of history itself, were internationalists who warred against the purity and race-consciousness of peoples—they were the capitalists, the Socialists, the pacifists, the liberals, all of whom Hitler identified with the Jews. This condemnation of Jews as a racial group made Nazism more dangerous than earlier forms of religious or economic anti-Semitism that had long been prevalent throughout Europe. For if the Jews, as Hitler thought, were like bacteria poisoning the bloodstream of the Aryan race, the only solution was their extermination. Nazism, in short, was the twisted product of a secular, scientific age of history.

Nazi anti-Semitism

Hitler's worldview dictated a unity of foreign and domestic policies based on total control and militarization at home, war and conquest abroad. In *Mein Kampf* he ridiculed the Weimar politicians and their "bourgeois" dreams of restoring the Germany of 1914. Rather, the German *Volk* could never achieve their destiny without Lebensraum ("living space") to support a vastly increased German population and form the basis for world power. Lebensraum, wrote Hitler in *Mein Kampf,* was to be found in the Ukraine and intermediate lands of eastern Europe. This "heartland" of the Eurasian continent (so named by the geopoliticians Sir Halford Mackinder and Karl Haushofer) was especially suited for conquest since it was occupied, in Hitler's mind, by Slavic *Untermenschen* (subhumans) and ruled from the centre of the Jewish-Bolshevik conspiracy in Moscow. By 1933 Hitler had apparently imagined a step-by-step plan for the realization of his goals. The first step was to rearm, thereby restoring complete freedom of maneuver to Germany. The next step was to achieve Lebensraum in alliance with Italy and with the sufferance of Britain. This greater Reich could then serve, in the distant third step, as a base for world dominion and the purification of a "master race." In practice, Hitler proved willing to adapt to circumstances, seize opportunities, or follow the wanderings of intuition. Sooner or later politics must give way to war, but because Hitler did not articulate his ultimate fantasies to the German voters or establishment, his actions and rhetoric seemed to imply only restoration, if not of the Germany of 1914, then the Germany of 1918, after Brest-Litovsk. In fact, his program was potentially without limits.

European responses to Nazism. European reaction to the rise of Nazism was cautious, but not at first overtly hostile. The Four-Power Pact and a concordat with the Vatican (July 20, 1933), negotiated by the Catholic Franz von Papen, conferred a certain legitimacy on the Nazi regime. (Hitler sought to end Vatican support for the Catholic Centre Party while he proceeded to subordinate the churches and to corrupt Christianity into a state-centred form of neo-paganism. Pope Pius XI, like every other European statesmen after him, thought that he could appease and moderate the Nazis.) On Jan. 26, 1934, Hitler shocked all parties by signing a nonaggression pact with Poland. This bit of duplicity neutralized France's primary ally in the east while helping to secure Germany over the dangerous years of rearmament. The new Polish foreign minister, Józef Beck, was in turn responding to the dilemma of Poland's central position between Germany and the U.S.S.R. He hoped to preserve a balance in his relations with the two giant neighbours (Poland signed a three-year pact with Moscow in July 1932) but feared the Soviets (from whom Poland had grabbed so much territory in 1921) more than the still-weak Germans. The pact with Germany was meant to run for 10 years.

The German-Polish pact

France was the nation most concerned by the Nazi threat and most able to take vigorous action. But fear of another war, the defeatist mood dating from the failure of the Ruhr occupation, the passivity engendered by the Maginot Line (due for completion in just five years), and domestic strife exacerbated by the Depression and the Stavisky scandal of 1933 all served to hamstring French foreign policy. As in the Weimar Republic, Communists and monarchists or Fascist groups like the Croix de Feu and Action Fran-

çaise battled in the streets. In February 1934 a crowd of war veterans and rightists stormed the parliament, and the Édouard Daladier Cabinet was forced to resign to head off a coup d'état. The new foreign minister, Louis Barthou, had been a friend of Poincaré and made a final effort to shore up France's security system in Europe: "All these League of Nations fancies—I'd soon put an end to them if I were in power. . . . It's alliances that count." But alliances with whom? The French Left was adamantly opposed to cooperation with Fascist Italy, the Right despised cooperation with the Communist Soviet Union. Britain as always eschewed commitments, while Poland had come to terms with Germany. Nevertheless, the moment seemed opportune; both Italy and the U.S.S.R. now made clear their opposition to Hitler and desire to embrace collective security.

To be sure, Mussolini was gratified by the triumph of the man he liked to consider his younger protégé, Hitler, but he also understood that Italy fared best while playing off France and Germany, and he feared German expansion into the Danubian basin. In September 1933 he made Italian support for Austrian Chancellor Engelbert Dollfuss conditional on the latter's establishment of an Italian-style Fascist regime. In June 1934 Mussolini and Hitler met for the first time, and in their confused conversation (there was no interpreter present) Mussolini understood the Führer to say that he had no desire for *Anschluss*. Yet, a month later, Austrian Nazis arranged a putsch in which Dollfuss was murdered. Mussolini responded with a threat of force (quite likely a bluff) on the Brenner Pass and thereby saved Austrian independence. Kurt von Schuschnigg, a pro-Italian Fascist, took over in Vienna. In Paris and London it seemed that Mussolini was one leader with the will and might to stand up to Hitler.

The first Hitler-Mussolini meeting

Stalin, meanwhile, had repented of the equanimity with which he had witnessed the Nazi seizure of power. Before 1933, Germany and the U.S.S.R. had collaborated, and Soviet trade had been a rare boon to the German economy in the last years of the Weimar Republic. Still, the behaviour of German Communists contributed to the collapse of parliamentarism, and now Hitler had shown that he, too, knew how to crush dissent and master a nation. The Communist line shifted in 1934–35 from condemnation of social democracy, collective security, and Western militarism to collaboration with other anti-Fascist forces in "Popular Fronts," alliance systems, and rearmament. The United States and the U.S.S.R. established diplomatic relations for the first time in November 1933, and in September 1934 the Soviets joined the League of Nations, where Maksim Litvinov became a loud proponent of collective security against Fascist revisionism.

Thus, Barthou's plan for reviving the wartime alliance and arranging an "Eastern Locarno" began to seem plausible—even after Oct. 9, 1934, when Barthou and King Alexander of Yugoslavia were shot dead in Marseille by an agent of Croatian terrorists. The new French foreign minister, the rightist Pierre Laval, was especially friendly to Rome. The Laval–Mussolini agreements of Jan. 7, 1935, declared France's disinterest in the fate of Abyssinia in implicit exchange for Italian support of Austria. Mussolini took this to mean that he had French support for his plan to conquer that independent African country. Just six days later the strength of German nationalism was resoundingly displayed in the Saar plebiscite. The small, coal-rich Saarland, detached from Germany for 15 years under the Treaty of Versailles, was populated by miners of Catholic or social democratic loyalty. They knew what fate awaited their churches and labour unions in the Third Reich, and yet 90 percent voted for union with Germany. Then, on March 16, Hitler used the extension of French military service to two years and the Franco-Soviet negotiations as pretexts for tearing up the disarmament clauses of Versailles, restoring the military draft, and beginning an open buildup of Germany's land, air, and sea forces.

The Saar plebiscite

In the wake of this series of shocks Britain, France, and Italy joined on April 11, 1935, at a conference at Stresa to reaffirm their opposition to German expansion. Laval and Litvinov also initialed a five-year Franco-Soviet alliance on May 2, each pledging assistance in case of unprovoked aggression. Two weeks later a Czech-Soviet pact complemented it. Laval's system, however, was flawed; mutual suspicion between Paris and Moscow, the failure to add a military convention, and the lack of Polish adherence meant that genuine Franco-Soviet military action was unlikely. The U.S.S.R. was in a state of trauma brought on by the Five-Year Plans, the slaughter and starvation of millions of farmers, especially in the Ukraine, in the name of collectivization, and the beginnings of Stalin's mass purges of the government, army, and Communist party. It was clear that Russian industrialization was bound to overthrow the balance of power in Eurasia, hence Stalin was fearful of the possibility of a preemptive attack before his own militarization was complete. But he was even more obsessed with the prospect of wholesale rebellion against his regime in case of invasion. Stalin's primary goal, therefore, was to keep the capitalist powers divided and the U.S.S.R. at peace. Urging the liberal Western states to combine against the Fascists was one method; exploring bilateral relations with Germany, as in the 1936 conversations between Hjalmar Schacht and Soviet trade representative David Kandelaki, was another.

Italy and Britain looked askance at the Franco-Soviet combination, while Hitler in any case sugar-coated the pill of German rearmament by making a pacific speech on May 21, 1935, in which he offered bilateral pacts to all Germany's neighbours (except Lithuania) and assured the British that he, unlike the Kaiser, did not intend to challenge them on the seas. The Anglo-German Naval Agreement of June 18, which countenanced a new German navy though limiting it to not larger than 35 percent the size of the British, angered the French and drove a wedge between them and the British.

The Anglo-German naval pact

Italian aggression. The Stresa Front collapsed as soon as Paris and London learned the price Mussolini meant to exact for it. By 1935 Mussolini had ruled for 13 years but had made little progress toward his "new Roman Empire" that was to free Italy from the "prison of the Mediterranean." What was more, Il Duce concluded that only the crucible of war could fully undermine the monarchy and the church and consummate the Fascist revolution at home. Having failed to pry the French out of their North African possessions, Mussolini fixed on the independent African empire of Abyssinia (Ethiopia). Italy had failed in 1896 to conquer Abyssinia, thus to do so now would erase a national humiliation. This spacious land astride Italy's existing coastal colonies on the Horn of Africa boasted fertile uplands suitable for Italy's excess rural population, and Mussolini promised abundant raw materials as well. The conquest of Abyssinia would also appear to open the path to the Sudan and Suez. Finally, this landlocked, semifeudal kingdom seemed an easy target. In fact, Emperor Haile Selassie had begun a modernization program of sorts, but this only suggested that the sooner Italy struck, the better.

The Italian army was scarcely prepared for such an undertaking, and Mussolini made matters worse by ordering ill-trained blackshirt brigades to Africa and entrusting the campaign to a Fascist loyalist, Emilio De Bono, rather than to a senior army officer. The military buildup at Mitsiwa left little doubt as to Italian intentions, and Britain tried in June to forestall the invasion by arranging the cession of some Abyssinian territories. But Mussolini knew that the British Mediterranean fleet was as unready as his own and expected no interference.

De Bono's absurdly large army invaded Ethiopia from Eritrea on Oct. 3, 1935. Adwa, the site of the 1896 debacle, fell in three days, after which the advance bogged down and Mussolini replaced De Bono with Marshal Pietro Badoglio. The League Council promptly declared Italy the aggressor (October 7), whereupon France and Britain were caught on the horns of a dilemma. To wink at Italy's conquest would be to condone aggression and admit the bankruptcy of the League; to resist would be to smash the Stresa Front and lose Italian help against the greater threat, Germany. The League finally settled on economic sanctions but shied away from an embargo on oil, which would have grounded the Italian army and air force, or closure of the Suez Canal, which would have cut the

Invasion of Ethiopia

Italian supply line. The remaining sanctions only vexed Italy without helping Abyssinia. Germany, no longer a League member, ignored the sanctions and so healed its rift with Rome.

In December, Laval and Sir Samuel Hoare, the British foreign secretary, contrived a secret plan to offer Mussolini most of Abyssinia in return for a truce. This Hoare–Laval Plan was a realistic effort to end the crisis and repair the Stresa Front, but it also made a mockery of the League. When it leaked to the press, public indignation forced Hoare's resignation. On the battlefield, the Italians finally took the fortress of Mekele on November 8, but their slow advance led Mussolini to order a major offensive in December. He instructed Badoglio to use whatever means necessary, including terror bombing and poison gas, to end the war.

The first German move. Hitler observed the Abyssinian War with controlled glee, for dissolution of the Stresa Front—composed of the guarantors of Locarno—gave him the chance to reoccupy the Rhineland with minimal risk. A caretaker government under Albert Sarraut was in charge of France during a divisive electoral campaign dominated by the leftist Popular Front, and Britain was convulsed by a constitutional crisis stemming from King Edward VIII's insistence on marrying an American divorcée. On March 7, 1936, Hitler ordered a token force of 22,000 soldiers back across the bridges of the Rhine. Characteristically, he chose a weekend for his sudden move and then softened the blow with offers of nonaggression pacts and a new demilitarized zone on both sides of the frontier. Even so, Hitler assured his generals he would retreat if the French intervened.

Occupation of the Rhineland

German reoccupation and fortification of the Rhineland was the most significant turning point of the interwar years. After March 1936 the British and French could no longer take forceful action against Hitler except by provoking the total war they feared. Why did the French, especially, not act to prevent this calamity to their defensive posture? They were not taken by surprise—Hitler's preparations had been noted—and Sarraut himself told French radio listeners that "Strasbourg would not be left under German guns." Moreover, the French army still outnumbered the German and could expect support from Czechoslovakia and possibly Poland. On the other hand, the French army commander, General Maurice Gamelin, vastly overestimated German strength and insisted that a move into the Rhineland be preceded by general mobilization. The French Cabinet also concluded that it should do nothing without the full agreement of the British. But London was not the place to look for backbone. Prime Minister Stanley Baldwin shrugged: "They might succeed in smashing Germany with the aid of Russia, but it would probably only result in Germany going Bolshevik," while the editor of *The Times* asked, "It's none of our business, is it? It's their own back-garden they're walking into." But by failing to respond to the violation, Britain, France, and Italy had broken the Locarno Treaties just as gravely as Germany.

The strategic situation in Europe now shifted in favour of the Fascist powers. In June, Mussolini appointed as foreign minister his son-in-law Galeazzo Ciano, who concluded an agreement with Germany on July 11 in which Italy acquiesced in Austria behaving henceforth as "a German state." The Rome–Berlin Axis followed on November 1, and the German–Japanese Anti-Comintern Pact, another vague agreement ostensibly directed at Moscow, on November 25. Finally, Belgium unilaterally renounced its alliance with France on October 14 and returned to its traditional neutrality in hopes of escaping the coming storm. As a direct result of the Abyssinian imbroglio, the militant revisionists had come together, and the status quo powers had splintered.

The Rome–Berlin Axis

Meanwhile, on May 5, 1936, Italian troops entered Addis Ababa and completed the conquest of Abyssinia. The country was never entirely pacified despite costly and brutal repression. But the Abyssinian war had been a disaster for the democracies, smashing both the Stresa Front and the credibility of the League. As historian A.J.P. Taylor wrote: "One day [the League] was a powerful body imposing sanctions, seemingly more effective than ever before; the next day it was an empty sham, everyone scuttling from it as quickly as possible." In December 1937, Italy, too, quit the League of Nations.

The blow to the League

BRITISH APPEASEMENT AND AMERICAN ISOLATIONISM

The rationale of appeasement. It is time to explore the roots of democratic lethargy in the face of Fascist expansionism in the 1930s. British policy, in particular, which Prime Minister Neville Chamberlain would proudly term "appeasement," conjures up images of naive, even craven surrender to Nazi demands. In the minds of British statesmen, however, appeasement was a moral and realistic expression of all that was liberal and Christian in British culture. First, 1914 cast a dark shadow on the opinion leaders of the 1930s, who determined this time to shun arms races and balance-of-power and commercial competition, and so spare the world another horrible war. Second, the overextended British Empire lacked the resources to confront threats from Japan in Asia, Italy in the Mediterranean, and Germany in Europe all at once. Wisdom dictated that Britain come to terms with the greatest and closest to home of its potential adversaries, Germany. Third, the British public was understandably provincial about central Europe and had no desire (in the popular French phrase) "to die for Danzig." This sentiment was even more pronounced in the British Dominions. Fourth, many Tory and Labour leaders, while put off by Hitler's ideology and brutality, shared his antipathy to Versailles and urged "fair play" in cases where German nationals were separated from the fatherland. Thus Wilsonian national self-determination perversely made the Nazis appear to be on the side of principle. Fifth, the appeasers also presumed that the Nazis would become less rambunctious once their grievances were removed. Sixth, some demoralized Englishmen believed the propagandistic claim that Fascism was the only bulwark against the spread of Bolshevism. Seventh, domestic opinion in Britain favoured a passive reliance on the League of Nations somehow to prevent another catastrophe—Baldwin's policy of sanctions without war in Abyssinia, as the chief case in point, earned his party a huge electoral victory in November 1935. Nor had pacifism flagged since 1933, when the Oxford Union "Resolved that this house refuses to fight for King and Country."

Voices of dissent existed. Some Left-Labourites warned that Fascism must be stopped sooner or later, while a few Tory backbenchers led by Winston Churchill demanded rearmament. In the mid-1930s a source in the Air Ministry leaked data to Churchill suggesting that Germany's air force was rapidly overtaking Britain's. But fear of the Luftwaffe only provided another excuse for appeasement, for aviation had developed to the point that theorists like the Italian Giulio Douhet could argue that air bombardment would win the next war in 48 hours by leveling enemy cities. In an air age, the English Channel no longer sheltered Britain from destruction.

Fear of war from the air

Many of these same considerations afflicted French policy: fear of another total war and destruction from the air, apathy toward eastern Europe, and ideological confusion. The election of May 3, 1936, brought victory for the Popular Front, which formed a Cabinet under the Socialist Léon Blum. But his economic policies threw France into a turmoil of strikes, capital flight, and recrimination. "Better Hitler than Blum," said some on the Right.

The civil war in Spain. The Spanish Civil War highlighted the contrast between democratic bankruptcy and totalitarian dynamism. In 1931 the Spanish monarchy gave way to a republic whose unstable government moved steadily to the left, outraging the army and church. After repeated provocations on both sides, army and air force officers proclaimed a Nationalist revolt on July 17, 1936, that survived its critical early weeks with logistical help from Portugal's arch-conservative premier António Salazar. The Nationalists, rallying behind General Francisco Franco, quickly seized most of Old Castile in the north and a beachhead in the south extending from Córdoba to Cádiz opposite Spanish Morocco, where the insurrection had begun. But the Loyalists, a Popular Front

composed of liberals, Socialists, Trotskyites, Stalinists, and anarchists took up arms to defend the Republic elsewhere and sought outside aid against what they styled as the latest Fascist threat. Spain became a battleground for the ideologies wrestling for mastery of Europe.

The civil war posed a dilemma for France and Britain, pitting the principle of defending democracy against the principle of noninterference in the domestic affairs of other states. The ineffectual Blum at first fraternally promised aid to the Popular Front in Madrid, but he reneged within a month for fear that such involvement might provoke a European war or a civil war in France. The British government counseled nonintervention and seemingly won Germany and Italy to that position. But Hitler, on well-rehearsed anti-Bolshevik grounds, hurriedly dispatched 20 transport planes that allowed Franco to move reinforcements from Morocco. Not to be outdone, Mussolini sent matériel, Fascist "volunteers," and, ultimately, regular army formations. The Italians performed miserably (especially at Guadalajara, March 1937), but German aid, including the feared Condor Legion, was effective. Hitler expected to be paid for his support, however, with economic concessions, and he also saw Spain as a testing-ground for Germany's newest weapons and tactics. These included terror bombing such as that over Guernica in April 1937, which caused far fewer deaths than legend has it but became an icon of anti-Fascism through the painting of Pablo Picasso. International aid to the Loyalists ran from the heroic to the sinister. Thousands of leftists and idealistic volunteers from all over Europe and America flocked to International Brigades to defend the Republic. Material support, however, came only from Stalin, who demanded gold payment in return and ordered Comintern agents and commissars to accompany the Soviet supplies. These Stalinists systematically murdered Trotskyites and other "enemies on the left," undermined the radical government of Barcelona, and exacerbated the intramural confusion in Loyalist ranks. The upshot of Soviet intervention was to discredit the Republic and thereby strengthen Western resolve to stay out.

The International Brigades

The war dragged on through 1937 and 1938 and claimed some 500,000 lives before the Nationalists finally captured Barcelona in January 1939 and Madrid in March. During the final push to victory, France and Britain recognized Franco's government. By then the fulcrum of diplomacy had long since shifted to central Europe. But the Nationalist victory did not, in the end, redound to the detriment of France, for Franco politely sent the Germans and Italians home and observed neutrality in the coming war, whereas a pro-Communist Spain might have posed a genuine threat to France during the era of the Nazi–Soviet pact.

The return of U.S. isolationism. The extreme isolationism that gripped the United States in the 1930s reinforced British appeasement and French paralysis. To Americans absorbed with their own distress, Hitler and Mussolini appeared as slightly ridiculous rabble-rousers on movie-house newsreels and certainly no concern of theirs. Moreover, the revisionist theory that the United States had been sucked into war in 1917 through the machinations of arms merchants or Wall Street bankers gained credence from the Senate's Nye Committee inquiries of 1934–36. But U.S. isolationism, like British appeasement, had many roots: liberal abhorrence of arms and war, the evident failure of Wilsonianism, the Depression, the revisionism of American historians, who were among the leaders in arguing that Germany was not solely responsible for 1914. Nor were isolationists restricted only to the Great Plains states or to one political party. Some members of Congress favoured punctilious defense of U.S. interests in the world but rejected involvement in the quarrels of others. Some were full-fledged pacifists even if it meant surrendering some U.S. rights abroad. Left-wing isolationists warned that another great war would push the United States in the direction of Fascism. Conservative isolationists warned that another great war would usher in socialism.

These factions disputed among themselves over the wording of legislation, but their collective strength was enough to carry a number of bills designed to prevent a recurrence of the events of 1914–17. The Johnson Act of 1934 forbade American citizens to lend money to foreign countries that had not paid their past war debts. The Neutrality acts of 1935 and 1936 prohibited sale of war matériel to belligerents and forbade any exports to belligerents not paid for with cash and carried in their own ships. Thus, the United States was not to acquire a stake in the victory of any side or expose its merchant ships to submarines. The effect of these acts, however, was to preclude American aid to Abyssinia, Spain, and China, and thus hurt the victims of aggression more than the aggressors.

The Neutrality acts

The United States did take steps in the 1930s, however, to mobilize the Western Hemisphere for the purposes of fighting the Depression and resisting European, especially German, encroachments. Roosevelt gave this initiative a name in his first inaugural address: the Good Neighbor Policy. Building on steps taken by Hoover, Roosevelt pledged nonintervention in Latin domestic affairs at the Montevideo Pan-American Conference of 1933, signed a treaty with the new Cuban government (May 29, 1934) abrogating the Platt Amendment, mediated a truce in the Chaco War between Bolivia and Paraguay in 1934 (a peace treaty following in July 1938), and negotiated commercial treaties with Latin-American states. As war approached overseas, Washington also promoted pan-American unity on the basis of nonintervention, condemnation of aggression, no forcible collection of debts, equality of states, respect for treaties, and continental solidarity. The Declaration of Lima (1938) provided for pan-American consultation in case of a threat to the "peace, security, or territorial integrity" of any state.

Japan's aggression in China. The first major challenge to American isolationism, however, occurred in Asia. After pacifying Manchukuo, the Japanese turned their sights toward North China and Inner Mongolia. But over the intervening years the KMT had made progress in unifying China. The Communists were still in the field, having survived their Long March (1934–35) to Yen-an in the north, but Chiang's government, with German and American help, had introduced modern roads and communications, stable paper currency, banking, and educational systems. How might Tokyo best round out its continental interests: by preemptive war or by cooperating with this resurgent China to expel Western influence from East Asia? The chief of the operations section of the Japanese general staff favoured collaboration and feared that an invasion of China proper would bring war with the Soviets or the Americans, whose economic potential he understood. But supreme headquarters preferred to take military advantage of apparent friction between Chiang and a North China warlord. In September 1936 Japan issued seven secret demands that would have made North China a virtual Japanese protectorate, but Chiang rejected them. In December Chiang was even kidnapped by one of his own generals, the commander of Nationalist forces from Manchuria, who tried to force him to suspend fighting the Communists and to declare war on Japan. This Sian Incident demonstrated the unlikelihood of Chinese collaboration with the Japanese program and strengthened the war party in Tokyo. As in 1931, hostilities began almost spontaneously and soon took on a life of their own.

The Sian Incident

An incident at the Marco Polo Bridge near Peking (then known as Pei-p'ing) on July 7, 1937, escalated into an undeclared Sino-Japanese war. Contrary to the Japanese analysis, both Chiang and Mao Zedong vowed to come to the aid of North China, while Japanese moderates failed to negotiate a truce or localize the conflict and lost all influence. By the end of July the Japanese had occupied Peking and Tientsin. The following month they blockaded the South China coast and captured Shanghai after brutal fighting and the slaughter of countless civilians. Similar atrocities accompanied the fall of Nanking on December 13. The Japanese now expected the Chinese to sue for peace, but Chiang moved his government to Han-k'ou and continued to resist the "dwarf bandits" with hit-and-run tactics that sucked the invaders in more deeply. The Japanese could occupy cities and fan out along roads and rails almost at will, but the countryside remained hostile.

World opinion condemned Japan in the harshest terms. The U.S.S.R. concluded a nonaggression pact with China

(Aug. 21, 1937), and Soviet-Mongolian forces skirmished with Japanese on the border. Britain vilified Japan in the League, while Roosevelt invoked the Stimson Doctrine in his "quarantine speech" of October 5. But Roosevelt was prevented by the Neutrality acts from aiding China even after the sinking of U.S. and British gunboats on the Yangtze.

Responses to Japan

On March 28, 1938, the Japanese established a Manchukuo-type puppet regime at Nanking, and spring and summer offensives brought them to the Wu-han cities (chiefly Han-k'ou) on the Yangtze. Chiang stubbornly moved his government again, this time to Chungking, which the Japanese bombed mercilessly in May 1939, as they did Canton for weeks before its occupation in October. Such incidents, combined with the Nazi and Fascist air attacks in Spain and Abyssinia, were omens of the total war to come. The United States finally took a first step in opposition to Japanese aggression on July 29, 1939, announcing that it would terminate its 1911 commercial treaty with Japan in six months and thereby cut off vital raw materials to the Japanese war machine. It was all Roosevelt could do under existing law, but it set in train the events that would lead to Pearl Harbor.

ANSCHLUSS AND THE MUNICH PACT

The German-Austrian union. Heightened assertiveness also characterized foreign policies in Europe in 1937. But while Hitler's involved explicit preparations for war, Britain's consisted of explicit attempts to satisfy him with concessions. The conjuncture of these policies doomed the independence of Austria, Czechoslovakia, and Poland, and set Europe on a slippery slope to war.

By the end of 1936, Hitler and the Nazis were total masters of Germany with the exceptions of the army and the foreign office, and even the latter had to tolerate the activities of a special party apparatus under the Nazi "expert" on foreign policy, Joachim von Ribbentrop. Nazi prestige, bolstered by such theatrics as the Berlin Olympics, the German pavilion at the Paris Exhibition, and the enormous Nürnberg party rallies, was reaching its zenith. In September 1936, Hitler imitated Stalin again in his proclamation of a Four-Year Plan to prepare the German economy for war under the leadership of Hermann Göring. With the Rhineland secured, Hitler grew anxious to begin his "drive to the east," if possible with British acquiescence. To this end he appointed Ribbentrop ambassador to London in October 1936 with the plea, "Bring me back the British alliance." Intermittent talks lasted a year, their main topic being the return of the German colonies lost at Versailles. But agreement was impossible, since Hitler's real goal was a free hand on the Continent, while the British hoped, in return for specific concessions, to secure arms control and respect for the status quo.

Germany's Four-Year Plan

Meanwhile, Stanley Baldwin, having seen the abdication crisis through to a finish, retired in May 1937 in favour of Neville Chamberlain. The latter now had the chance to pursue what he termed "active appeasement": find out what Hitler really wants, give it to him, and thereby save the peace and husband British resources for defense of the empire against Italy and Japan. By the time of Lord Halifax's celebrated visit to Berchtesgaden in November 1937, Hitler had already lost interest in the talks and begun to prepare for the absorption of Austria, a country in which, said Halifax, Britain took little interest. Hitler had also taken measures to complete the Nazification of foreign and defense policy.

On November 5, Hitler made a secret speech in the presence of the commanders of the three armed services, War Minister Werner von Blomberg, Foreign Minister Konstantin von Neurath, and Göring. The Führer made clear his belief that Germany must begin to expand in the immediate future, with Austria and Czechoslovakia as the first targets, and that the German economy must be ready for full-scale war by 1943–45. On November 19, Hitler replaced Schacht as minister of economics. Two months later he fired generals Blomberg and Werner von Fritsch in favour of the loyal Walther von Brauchitsch and Wilhelm Keitel and replaced Neurath with Ribbentrop. Historians have debated whether the November 5 speech was a blueprint for aggression, a plea for continued rearmament, or preparation for the purges that followed. But there is no denying that the overheated Nazi economy had reached a critical turn with labour and resources fully employed and capital running short. Hitler would soon have to introduce austerity measures, slow down the arms program, or make good the shortages of labour and capital through plunder. Since these material needs pushed in the same direction as Hitler's dynamic quest for Lebensraum, 1937 merely marked the transition into concrete timetables of what Hitler had always desired. Nazification of the economy, the military, and the foreign service only removed the last vestige of potential opposition to a risky program of ruthless conquest.

German intrigues in Austria had continued since 1936 through the agency of Arthur Seyss-Inquart's Nazi movement. When Papen, now ambassador to Vienna, reported on Feb. 5, 1938, that the Schuschnigg regime showed signs of weakness, Hitler invited the Austrian dictator to a meeting on the 12th. In the course of an intimidating tirade Hitler demanded that Nazis be included in the Vienna government. Schuschnigg, however, insisted that Austria remain "free and German, independent and social, Christian and united," and scheduled a plebiscite for March 13 through which Austrians might express their will. Hitler hurriedly issued directives to the military, and when Schuschnigg was induced to resign, Seyss-Inquart simply appointed himself chancellor and invited German troops to intervene. A last-minute Italian demarche inviting Britain to make colonial concessions in return for Italian support of Austria met only "indignant resignation" and Anthony Eden's irrelevant complaints about Italy's troops in Spain. A French plea for Italian firmness, in turn, provoked Ciano to ask: "Do they expect to rebuild Stresa in an hour with Hannibal at the gates?" Still, Hitler waited nervously on the evening of March 11 until he was informed that Mussolini would take no action in support of Austria. Hitler replied with effusive thanks and promises of eternal amity. In the nighttime invasion, 70 percent of the vehicles sent into Austria by the unprepared Wehrmacht broke down on the road to Vienna, but they met no resistance. Austrians cheered deliriously on the 13th, when Hitler declared Austria a province of the Reich.

The end of Austrian resistance

The taking of Czechoslovakia. The *Anschluss* outflanked the next state on Hitler's list, Czechoslovakia. Once again Hitler could make use of national self-determination to confuse the issue, as 3,500,000 German-speakers organized by another Nazi henchman, Konrad Henlein, inhabited the Czech borderlands in the Sudeten Mountains. Already on February 20, before the *Anschluss,* Hitler had denounced the Czechs for alleged persecution of this German minority, and on April 21 he ordered Keitel to prepare for the invasion of Czechoslovakia by October even if the French should intervene. Chamberlain was intent on appeasing Hitler, but this meant "educating" him to seek redress of grievances through negotiation, not force. He issued a stern warning to Germany during the spring war scare while pressuring Beneš to compromise with Henlein. Germany, however, had instructed Henlein to display obstinacy so as to prevent agreement. In August a worried British Cabinet dispatched the elderly Lord Walter Runciman to mediate, but Henlein rejected the program of concessions he finally arranged with Beneš. As the prospect of war increased, the British appeasers grew more frantic. In the spring the editor of the leftist *New Statesman* thought "armed resistance to the dictators was now useless. If there was a war we should lose it." General Edmund Ironside, ruing the prime minister's reluctance to rearm, sneered that "Chamberlain is of course right. . . . We cannot expose ourselves now to a German attack. We simply commit suicide if we do." And a shocking *Times* editorial called for the partition of Czechoslovakia, a view shared by Hitler at the Nürnberg party rally, where he condemned "Czechia" as an "artificial state." Chamberlain then journeyed to Berchtesgaden and proposed to give the Germans all they demanded. Hitler, nonplussed, spoke of the cession of all Sudeten areas at least 80 percent German and agreed not to invade while Chamberlain won over Paris and Prague.

The abandonment of the Czechs

The French Cabinet of Édouard Daladier and Georges-Etienne Bonnet agreed, after the latter's frantic pleas to Roosevelt failed to shake American isolation. But the Czechs resisted handing over their border fortifications to Hitler until September 21, when the British and French made clear they would not fight for the Sudetenland. Chamberlain flew to Bad Godesberg the next day only to be met with a new demand that the entire Sudetenland be ceded to Germany within a week. The Czechs, fully mobilized as of the 23rd, refused, and Chamberlain returned home in a funk: "How horrible, fantastic, incredible it is that we should be digging trenches and trying on gas masks here because of a quarrel in a far-away country between people of whom we know nothing." But his sorrowful address to Parliament was interrupted by the news that Mussolini had proposed a conference to settle the crisis peacefully. Hitler agreed, having seen how little enthusiasm there was in Germany for war and on the advice of Göring, Goebbels, and the generals. Chamberlain and Daladier, elated, flew to Munich on September 29.

The awkward and pitiful Munich Conference ended on the 30th in a compromise prearranged between the two dictators. The Czechs were to evacuate all regions indicated by an international commission (subsequently dominated by the Germans) by October 10 and were given no recourse—the agreement was final. Poland took the opportunity to grab the Teschen district disputed since 1919. Czechoslovakia was no longer a viable state, and Beneš resigned the presidency in despair. In return, Hitler promised no more territorial demands in Europe and consultations with Britain in case of any future threat to peace. Chamberlain was ecstatic.

Why did the Western powers abandon Czechoslovakia, which, by dint of its geography, democracy, military potential (more than 30 divisions and the Škoda arms works), and commitment to collective security, could rightly be called "the keystone of interwar Europe"? No completely persuasive answer is possible, but this height of appeasement can be accounted for by politics, principles, and pragmatism. There is no question that the Munich settlement was extremely popular. Chamberlain returned to London claiming "peace in our time" and was greeted by applauding throngs. So was Daladier. The relief was so evident even in Germany that Hitler swore he would allow no more meddling by "English governesses" to cheat him of his war. Of course, the euphoria was not universal: Aside from the Czechs, who wept in the streets, Churchill spoke for a growing minority when he observed that the British Empire had just suffered its worst military defeat and had not fired a shot.

Could Czechoslovakia have been defended? Or was Munich a necessary evil to buy time for Britain to rearm? Certainly British air defenses were unready, while France's scarcely existed, and the strength of the Luftwaffe, so recently discounted by the British Cabinet, was now exaggerated. The French and Czech armies still outnumbered the German, but French intelligence also magnified German strength, while the army had no plans for invading Germany in support of the Czechs. The Munich powers were criticized for ignoring the U.S.S.R., which had claimed readiness to honour its alliance with Prague. But the U.S.S.R. would hardly confront Germany unless the Western powers were already engaged, and the ways open to them were few without transit rights across Poland. The West discounted Soviet military effectiveness in light of Stalin's 1937 purge of his entire officer corps down to battalion level. The Soviets were also distracted by divisional-scale fighting that broke out with Japanese forces on the Manchurian border in July–August 1938. At best, a few squadrons of Soviet planes might have been sent to Prague.

Of course, the moral cause of liberating the Sudeten-Germans was ludicrous in view of the nature of the Nazi regime and was far outweighed by the moral lapse of deserting the doughty Czechs (French ambassador André François-Poncet, upon reading the Munich accord, choked, "Thus does France treat her only allies who had remained faithful to her.") That betrayal, in turn, seemed more than outweighed by the moral cause of preventing another war. In the end, the war was only delayed a year, and whatever the military realities of 1938 versus 1939, the appeasement policy was an exercise in self-delusion. Chamberlain and his ilk did not begin their reasoning with an analysis of Hitlerism and then work forward to a policy. Rather, they began with a policy based on abstract analysis of the causes of war, then worked backward to an image of Hitler that suited the needs of that policy. As a result, they gave Hitler far more than they ever gave the democratic statesmen of Weimar and, in the end, the freedom to launch the very war they slaved to prevent.

Hitler had no intention of honouring Munich. In October the Nazis encouraged the Slovak and Ruthene minorities in Czechoslovakia to set up autonomous governments and then in November awarded Hungary the 4,600 square miles north of the Danube taken from it in 1919. On March 13, 1939, Gestapo officers carried the Slovak leader Monsignor Jozef Tiso off to Berlin and deposited him in the presence of the Führer, who demanded that the Slovaks declare their independence at once. Tiso returned to Bratislava to inform the Slovak Diet that the only alternative to becoming a Nazi protectorate was invasion. They complied. All that remained to the new president in Prague, Emil Hácha, was the core region of Bohemia and Moravia. It was time, said Hácha with heavy sarcasm, "to consult our friends in Germany." There Hitler subjected the elderly, broken-spirited man to a tirade that brought tears, a fainting spell, and finally a signature on a "request" that Bohemia and Moravia be incorporated into the Reich. The next day, March 16, German units occupied Prague, and Czechoslovakia ceased to exist.

Nazi incitement of minorities

TECHNOLOGY, STRATEGY, AND THE OUTBREAK OF WAR

Rearmament and tactical planning. The Anglo-French defection from east-central Europe doomed the balance of power of interwar Europe. That the Western powers were unwilling and unable to defend the balance was in part the product of inadequate military spending and planning over the course of the decade. Still, decisions were taken in the last 24 months of peace that would shape the course of World War II.

The central problem posed for all defense establishments was how to respond to the lessons of the 1914–18 stalemate. The British simply determined not to send an army to the Continent again, the French to turn their border into an impregnable fortress, and the Germans to perfect and synthesize the tactics and technologies of the last war into a dynamic new style of warfare: the Blitzkrieg ("lightning war"). Blitzkrieg was especially suited to a country whose geostrategic position made likely a war on two fronts and dictated an offensive posture: a Schlieffen solution made plausible by the internal-combustion engine. Whether or not Hitler actually planned for the type of war with which the general staff was experimenting is debatable. Perhaps he only made a virtue of necessity, for the Nazis by no means created a full war economy in the 1930s. Since Blitzkrieg attacks by tank columns, motorized infantry, and aircraft permitted the defeat of enemies one by one with lightning speed, it required only "armament in width," not "armament in depth." This in turn allowed Hitler to mollify the German people with a "guns and butter" economy, with each new conquest providing the resources for the next. Blitzkrieg also allowed Hitler to conclude that he might successfully defy other Great Powers whose combined resources dwarfed those of Germany. After Munich, German rearmament accelerated, but not nearly to the degree of likely enemies. Hitler may have been right to launch his war as soon as possible, on the calculation that only by seizing the resources of the entire continent could the Reich prevail against the British Empire or the Soviet Union.

Invention of Blitzkrieg

After Versailles the British government had established the Ten-Year Rule as a rationale for holding down military spending: Each year it was determined that virtually no chance existed of war breaking out over the next decade. In 1931 expenditures were cut to the bone in response to the worldwide financial crisis. The following year, in response to Japanese expansion, the Ten-Year Rule was abolished, but Britain did not make even a gesture toward

rearmament until 1935. These were "the years the locust hath eaten," said Churchill. Understandably, British strategy fixed on the imperial threats from Japan and Italy and envisioned the dispatch of the Mediterranean fleet to Singapore. But Britain's defensive posture, budgetary limits, and underestimation of Japan's capabilities, especially in the air, made for a desultory buildup in battleships and cruisers rather than aircraft carriers. The British army in turn was tied up in garrisoning the empire; only two divisions were available for the Continent.

Britain's home air defense

After March 1936 the Defence Requirements Committee recognized that home air defense must become Britain's top priority and commanded development of a high-speed, single-wing fighter plane. But two years passed before Sir Warren Fisher finally persuaded the Air Ministry to concentrate on fighter defense in its Scheme M, adopted in November 1938. At the time of Munich, therefore, the Royal Air Force possessed only two squadrons of Spitfires and Hurricanes, lacked oxygen masks sufficient to allow pursuit above 15,000 feet, and had barely begun deployment of that new wonder, radar. Only after Hitler's occupation of Prague was conscription reinstated (April 27, 1939) and a continental army of 32 divisions planned. Throughout the era of appeasement the British expected to resist Japan and come to terms with Germany. Instead, by dint of the mistaken choices in naval technology and the eleventh-hour attention to air defense, Britain would be humiliated by Japan and withstand Germany.

The Maginot Line

Of all the Great Powers, France most expected the next war to resemble the last and so came to rely on the doctrine of the continuous front, the Maginot Line, and the primacy of infantry and artillery. The Maginot Line was also a function of French demographic weakness vis-à-vis Germany, especially after military service was cut to one year in 1928. This siege mentality was the polar opposite of the French "cult of the attack" in 1914 and ensured that Colonel Charles de Gaulle's 1934 book depicting an all-mechanized army of the future would be ignored. As late as 1939 the French war council insisted that "no new method of warfare has been evolved since the termination of the Great War." Even though French military spending held steady through the Depression, France's army and air force were ill-designed and not deployed for offense or mobile defense, even if their aged and hidebound commanders had had the will to conduct them.

Soviet preparations and technical choices also presaged the defeats to come in the early years of the war. Communist doctrine decreed that matériel, not generalship, was decisive in war, and Stalin's Five-Year plans concentrated on steel, technology, and weapons. Soviet planners also benefited from the work of some outstanding aviation designers, whose experimental planes broke world records and whose fighters performed well in the early days of the Spanish war. But Stalin's obsession with domestic security outweighed rational planning for national security. In 1937 Marshal Mikhail Tukhachevsky and his weapons research teams were liquidated or consigned to the *gulag*. Then Stalin ordered the 1936-vintage fighter planes into mass production at the very time the Germans were upgrading their Messerschmidts. The Soviets were sufficiently impressed by Douhet's theories to invest in heavy bombers that would be of marginal use against a Blitzkrieg and defenseless without fighter cover. Stalin's advisers also misunderstood the use of tanks, placing them in the front line rather than in mobile reserves. These mistakes almost spelled the death of Bolshevism in 1941.

Italian unreadiness

Little need be said of Italian preparations. Italy's industrial base was so small, and its leaders so inept, that Mussolini had to order local Fascists to make a visual count of airplanes on fields around the country to contrive an estimate of his air strength. In August 1939, Ciano appealed to Mussolini not to join Hitler in unleashing war, given the deplorable state of Italian armed forces. This apprehensiveness was shared by the Italian generals and indeed by most military leaders of the 1930s. The Great War had revealed the vanity of planning, the vagaries of technical change, and the terrible cost of industrial war. In 1914 the generals had pushed for war while civilian leaders hung back; in the 1930s the roles were reversed. Only in Japan, which had won easy victories at little cost in 1914, did the military push for action.

Poland and Soviet anxiety. Hitler's cynical occupation of Prague, giving the final lie to all his peaceful protestations after Munich, prompted much speculation about the identity of his next victim: Romania with its oil reserves, the Ukraine, Poland, or even the "Germanic" Netherlands, which suffered an invasion scare in January? Chamberlain himself, offended in conscience and ego, attacked Hitler's mendacity and evident intention of dominating the continent by force. In a speech on March 17, 1939, he gave voice to the new conviction of "the man on the street" that Hitler could not be trusted and must be stopped. Three days later Hitler renewed his demand for a "corridor across the [Polish] Corridor" to East Prussia and restoration of Danzig to the Reich. On the 22nd he underscored his seriousness by forcing Lithuania to cede Memel (Klaipėda).

After 10 days of hand wringing, during which Colonel Beck repeated Poland's opposition to seeking help from Moscow, the British Cabinet declared a unilateral military guarantee of Polish security on March 31, solemnized in a bilateral treaty on April 6. It seemed an extraordinary turnaround in British policy: the apparent end of appeasement. In fact, it was a last desperate effort by Chamberlain to preserve appeasement and teach Hitler to settle foreign disputes by diplomacy, as at Munich, and not by force, as at Prague. But the pace of Fascist expansion was irreversible and even contagious. Mussolini had grown irritable over Hitler's succession of coups and his own junior-partner status, so Italy occupied Albania on April 7 and expelled its erstwhile client King Zog. Hitler, who reacted to the British guarantee with the oath, "I'll cook them a stew they'll choke on!" renounced his 1934 pact with Poland and the Anglo-German Naval Treaty on the 28th. Germany and Italy then turned their Axis into a military alliance known as the Pact of Steel on May 22.

The Pact of Steel

How could Britain and France ever make good on their pledges to defend Poland? British planning called only for a naval blockade in the early stages of war, while the French (despite a promise to attack) contemplated no action beyond French soil. The answer was that the Polish guarantee was a military bluff unless the Red Army could somehow be enlisted. So finally, in the late spring of 1939, the Western allies went in search of collaboration with Moscow.

Stalin had witnessed events during the era of appeasement with growing suspicion and moved his pieces on the chessboard with deftness and cynicism. His overriding purpose was to deflect the thrusts of Germany and Japan elsewhere or—if the U.S.S.R. were forced to fight—make certain that the Western powers were likewise engaged. German reoccupation of the Rhineland had been a military setback, since it freed Germany for adventures to the east, but a diplomatic boon, since it enhanced the value of the Soviet alliance for France. The Anti-Comintern Pact had opened the terrible possibility for the Soviet Union of a war on two fronts, but it soon developed that Berlin and Tokyo were both expecting the other to stand guard over Russia while they pursued booty in central Europe and China respectively. Now Britain and France were promising to fight Hitler over Poland, thereby handing Stalin the choice of joining the Western powers in war or dealing separately with Germany to avoid conflict entirely. Fearing that war might unleash rebellion at home, Stalin chose to become the greatest appeaser of all.

Stalin's choice

It is often said that Munich forced Stalin to conclude that the Western powers were pushing Nazi Germany to the east and thus reluctantly to consider rapprochement with Hitler. But one might just as well interpret Litvinov's passionate pleas for collective security as a ploy to provoke conflict between Germany and the West while the U.S.S.R. huddled in safety behind its Polish buffer. The incident that made possible the union of the two dictators, as historian Adam Ulam has shown, was not Munich but the British guarantee of Poland. Before that act Stalin faced the prospect of an unopposed German march into Poland, whereupon the U.S.S.R. would be in mortal danger. After that act, Hitler could seize Poland only at the

cost of war with the West, whereupon Hitler would need the U.S.S.R. as an ally. The British guarantee thus made Stalin the arbiter of Europe.

In a contest for Soviet friendship, however, the Allies were at a distinct disadvantage. All they could offer Stalin was the likelihood of war, albeit in alliance with them. On May 3, Stalin replaced Foreign Minister Litvinov, pro-Western and a Jew, with Vyacheslav Molotov—a clear signal of his willingness to improve relations with the Nazis. The Western powers accordingly stepped up their appeals to Moscow for an alliance, but they faced two lofty hurdles. First, Stalin demanded the right to occupy the Baltic states and portions of Romania. While Westerners could scarcely expect to enlist the Red Army in their cause without giving something in return, they could scarcely justify turning free peoples over to Stalinist tyranny. Second, the Poles, as always, refused to invite the Red Army onto lands they had wrested from that same army just 18 years before. By July, Stalin was also demanding that a military convention precede the political one to ensure that he was not left in the lurch. Ironically, the only ploy likely to persuade Stalin of Western sincerity was a blunt threat that the West would *not* fight for Poland *unless* the U.S.S.R. participated.

Since the spring of 1939 the U.S.S.R. had been sending signals to Berlin that Hitler alternately acknowledged and ignored. His hatred for the Moscow regime was overcome, however, by the urgings of Ribbentrop and the unease of his generals. The Soviets, for their part, were again fighting heavy battles along the Manchurian border and in need of security in Europe. Soviet bargaining power was enhanced by the fact that Hitler had a timetable: He had ordered the invasion of Poland by August 26. Negotiations dragged on from July 18 to August 21, when Hitler insisted that Stalin receive Ribbentrop and conclude their business two days hence. On Aug. 23, 1939, therefore, Ribbentrop and Molotov signed the German–Soviet Nonaggression Pact in Moscow, then raised their glasses as Stalin, the leader of world Communism, toasted the German people and their beloved Führer and vowed never to betray them. This nonagression pact was in fact a pact of aggression against Poland, which was to be partitioned, roughly along the old Curzon Line. Hitler also granted the U.S.S.R. a free hand in Finland, the Baltic states, and Bessarabia.

The German–Soviet Nonaggression Pact

Hitler expected that his successful wooing of Russia would oblige Britain and France to withdraw their pledge to Poland. The free peoples were indeed shocked by the news from Moscow, but far from succumbing, they steeled their will to resist. The world situation, so cloudy since 1933, suddenly seemed clear, and scales fell from many eyes. The abstract and often effete ideological debate over democratic decadence and the relative merits of Fascism and Communism came suddenly to an end. Both vaunted ideologies now seemed so much lying propaganda, and their patrons so many gangsters. The day after the pact Chamberlain wrote to Hitler to warn that British resolve was as firm as ever, and on the 25th he signed a full alliance with Poland. British determination and the news that Italy was not ready for war prompted Hitler to delay his invasion a week in hopes of detaching Britain with promises of treaties and guarantees of the British Empire. When Chamberlain refused, Hitler demanded a Polish plenipotentiary be sent to Berlin on August 30 to settle the matter of Danzig and the Corridor. Should the Poles refuse, their obstinancy might give London an excuse to leave them to their fate. But Colonel Beck had seen the fate of Schuschnigg and Hácha; he would not submit to a Hitlerian kidnapping nor another Munich. When Hitler's ultimatum expired, the German army staged a border incident and invaded Poland in force on the morning of Sept. 1, 1939.

Invasion of Poland

The British and French parliaments, confident that their governments had turned every stone in search of peace, declared war on Germany on September 3.

HITLER'S WAR OR CHAMBERLAIN'S?

For two decades after 1939, German guilt for the outbreak of World War II seemed incontestable. The Nürnberg war-crimes trials in 1946 brought to light damning evidence of Nazi ambitions, preparations for war, and deliberate provocation of the crises over Austria, the Sudetenland, and Poland. Revelation of Nazi tyranny, torture, and genocide was a powerful deterrent to anyone in the West inclined to dilute German guilt. To be sure, there were bitter recriminations in France and Britain against those who had failed to stand up to Hitler, and the United States and the U.S.S.R. alike were later to invoke the lessons of the 1930s to justify Cold War policies: Appeasement only feeds the appetite of aggressors; there must be "no more Munichs." But World War II was undeniably Hitler's war, as the ongoing publication of captured German documents seemed to prove.

The British historian A.J.P. Taylor challenged the thesis of sole Nazi guilt in 1961, coincidently the same year Fritz Fischer revived the notion of German guilt for World War I. Taylor boldly suggested that Hitler's "ideology" was nothing more than the sort of nationalist ravings "which echo the conversation of any Austrian cafe or German beer-house"; that Hitler's ends and means resembled those of any "traditional German statesman"; and that the war came because Britain and France dithered between appeasement and resistance, leading Hitler to miscalculate and bring on the accident of September 1939. Needless to say, revisionism on a figure so odious as Hitler sparked vigorous rebuttal and debate. If Hitler had been a traditional statesman, then appeasement would have worked, said some. If the British had been consistent in appeasement—or resisted earlier—the war would not have happened, said others.

Fischer's theses on World War I were also significant, for if Germany at that time was bent on European hegemony and world power, then one could argue a continuity in German foreign policy from at least 1890 to 1945. Devotees of the "primacy of domestic policy" even made comparisons between Hitler's use of foreign policy to crush domestic dissent and similar practices under the Kaiser and Bismarck. But how, critics retorted, could one argue for continuity between the traditional imperialism of Wilhelmine Germany and the fanatical racial extermination of Nazi Germany after 1941? At bottom, Hitler was not trying to preserve traditional elites, but to destroy the domestic and international order alike.

Soviet writers tried, without success, to draw a convincing causal chain between capitalist development and Fascism, but the researches of British Marxist T.W. Mason did expose the German economic crisis of 1937, suggesting that the timing of World War II was partly a function of economic pressures. Finally, Alan Bullock suggested a synthesis: Hitler knew where he wanted to go—his will was unbending—but as to how to get there he was flexible, an opportunist. Gerhard Weinberg's exhaustive study of the German documents then confirmed a neo-traditional interpretation to the effect that Hitler was bent on war and Lebensraum and that appeasement only delayed his gratification.

Hitler's will to war

Publication of British and French documents, in turn, enabled historians to sketch a subtler portrait of appeasement. Chamberlain's reputation improved during the 1970s as American historians, conscious of U.S. overextension in the world and sympathetic to detente with the Soviets, came to appreciate the plight of Britain in the 1930s. But financial, military, and strategic rationalizations could not erase the gross misunderstanding of the nature of the enemy that underlay appeasement. The British historian Anthony Adamthwaite concluded in 1984 that despite the accumulation of sources the fact remains that the appeasers' determination to reach agreement with Hitler blinded them to reality. If to understand is not to forgive, neither is it to give the past the odour of inevitability. Hitler wanted war, and Western and Soviet policies throughout the 1930s helped him to achieve it.

World War II, 1939–45

War once again broke out over nationality conflicts in east-central Europe, provoked in part by a German drive for continental hegemony, and it expanded, once again, into a global conflict whose battle zones touched the waters or

heartlands of almost every continent. The total nature of World War II surpassed that of 1914–18 in that civilian populations not only contributed to the war effort but also became direct targets of aerial attack. Moreover, in 1941 the Nazi regime unleashed a war of extermination against Slavs, Jews, and other elements deemed inferior by Hitler's ideology, while Stalinist Russia extended its campaign of terror against the Ukrainians to the conquered Poles. The Japanese–American war in the Pacific also assumed at times the brutal aspect of a war between races. This ultimate democratization of warfare eliminated the age-old distinction between combatants and non-combatants and ensured that total casualties in World War II would greatly exceed those of World War I and that civilian casualties would exceed the military.

The broadening of war

Once again the European war devolved into a contest between a German-occupied *Mitteleuropa* and a peripheral Allied coalition. But this time Italy abandoned neutrality for the German side, and the Soviet Union held out in the east, while France collapsed in the west. Hence Soviet dictator Joseph Stalin took France's place in meetings of the "Big Three," together with Franklin Roosevelt and Winston Churchill. The Japanese chose to remain neutral vis-à-vis the U.S.S.R., while the Grand Alliance of anti-Fascist states simmered with conflicts over strategy and war aims. World War II, therefore, comprised several parallel or overlapping wars, while the war in Europe became a kind of three-way struggle among the forces of democracy, Nazism, and Communism. As soon as German and Japanese power were effaced, the conflicts among the victors burst into the open and gave birth to the Cold War. World War II completed the destruction of the old Great Power system, prepared the disintegration of Europe's overseas empires, and submerged Europe itself into a world arena dominated by the Soviet Union and the United States.

THE LAST EUROPEAN WAR, 1939–41

Poland and the northern war. At first glance Germany might have seemed the underdog in the war launched by Hitler. The Wehrmacht numbered 54 active divisions, compared to 55 French, 30 Polish, and two British divisions available for the Continent. But the combination of German Blitzkrieg tactics, French inactivity, and Russian perfidy doomed Poland to swift defeat. The German army command deployed 40 of its divisions, including all six panzer (armoured) divisions and two-thirds of its 3,500 aircraft in the east. The so-called Siegfried Line in the west, manned by 11 active divisions and reserve units as they became available, sufficed to block a French advance. Beginning on September 1, General Fedor von Bock's northern army corps pinched off the Polish Corridor from East and West Prussia, while General Gerd von Rundstedt's more powerful southern army corps drove across the border from Silesia and Slovakia. Polish Marshal Edward Śmigły-Rydz tried vainly to defend Poland's industrial regions along the frontier, increasing his army's vulnerability to Blitzkrieg. German tanks quickly burst into the rear, while dive-bombing Stukas disrupted Polish supply and reinforcements. The Polish air force was destroyed in 48 hours. Within a week two panzer corps advanced 140 miles to the outskirts of Warsaw and the Bug River line to the south. Śmigły-Rydz's order for a general retreat on the 10th came too late; most Polish forces were already outflanked on the north by General Heinz Guderian's rapid thrust to Brest-Litovsk and on the south by Paul von Kleist's panzers advancing from Lvov. On September 17 the pincers closed, the Soviet army invaded from the east, and the Polish government fled to Romania, whence it made its way to London as the first of many European governments-in-exile. The Warsaw garrison surrendered on the 27th.

The fall of Warsaw

In a protocol of May 15, 1939, the French had promised to take the offensive two weeks after mobilization. Instead, General Maurice Gamelin contented himself with a brief sortie into the Saar, after which the French withdrew to the Maginot Line. The regime most upset by the German walkover in Poland was Hitler's new ally, the Soviets. On September 10, Stalin ordered partial mobilization and loudly boasted of the Red Army's "three million men." Since a callup of reserve troops was scarcely needed merely to occupy Moscow's share of Poland under the German–Soviet pact, this maneuver must have reflected Stalin's fear that the Germans might not stop at the prearranged line. Stalin told the German ambassador on September 25: "In the final settlement of the Polish question anything that in the future might create friction between Germany and the Soviet Union must be avoided." Three days later Molotov signed a new agreement granting Germany a somewhat larger share of Poland as well as extensive Soviet trade in return for a free hand in Lithuania. Only after this second German–Soviet pact did Communist parties in the West fully embrace their new Nazi ally and oppose Western military resistance to Hitler. Henceforth, Stalin was a fearful and solicitous neighbour of the Nazi empire, and he moved quickly to absorb the regions accorded him. By October 10, Latvia, Lithuania, and Estonia had been forced to accept Soviet occupation. When Finland resisted Soviet demands for border rectifications and bases, Stalin ordered the Red Army to attack on November 30. He expected a lightning victory of his own that would impress Hitler and increase Soviet security in the Baltic. Instead, the Finns resisted fiercely in this "Winter War," holding the fortified Mannerheim Line in the south and cutting off the road-bound Soviet columns in the north with their mobile ski troops. The disorganized Red Army, by contrast, showed the effect of the recent military purges. In some cases only the machine guns of NKVD (political police) units kept the soldiers at the front. Soviet military prestige suffered a devastating blow.

No major fighting broke out in the West during this period, sardonically dubbed the "Sitzkrieg," or Phony War. After the fall of Poland, while hope still existed that a repetition of World War I might be avoided, Hitler sought to persuade Britain to renege on its commitment to Poland's defense. In secret contacts and in his "Peace Address" to the Reichstag of October 6 he even hinted at the possibility of restoring a rump Polish state. The Chamberlain Cabinet, betrayed so often by Hitler, refused to acknowledge the demarches, however, and Hitler ordered preparations for an attack in the west by November 12. The army high command protested vigorously against a winter campaign, and bad weather did force a postponement first to January 1940 and then to the spring. Since the French and British were loath to take initiative, the Phony War dragged on. Gamelin's lame proposal of an advance through the Low Countries was moot given the Dutch and Belgian commitments to neutrality. Combat occurred only at sea. In 1939 alone U-boats sank 110 merchant vessels as well as the aircraft carriers *Ark Royal* and *Courageous* (September 14 and 17) and the battleship *Royal Oak* (October 14). The battle cruisers *Scharnhorst* and *Gneisenau* and pocket battleship *Deutschland* eluded British pursuit and returned safely to port. The *Graf Spee,* however, caught in the South Atlantic, sank nine merchantmen before sustaining damage from British cruisers. It then put in at Montevideo, causing a diplomatic crisis for the South American states. The naval situation, therefore, came quickly to resemble that of World War I, with the British fleet maintaining a distant blockade in the North Sea and the Germans waging a submarine war against British shipping.

The "Phony War"

The Russo-Finnish War, however, suggested that Scandinavia might provide a theatre in which to strike a blow at the German–Russian alliance. Beyond the feckless expulsion of the Soviet Union from the League of Nations on December 14, Britain and France contemplated helping the brave Finns—even at the risk of war with Russia—and perhaps cutting the flow of Swedish iron to Germany. The French wanted to send several divisions to Narvik in Norway and thence by land to Finland. The British demurred at such a violation of neutral rights, but Churchill, now first lord of the Admiralty, insisted that "humanity, rather than legality, must be our guide." In the event, the Allies dithered (as did the United States, which debated granting a loan to Finland, the only nation to pay interest on its World War I debt) until a massive Soviet offensive broke the Mannerheim Line in February. Stalin had given

The Russo-Finnish War

a hint of the future by setting up a Finnish Democratic Republic during the war, under the Comintern agent Otto Kuusinen, but he settled for a treaty with Helsinki on March 12, 1940, in which Finland ceded the Karelian isthmus and leased a naval base to the U.S.S.R. on the Hangö peninsula.

The Finnish fiasco toppled Daladier's government in favour of a Cabinet under Paul Reynaud. He and Neville Chamberlain hoped at least to deny the Germans possible U-boat bases by mining or occupying Norwegian ports. But the German navy, too, had persuaded Hitler of the strategic importance of Norway, and on April 9, the day after British minelaying began, the Germans suddenly seized the ports from Oslo to Narvik in a brilliant sea and air operation, and occupied Denmark by Blitzkrieg. British troops contested Norway and managed to capture Narvik on May 27, but by then greater events were unfolding on the Continent. The British evacuated Narvik on June 6, and Vidkun Quisling's collaborationists assumed control of Norway.

The Western front. The Allies' bungling in Scandinavia lost Chamberlain the confidence of Parliament, and King George VI selected Winston Churchill to head the War Cabinet. In the first of many ringing speeches that would sustain the British spirit, Churchill told his nation: "I have nothing to offer but blood, toil, tears, and sweat."

In eight months of warfare all the belligerents had vastly expanded their frontline strength. In May 1940 the German army concentrated 134 divisions on the Western front, including 12 panzer divisions, 3,500 tanks and 5,200 warplanes. The French army totalled 94 divisions, the British 10, and the neutral Belgians and Dutch 22 and eight respectively. The French army possessed some 2,800 tanks, but less than a third were concentrated in armoured units. The French air force, disrupted during the Popular Front, was in any case antiquated, and 90 percent of the artillery dated from World War I. More important, French morale was low, sapped by the memory of the first war's carnage, by political decadence, and by over-reliance on the Maginot Line. Britain's Royal Air Force had become a prodigious force thanks to 1,700 new planes, but commanders were loath to deflect them from home defense to the Continent. The German plan of attack in the west, meanwhile, had evolved since the previous autumn. Originally favouring a Schlieffen-type attack with the mass concentrated on the right wing in Belgium, the Führer had been won to General Erich von Manstein's scheme for a panzer attack through the rugged Ardennes Forest of southern Belgium and Luxembourg. Either route bypassed the Maginot Line, but the latter plan took advantage of the panzer army's ability to pierce French defenses, disrupt the enemy rear, and split Allied forces in two. The concomitant risk was that Allied counterattacks might pinch off and destroy the armoured spearheads at a blow.

French unreadiness for war

The German offensive struck with devastating effect on May 10. Within days the Dutch surrendered. Göring's Luftwaffe did not get the message and proceeded to devastate the central city of Rotterdam, killing numerous civilians and sending a signal to the city of London. Meanwhile, General Gerd von Rundstedt's panzer army picked its way through the Ardennes and emerged in force at Sedan. By May 20, German tanks reached the coast at Abbeville and cut the Allied armies in two. On the 28th, King Leopold III instructed the Belgian army to surrender, while the British government ordered Lord Gort, commanding the British Expeditionary Force, to make for Dunkirk and prepare for evacuation by sea.

As the Blitzkrieg in Poland had shocked Stalin, so the German victory in France shocked Mussolini. For 17 years he had preached the necessity and beauty of war, believing that a neutral Italy would cease to be regarded as a Great Power and that he needed war in order to fulfill his expansionist fantasies and permit the full triumph of Fascism at home. Yet in August 1939 he demanded from Germany 6,000,000 tons of coal, 2,000,000 tons of steel, and 7,000,000 tons of oil before he could honour the Pact of Steel. In fact, war preparations under the corrupt and incompetent Fascists remained feeble, and during these months of nonbelligerence, Mussolini himself took sick and at times even considered joining the Allies. On March 18 he met Hitler at the Brenner Pass and was told that the Germans did not need him to win the war but that he would be allowed to participate and thus escape second-rate status in the Mediterranean. Still Mussolini tried to have it both ways, telling his military chiefs that Italy would not fight Hitler's war, but a "parallel war" to forge "a new Roman Empire." In reality, he would enter the war only when it seemed clear the Allies were finished and his regime would not be put to the test.

That moment seemed to arrive in June 1940. With French defeat assured, Mussolini declared war on France and Britain on the 10th. "The hand that held the dagger," said President Roosevelt, "has struck it into the back of its neighbor." As Mussolini put it to Marshal Pietro Badoglio, "All we need is a few thousand dead" to win a place at the peace conference. The Italian offensive on the Alpine front met contemptuous resistance from the French—Italy's gains were measured literally in yards—but Mussolini was right about the proximity of victory. With German forces streaming east and south, the French government fled on the 11th to Bordeaux and debated three courses of action: request an armistice; transfer the government to North Africa and fight on from the colonies; ask Germany for its terms and temporize. The choice was complicated by a French promise to Britain not to exit the war without London's consent. Churchill, concerned that the French fleet not fall into German hands, went so far as to offer Anglo-French political union on June 16. Reynaud wanted to continue the war but was outvoted. He resigned on the 16th, whereupon the ancient Marshal Pétain asked for an armistice. From London, General Charles de Gaulle broadcast a plea to the French people to fight on and set about organizing Free French forces in France's sub-Saharan colonies. But the armistice was signed at Compiègne, in the same railway car used for the German armistice of 1918, on June 22. The Germans occupied all of northern France and the west coast—60 percent of the country—and the rest was administered by Pétain's quasi-Fascist collaborationist regime at Vichy. The French navy and air force were neutralized. In another meeting of dictators on the 18th, Hitler disappointed Mussolini with his talk of a mild peace lest French forces be driven to defect to Britain. Instead, Pétain broke relations with London on July 4, following a British attack on the French fleet moored at Mers el-Kebir in Algeria. Hitler at once toyed with the notion of winning the Vichy French to an active alliance, thrusting Mussolini farther into the background.

Italy enters the war

The French armistice

Britain's refusal to give up frustrated Hitler, especially since his ultimate goal—Lebensraum—lay in the east. The chief of the army general staff quoted Hitler on May 21 as saying that "we are seeking contact with Britain on the basis of partitioning the world." But when the carrot failed, Hitler tried the stick, authorizing plans on July 2 for Operation Sea Lion, the cross-Channel invasion. Such an operation required complete air superiority, and Göring promised that the Luftwaffe could smash British air defenses in four days. The Battle of Britain that followed in August 1940 was a massive air duel between Germany's 1,200 bombers and a thousand fighter escorts and the RAF's 900 interceptors. But the British Hurricanes and Spitfires were technically superior to all the German fighters except the Me-109, which was restricted in its range to the zone south of London. The British radar screen and ground control network permitted British fighters to concentrate on each German attack. On September 7 Göring made the fatal error of shifting the attack from airfields to London itself (in retaliation for a September 4 raid on Berlin). For 10 days the blitz continued night and day over London, the climax coming on the 15th when nearly 60 German planes were shot down. Two days later Hitler granted that air superiority was not to be had and postponed Operation Sea Lion.

The Battle of Britain

For a full year—June 1940 to June 1941—the British Empire fought on alone (though with growing U.S. aid) against Germany, Italy, and the threat of Japanese action in Asia. Frustrated on sea and in the air, Hitler pondered how his overwhelming land power might be used

to persuade Britain to call it quits. A Mediterranean strategy based on the capture of Gibraltar, Malta, and the Suez Canal, did not seem likely to be decisive, nor did it satisfy the Nazis' *Blut und Boden* ("blood and earth") lust for Lebensraum. To be sure, the Germans raised the prospect of an occupation of Gibraltar numerous times with Franco, but the latter always found an excuse to remain neutral. In fact, Franco knew that the Spanish were exhausted after their civil war and that Spain's Atlantic islands would be lost to the British if it joined the Axis. A Catholic authoritarian, he was also contemptuous of the neo-pagan Fascists. After their last meeting, Hitler confessed that he would rather have his teeth pulled than go through another bout with Franco. Hitler also negotiated with Pétain in July and October 1940 and May 1941, in hopes of enticing France into alliance. But Pétain, too, played a double game, pledging "genuine collaboration" with Germany but reassuring the British that he sought a "cautious balance" between the belligerents.

Spanish neutrality

Hitler's troublesome ally Italy, however, ensured that Germany would be involved in complications to the south. On July 7, 1940, Ciano visited Hitler seeking approval for an expansion of the war to Yugoslavia and Greece. The Führer instead encouraged the occupation of Crete and Cyprus, which would further the war against Britain. But three days later Italy's inability to chase the British out of the Mediterranean became apparent when a British convoy off Calabria bumped into an Italian force that included two battleships and 16 cruisers. The Italian commander broke off the action after one hit on one of his battleships, whereupon the Fascist air force arrived to bomb indiscriminately friend and foe alike, doing little damage to either. Frustrated in the Balkans and at sea, Mussolini ordered his Libyan army to cross the Western desert and conquer Egypt. This adventure soon turned to disaster.

The Eastern front. The end of hostilities in western Europe also provoked a jockeying for position in eastern Europe, where Stalin's fear of the all-conquering Nazis had grown apace. In 1940 Germany signed a pact with Romania for oil and arms transfers. Stalin then forced the Romanian government to hand over Bessarabia and northern Bukovina (June 26, 1940), and annexed Estonia, Latvia (July 12), and Lithuania (August 3) to the U.S.S.R. Hungary and Bulgaria now demanded Romanian territories for themselves, but Hitler intervened to prevent hostilities, lest Stalin see the chance to occupy the Romanian oil fields around Ploieşti. The Treaty of Craiova (August 21) awarded the Southern Dobruja to Bulgaria, and the so-called Vienna Award by Hitler and Mussolini ceded northern Transylvania to Hungary. Romania's King Carol II abdicated in protest, General Ion Antonescu took power, and a German military mission arrived in Bucharest on October 12.

The fall of Romania

The Romanian coup provoked Mussolini's next rash act. "Hitler always faces me with *faits accomplis,*" he raged. "This time I will pay him back in his own coin." On October 13, Mussolini ordered Marshal Badoglio to prepare the long-desired attack on Greece for two weeks hence. He would declare his independence from Hitler and consummate his "parallel war." On Oct. 28, 1940, seven Italian divisions crossed the Albanian border into Greece, provoking Hitler's adjutant to record: "Führer enraged . . . this is revenge for Norway and France." In fact, Mussolini's impetuous attack, combined with the reversals in Africa, would only ensure his humiliation and utter dependence on his northern ally. For the Greek campaign was predictably disastrous, given Italy's bare numerical superiority and lack of planning and equipment, the rough terrain, and the determination of the Greeks. On November 8, General Alexandros Papagos counterattacked, and within a month the Greeks had turned the tables, occupying one-third of Albania. Prime Minister Ioannis Metaxas refused to let the British into Greece for fear of provoking the Germans; indeed, he hoped to drive Italy out of the Balkans before German help might arrive, and to induce Yugoslavia and Turkey to make common cause with Greece against the Fascists.

The Balkan situation seriously interfered with Hitler's evolving continental strategy. Ribbentrop still hoped to persuade him that Britain could be induced to relent through diplomacy, and his last achievement was the Tripartite (or Axis) Pact between Germany, Italy, and Japan on Sept. 27, 1940. Presumably, this alliance would deflect U.S. attention from Europe, threaten the U.S.S.R. with a war on two fronts, and thus drive the British to despair over the prospect of facing Germany alone. But London stood firm, and Hitler grew impatient to get on with his real chore of seizing a Ukrainian empire for the German master race. Upon his return from unsuccessful conferences with Franco at Hendaye (October 23) and Pétain at Montoire (24th), Hitler played host to Soviet Foreign Minister Molotov in Berlin (November 12–14). Though Stalin had meticulously observed his pact with Hitler, their rivalry in the Balkans strained relations. Hitler and Ribbentrop tried to persuade the Soviets to pursue their "natural tendency" to expand in the direction of the Indian Ocean, but Molotov repeatedly interrupted to ask the Germans why they were sending troops to Finland and Romania. These conversations confirmed Hitler's intention to turn his idle military machine to the east. Conquest of the U.S.S.R. might serve now as both means and end, convincing the British of the hopelessness of their situation, allowing Hitler to realize Nazi racist fantasies, and forging a territorial basis for global empire. On December 18 he ordered the army to prepare Operation Barbarossa by May 15, 1941.

The Tripartite Pact

This latest timetable, however, fell victim to Mussolini's folly and the need to secure Germany's flank in the Balkans. German troops entered Romania on Jan. 7, 1941, and Bulgaria on February 27. But Italy's disasters brought into question the very survival of the Fascist regime. Mussolini made Badoglio a scapegoat and in November 1940 issued the first of his pitiful appeals to Hitler to bail him out. At their Berghof meeting on Jan. 20, 1941, Hitler informed Mussolini of his plans to invade Greece. The death of Metaxas in the following days, in turn, led the Greeks to accept a British expeditionary force. Accordingly, Hitler pressured Yugoslavia to permit the passage of German troops, but air force officers in Belgrade staged a coup on March 27 and signed a treaty with Moscow. Furious over such defiance, Hitler ordered a Blitzkrieg for April 6 that broke Yugoslav resistance in five days and overran Greece by the 22nd. Crete then succumbed to a spectacular German airborne assault (May 20–31). Hitler set up puppet regimes in Serbia and "Greater Croatia" and partitioned the rest of Yugoslavia among his client states.

Greece and Yugoslavia

The Balkan campaign postponed "Barbarossa" for six weeks. This did not overly perturb Hitler, who promised his generals victory within a month and denied the need to prepare for cold-weather warfare in Russia. But some generals were skeptical of Blitzkrieg in the vastness of Russia, while others debated whether to force narrow spearheads deep into Russia, emulating the campaign in France, or fight classic battles of envelopment close to the frontier. Hitler's "infallible intuition" dictated the latter, lest his armies, like Napoleon's, be sucked too deep into Russia before enemy forces were destroyed. In the spring of 1941 the Wehrmacht assembled 4,000,000 men—the greatest invasion force in history—including 50 Finnish and Romanian and 207 German divisions armed with 3,300 tanks. They faced a Red Army of some 4,500,000 men and perhaps 15,000 tanks. German success depended heavily on surprise, but preparations of such magnitude could scarcely be hidden. Stalin seemed alive to the danger when he signed a neutrality pact with Japan on April 13 (knowing of Japan's preference for a southern strategy from the espionage of Richard Sorge in Tokyo), then pleaded with Foreign Minister Matsuoka Yosuke: "We must remain friends and you must now do everything to that end." Yet Stalin also redoubled his efforts to assure Hitler of his good intentions and discounted British warnings of a German attack (they had been making such predictions since June 1940, and even the British thought a German strike against Turkey or England more likely). Stalin may also have dismissed the warnings as attempts to poison his relations with Germany. In any case, the Germans achieved complete tactical surprise, while the

Soviets' forward deployments exposed them to the full force of Blitzkrieg.

Invasion of the Soviet Union

The Germans struck on June 22, 1941, along a 2,000-mile-front. Three army groups drove deep into the Soviet Union, occupying vast territories and capturing huge numbers of Soviet troops. But gradually the momentum deserted the invaders. Many myths surround the 1941 campaign. It is said that the Germans were wrong in making for Moscow like Napoleon. But Moscow was of far more military value in 1941 than in 1812; it was the hub of Soviet railroads, communications, and government, and its capture might have crippled the Soviet effort to reinforce the front from the Asian hinterland or have undermined the Communist regime. It is also said that winter defeated the Germans. But they would have had ample time to reach Moscow before winter had they not wasted almost two months in diversions and debate. It is also said that the size of the Soviet Union made swift German victory impossible. But the endless Russian plain actually aided the panzer armies by giving them limitless room to maneuver and form the huge pockets that cost the Red Army 2,500,000 men in the first six months. What did stop the Germans was their own dilatoriness, the mud and unpaved roads, their underestimation of Soviet reserves and resilience, and the Nazis' own brutality, which alienated a population otherwise hostile to Stalinism.

By December 1941 the German invasion of the U.S.S.R. and the latter's survival had confirmed precisely that British hope which Hitler had meant to quash. The entry into the war of the United States that same month made German defeat virtually certain—and also brought to a close the last purely European war.

ORIGINS OF AMERICAN BELLIGERENCE

From neutrality to active aid. The outbreak of war brought a swift change of mood to the United States. While isolationism was still widespread, the vast majority of Americans were sympathetic to Britain, and Roosevelt did not follow Wilson in asking Americans to be neutral in thought as well as deed. Instead he set out to lead public opinion and gradually expand his ability to aid the Allies. On Sept. 21, 1939, his brilliant speech to Congress laid the groundwork for passage of the Pittman Bill, which became law on November 4 and repealed the arms embargo on belligerent nations. Henceforth, the United States might trade with Britain and France, but only on a "cash and carry" basis. Senator Arthur Vandenberg rightly noted that the United States could not "become the arsenal for one belligerent without becoming the target for another." Still, the President made clear to Churchill (with whom he struck up close relations by correspondence) his desire to aid Britain in every way consonant with the American mood. Only once did Roosevelt make a feint at mediation: In March 1940 he sent Undersecretary of State Sumner Welles to Europe on a fact-finding mission that revealed "scant immediate prospect" of peace. When Hitler's Western offensive followed, even that dubious prospect disappeared, and Churchill assured his House of Commons that Britain would fight on "until, in God's good time, the New World, with all its power and might, steps forth to the rescue and liberation of the Old."

Cash and carry

In January 1940, Roosevelt asked for a mere $2,000,000,000 in defense spending, a slight increase over the year before. But the fall of France pushed the pace of U.S. rearmament up to $10,500,000,000 by September. Opinion polls showed the American public heavily favouring a policy of "all aid short of war" to Britain. On May 15, Churchill sought to capitalize on the shifting sentiment with an emergency request for 40 or 50 overage destroyers with which to counter German U-boats. Roosevelt hesitated because of the legal complications, while continuing his efforts to shape opinion by encouraging William Allen White's Committee to Defend America to foster the idea that "Between Us and Hitler Stands the *British* Fleet!" On September 2 the United States transferred 50 warships to Britain in return for long-term leases on British naval bases in the Western Hemisphere. Despite Roosevelt's public relations, isolationist sentiment remained strong. On September 4 the America First Committee arose to challenge Roosevelt's deceptive campaign for intervention, and Wendell Willkie charged during the presidential campaign that Roosevelt's reelection would surely mean war. The president responded that "your boys are not going to be sent into any foreign wars," gliding over the fact that if the United States were attacked, it would no longer be a foreign war.

Lend-lease

The next step in U.S. involvement stemmed from Churchill's warning of Dec. 9, 1940, that Britain was near bankruptcy. Roosevelt responded with lend-lease, a plan to "eliminate the dollar sign" by lending, not selling, arms. If your neighbour's house is on fire, he argued, you do not sell him a hose, you lend it to him until the fire is out. "If Great Britain goes down," he warned, "all of us in the Americas would be living at the point of a gun. . . . We must be the great arsenal of democracy." Churchill added his own ringing appeal on Feb. 9, 1941: "Give us the tools and we will finish the job." Willkie asked Republicans to back lend-lease, which became law on March 11.

Unknown to the public, Roosevelt authorized joint U.S.–British staff talks. The two countries also collaborated on how to meet the U-boat menace. Admiral Karl Dönitz's wolfpack technique, by which eight to 10 U-boats would strike a convoy from the surface at night (thereby avoiding the British Anti-Submarine Detection Investigation Committee device [ASDIC sonar]), cost the British and Americans 320,048 tons of shipping in January 1941 and 653,960 tons in April. American Admiral Harold R. Stark considered the situation "hopeless except as [the United States] take strong measures to save it." In Hemispheric Defense Plan No. 1 (April 2) Roosevelt authorized the navy to attack German submarines west of 25° longitude and by executive agreement with the Danish government-in-exile placed Greenland under American protection (April 9). U.S. marines also occupied Iceland in July.

The German invasion of the Soviet Union posed the problem of whether to extend lend-lease to the U.S.S.R. Only 35 percent of Americans polled favoured underwriting the Communist regime, but Roosevelt, supporting his acting secretary of state, Sumner Welles, said "Of course we are going to give all aid we possibly can to Russia," on the theory that anything that contributed to the defeat of Germany enhanced the security of the United States. Aid to the Soviet Union began in July, and a formal agreement followed on August 2. But the initial supplies were too meagre to affect the battles of 1941. Roosevelt meanwhile pressed for amendments to the Selective Service Act to remove the ceiling of 900,000 men on U.S. armed forces and the ban on use of troops beyond the Western Hemisphere and to permit the president to retain draftees in service. This provoked the last great Congressional debate on isolationism versus interventionism; the House passed the bill by a single vote on August 12.

The Atlantic Charter

It was during this debate that Roosevelt and Churchill met secretly off the coast of Newfoundland and drafted a manifesto of the common principles that bound their two countries and all free peoples. In this eight-point Atlantic Charter (announced on August 14), reminiscent of Wilson's Fourteen Points, the signatories renounced territorial aggrandizement and endorsed the restoration of self-government to all captured nations and equal access to trade and raw materials for all. According to Churchill, Roosevelt also promised to "wage war but not declare it" and to look for an incident that would justify open hostilities. When the Congress voted on November 7 to arm merchant ships and allow them into the war zone, it seemed that submarine warfare would again be casus belli for the United States. U-boats had already torpedoed the destroyers *Kearney* and *Reuben James* (the latter was attacking the submarine, but sank with 115 hands on October 31). But in fact it took dramatic events in another theatre altogether to make Roosevelt's undeclared war official.

Japan's challenge. When war broke out in Europe, the Japanese occupation of China was nearing its greatest extent, and there was no sign of Chinese capitulation. Japan was understandably incensed when its ally in the Anti-Comintern Pact, Germany, joined with Moscow at a time when the Japanese were fighting the Soviets in

Manchuria and Mongolia. On the other hand, the German victories of 1940 made orphans of the French and Dutch colonies in Southeast Asia, including mineral-rich Indochina and oil-rich Indonesia. These sources of vital raw materials were all the more tempting after the United States protested Japan's invasion of China by allowing its 1911 commercial treaty with Japan to expire in January 1940. Thereafter trade continued on a day-to-day basis while U.S. diplomacy sought peaceful ways to contain or roll back Japanese power. But the territorial and trade hegemony that Japan would come to term the "Greater East Asia Co-Prosperity Sphere" in 1941 increasingly appeared to be a cover for brutal imperialism and exclusionist trade policies. In June 1940, as France was crumbling, Japan insisted that the new Vichy regime cut off the flow of supplies to China over Indochinese railways. The beleaguered British, fearful of simultaneous war in Asia and Europe, also agreed to close down the Burma Road to China for three months, isolating Chiang Kai-shek.

Toward Asian hegemony

Japanese militarists then arranged a new government in Tokyo under the weak Konoe Fumimaro, expecting that Foreign Minister Matsuoka and War Minister Tōjō Hideki would dominate. On July 27 the Cabinet decided to ally with the Axis and strike into Southeast Asia even as it sought to resume normal trade with the United States.

Japanese assertion posed a dilemma for Washington. Secretary of War Henry Stimson and Treasury Secretary Henry Morgenthau, Jr., believed an embargo on oil and scrap iron would cripple the Japanese war machine, but Secretary of State Cordell Hull feared an embargo would provoke Japan into seizing Southeast Asia. On July 26, 1940, after lengthy debate, the United States banned export of high-grade scrap iron and aviation fuel to Japan. On August 1, Japan forced Vichy to permit a limited occupation of northern Indochina, and the following month it signed the Tripartite (Axis) Pact in which Germany, Italy, and Japan pledged aid to each other should any be attacked by a power not at present involved in the Pacific War (*i.e.,* the United States). But this act of defiance only stoked American indignation. In November, Roosevelt approved a loan of $100,000,000 to the Nationalist Chinese and began to allow American pilots to volunteer for Chinese service in Claire Chennault's Flying Tigers. In December and January all forms of iron, copper, and brass were added to the embargo.

Civilian government had eroded in Japan until censorship, propaganda, and intimidation overwhelmed moderates and placed policy in the hands of militarists devoted to traditional Japanese exclusivism, xenophobia, and the Bushidō code of combat. Of the latter mentality Americans had barely a clue, just as the Japanese looked upon Western notions of self-determination and the Open Door as so much hypocrisy. But although reciprocal misunderstanding and racialist thinking inhibited the quest for peace in the Pacific, Japan's determination to carve out an Asian empire was clearly the source of the crisis, while American policy was essentially reactive.

The latest U.S. trade restrictions sparked the final peace initiative of the moderate faction composed of Konoe and leading Japanese industrialists. Two American Catholic missionaries served as intermediaries for an alleged Japanese offer to evacuate China and break the Tripartite Pact in return for normal trade with the United States. This was exactly what Roosevelt wanted, and he urged that the offer be placed in writing. A new Japanese ambassador, Nomura Kichisaburo, then arrived in Washington and met privately with Hull 40 times after March 1941. On April 9 the Catholic missionaries delivered a written offer, but it contained no promise of troop withdrawals and instead asked the United States to cut off aid to China. Hull clearly informed Nomura that any accord must be founded on four principles: respect for territorial integrity, noninterference in the internal affairs of other countries, commercial equality, and respect for the status quo in the Pacific. Nomura unfortunately failed to understand and reported that the United States had accepted the April 9 proposal. The Tokyo Cabinet then drafted an even tougher note as a basis for negotiation, prompting Hull to conclude that the Japanese were incorrigible.

Japanese-U.S. diplomacy

Meanwhile, the Japanese military debated the merits of a northern advance against the Soviet Union's maritime provinces or a southern advance against the French, Dutch, and British colonies. The Russo-Japanese neutrality pact of April 1941 indicated a southern advance, but the German invasion of the Soviet Union indicated a northern one. The course of the war—and the survival of the U.S.S.R.—hung in the balance. Heretofore, Hitler had been at pains to keep Japan out of his Soviet sphere of influence, but at the height of German success in the Soviet Union, Hitler suggested to Ambassador Oshima Hiroshi that the two join forces to liquidate the Soviet empire, a plan endorsed by Matsuoka. If Hitler meant it, he was too late, for the Cabinet in Tokyo decided again after the invasion of the Soviet Union (June 22) to exploit German victories rather than take part in them. The Japanese army and navy would move south and establish the Greater East Asia Co-Prosperity Sphere. The Emperor endorsed the plan on July 2, and the Americans, having broken the Japanese code with the MAGIC process, knew of the decision at once. On July 26, Japan occupied all of French Indochina, and the United States impounded Japanese assets. On September 5, Hull sanctioned a complete embargo on petroleum.

Japan now faced a choice of abandoning all the conquests made since 1931 or seizing the necessary war matériel to defend its empire. Konoe tried desperately to reverse the tide and requested a summit meeting with Roosevelt. But Roosevelt, on Hull's advice, insisted on prior Japanese acceptance of the four principles. Konoe was obliged on September 7 to make a deal with his militarists: He could try once more for an agreement, but if the United States did not relent by early October, Konoe would then support the military solution. When the deadlock was confirmed Konoe in fact resigned on October 16, and Tōjō became prime minister. The veteran diplomat Kurusu Saburo then flew to Washington with two final options, Plan A and Plan B. The latter held out some hope, since in it Japan at least promised to make no military moves to the south. But MAGIC deciphered a cable revealing the secret deadline of November 29, while the British, Dutch, and Chinese vetoed any modus vivendi that left Japan a free hand in China. On November 27, American warnings of war were dispatched to the Pacific, and on December 1 a Japanese Imperial conference ratified Tōjō's conclusion that "Japan has no other way than to wage war . . . to secure its existence and self-defense."

Plan A and Plan B

The final diplomatic exchanges were superfluous, but they included a 10-part American note of November 26 and Roosevelt's personal appeal to the Emperor on December 6. That same day a 13-part Japanese reply arrived in Washington, which MAGIC deciphered even before the Japanese embassy did. That war was imminent was clear; where the first blow would fall was not. On Sunday, December 7, a 14th part arrived, which the Japanese embassy was slow in translating and typing. By the time the diplomats arrived at Hull's office at 2:00 PM, news of the treacherous attack on Pearl Harbor, Hawaii, had already arrived. Hull delivered his opinion of Japanese diplomacy in vitriolic terms and told the ambassadors to get out. The following day Roosevelt named it "a day which will live in infamy" and asked Congress for a declaration of war.

Revisionist historians have argued that Roosevelt should have known of the danger of Japanese attack from the secret intercepts and reports of Japanese fleet movements, or that he did know and purposely suppressed the information so that the United States might enter the European war, unified and irate, "through the back door." To be sure, American blunders marked the final years of neutrality, and a cover-up of those blunders may have occurred. But certainly no one forced the Japanese to make a direct attack on U.S. territory, nor did anyone expect an attack so bold as that on Hawaii. Nor did the Congress even take that opportunity to enter the European war. That was accomplished on December 11, when Hitler and Mussolini, honouring the Tripartite Pact, declared war on the United States. Hitler considered the "half-Judaized and half-negrified" Americans to be of little military account, especially since, he believed, the Japanese war would prevent U.S.

The Axis declares war

intervention in Europe. His gratuitous declaration of war was in fact a folly surpassing Ludendorff's provocations of the United States in 1917.

Japan's war plan was marked by operational brilliance but strategic folly. The notion that Japan could take on the British Empire and the United States at the same time, and win, was the equivalent (in the Japanese simile for courage) of "jumping with eyes closed off the veranda of Kiyomizu Temple." Still, Admiral Yamamoto devised a bold campaign to destroy Allied striking power for the foreseeable future, whereupon the Americans would presumably sue for peace. He assigned all six of his aircraft carriers to a surprise attack on the U.S. Navy base at Pearl Harbor. The rest of the navy—eight battleships, four auxiliary carriers, 20 cruisers, and 112 destroyers—was earmarked for the south, together with 11 infantry divisions and 795 planes. The first force struck at dawn, its dive-bombers penetrating Pearl Harbor's defenses through the mountain passes of Oahu. They sank four of eight U.S. battleships, damaged four others, sank or disabled 10 other ships and 140 planes, and killed 2,330 troops. By chance, the three U.S. aircraft carriers were at sea and escaped destruction. A second Japanese force destroyed 50 percent of the U.S. aircraft in the Philippines, landed on Luzon on December 10, took Manila on Jan. 2, 1942, and drove the remaining U.S. and Filipino forces into redoubts on the Bataan Peninsula and Corregidor Island. The Japanese also bombed Hong Kong on December 8, took the British outpost from the mainland on the 25th and occupied Bangkok on December 9 and southern Burma on the 16th. Most damaging to the British were the Japanese landings in Malaya after December 8 and the advance through the jungle to Singapore. This mighty fortress, considered impregnable, was the keystone of British strategy in Asia, and Churchill had ordered out the battleship *Prince of Wales* and battle cruiser *Repulse* in the expectation of intimidating the Japanese. Instead, Japanese aircraft sank the two ships on December 10. On Feb. 9, 1942, three Japanese divisions overran Singapore, whose defenses were directed seaward, and captured the 90,000-man force. The fall of Singapore crippled British communications and naval power in Asia.

Pearl Harbor

The fall of Singapore

Supporting the assault on the Philippines, the Japanese bombed Wake Island on December 8 and overcame fierce resistance from the tiny U.S. garrison on December 23. By February 10, Guam and Tarawa in the Gilberts and Rabaul and Gasmata on New Britain were occupied. Japan was now master of a vast empire stretching from Manchuria to the East Indies and the border of India deep into the western Pacific.

THE TURNING POINT, 1942

Within a year after American entry into the war Axis power crested and began to ebb, for critical battles were fought in 1942 in every major theatre. The year also saw the forging of a Grand Alliance among the United States, Britain, and the U.S.S.R. and the first sign of disagreement on strategy and war aims.

After Pearl Harbor, Churchill requested an immediate conference with Roosevelt. The two met for three weeks at the Arcadia Conference in Washington after Dec. 22, 1941. They reaffirmed the "Europe first" strategy and conceived "Gymnast," a plan for Anglo-American landings in North Africa. They also created a Combined Chiefs of Staff Committee and issued, on Jan. 1, 1942, the United Nations Declaration in the spirit of the Atlantic Charter. But Sir Anthony Eden had traveled to Moscow in late December and returned with troubling news: Stalin demanded retention of all the territory gained under the German–Soviet Nonaggression Pact and grumbled that the Atlantic Charter was apparently directed against him, not Hitler. The Soviets also first made what was to become their incessant demand that the Allies open a second front in France to take the pressure off the Red Army. Roosevelt sent Army Chief of Staff George C. Marshall to London to argue for a cross-Channel invasion by April 1943, but the British deemed it impossible. London reassured Molotov by concluding an Anglo-Soviet alliance (May 26, 1942) to last for 20 years. In late June, Churchill and Roosevelt met again in Washington, D.C., and confirmed plans for a joint operation in Africa despite the misgivings of American generals, who suspected the British of being more concerned for the defense of their empire than the rapid defeat of Hitler. In the end the British won, and on July 25 the Allies approved the renamed operation "Torch"—a combined invasion of North Africa planned for the autumn. Churchill then traveled to Moscow in August 1942, where Stalin berated him for postponing the second front and suspending Arctic convoys because of German naval action. Despite his suspicions and fears, Stalin could take grim satisfaction from the events of 1942, for by December of that year the German advance into the Soviet Union had been stopped, though at enormous cost.

The Allied landings in North Africa, where British forces had finally turned back General Erwin Rommel's Afrika Korps at el-Alamein, were targeted for Casablanca, Oran, and Algiers. (Hence, the first American initiative in the war was to be an unprovoked and undeclared attack against neutral territory.) Vichy France promptly severed diplomatic relations with Washington and ordered French forces in North Africa to resist. Brief but serious fighting resulted at Oran and Casablanca. The allies had been seeking a French leader with the prestige and willingness to rally French Africa against the Axis, but the nominal commander was Admiral François Darlan, an ardent collaborationist in the Vichy Cabinet. The Allies preferred General Henri Giraud, a heroic escapee from a prison camp, but he insisted on being given command of the whole Allied invasion force. When Darlan surprisingly turned up in Algiers, U.S. Ambassador Robert Murphy negotiated a deal whereby Eisenhower recognized Darlan as political chief of North Africa in return for Darlan's ordering French forces to cease resistance. The Americans soon escaped the embarrassment of having bargained with a leading Fascist when a French royalist shot Darlan on December 24. De Gaulle was able to outmaneuver the vain but inept Giraud to become de facto leader of Free French forces.

Invasion of North Africa

In the Pacific, the naval Battle of Midway in June, the landing of U.S. forces on Guadalcanal in August, and the creation of an "island-hopping" strategy against Japan's sudden and far-flung empire similarly blunted the string of the Axis' early victories. Meanwhile, General Douglas MacArthur rallied Allied forces in Australia in anticipation of fulfilling his departing promise to the Filipinos: "I shall return." A Japanese invasion force landed near Gona at the southeastern end of New Guinea in July 1942 and drove Australian troops back to within 32 miles of Port Moresby. But MacArthur executed a series of landings behind the Japanese and secured the entire Papuan coast by late January 1943. Thenceforth Japan, too, went on the strategic defensive.

THE ECONOMIC AND SCIENTIFIC WARS

How could the Axis powers have imagined that they might win the war, given their narrow base of land area, population, and production, and the size and strength of the enemies they themselves forced into the war? The answer was Blitzkrieg, which involved more than simply a set of tactics for mobile combat but was rather an encompassing theory of total war. The theory posited a strategically mobilized and organized economy meant to avoid a repetition of the war of attrition that wore Germany down in 1914–18. By overrunning their neighbours one by one in swift assaults, the Germans constantly added to their own manpower and resource base while shrinking that available to the enemy. In addition, armament in breadth rather than depth provided the flexibility necessary to shift production from one set of weapons to another depending on the needs of the next campaign, and it permitted constant innovation of weapons systems. Most tellingly, Blitzkrieg shifted the burdens of war from Germany to the conquered peoples. By June 1940 the British were unable to budge a Nazi empire that drew on the resources of the entire continent. But Hitler also realized by late 1940 that all the resources of America would eventually be made available to Britain; hence his decision to break the stalemate by unleashing Blitzkrieg against the Soviet

Union. Soviet survival, however, turned the Blitzkrieg into a gigantic war of attrition after all, one in which Germany could never prevail.

The German economy and the Jews. Cut off from foreign sources of capital, Germany paid for World War II through taxes and ruthless exploitation of occupied regions. Levies on conquered peoples amounted to 40 percent of the income raised by internal taxation, and 42 percent of that tribute came from France. The number of slave labourers deployed by various arms of the regime peaked at 7,100,000 in 1944; this figure included prisoners of war and "racial enemies" condemned to slavery until death in SS camps.

Seen only in cold economic terms, Nazi genocide against Jews and other groups, racially or ideologically or otherwise defined, was the height of irrationality. As early as January 1939 Hitler gave vent to his pathological hatred and fear of the Jews before the Reichstag: "If the international Jewish financiers . . . succeed in plunging the nations once more into a world war the result will be the obliteration of the Jewish race in Europe." The war gave Hitler the opportunity to seek a "final solution." In 1939–40 the Nazis considered using Poland or Madagascar as dumping grounds for Jews. But the invasion of the U.S.S.R. emboldened Hitler, Göring, and SS leaders Heinrich Himmler and Reinhard Heydrich to decide instead on mass extermination in camps at Belzec, Majdanek, Sobibor, Treblinka, and Auschwitz. Large numbers of SS troops, as well as railroads and rolling stock, were absorbed in capturing, transporting, and putting to death as many as 12,000 Jews per day. The total by war's end would reach 6,000,000, almost half from Poland, and some 2,000,000 others including Gypsies, clergy, Communists, and other resisters. SS troops accompanied the regular army into the Soviet Union in 1941 and made racial war on the Slavs as well in order to prepare the farmlands of the Ukraine for German settlement.

The death camps

News of the Holocaust reached the West slowly but surely, although Auschwitz was able to keep its monstrous secret for more than two years after the first gassings in May 1942. Richard Lichtheim of the Jewish Agency in Geneva served as a conduit for information about what was occurring in Nazi Europe, but his and others' efforts to promote action on the part of the Allies broke against political and practical barriers. The British, worried by the prospect of Arab revolt, limited Jewish emigration to Palestine, while quotas elsewhere in the world meant that even those Jews who managed to escape Europe sometimes had nowhere to go. Reports appearing in Western newspapers inspired the Allies to make a declaration on Dec. 17, 1942, condemning "this bestial policy of cold-blooded extermination," and on Jan. 22, 1944, Roosevelt established a War Refugee Board "to forestall the plan of the Nazis to exterminate all Jews and other minorities." But the Allies were unable to take direct action of any sort until the capture of Italy brought Allied bombers within range of the camps. Jewish leaders were then misled by hints that the Germans might negotiate about the Jews. Finally, after June 1944, when escapees confirmed the existence and nature of Auschwitz, the World Jewish Congress requested bombing of the gas chambers. But the Allied Bomber Command judged that its efforts should be directed only at military targets and that the best way of helping the Jews was to hasten the defeat of Nazi Germany.

Strategic bombing. Allied strategic bombing was the most deadly form of economic warfare ever devised and showed another side of the indiscriminateness of industrial war. But in mid-1941 the British Chiefs of Staff soberly concluded that morale, not industry, was Germany's most vulnerable point and ordered Sir Arthur Harris of the RAF Bomber Command to concentrate on "area bombing" of cities. Churchill's scientific adviser Professor L.A. Lindemann of Oxford (later Lord Cherwell) concurred in April 1942 that one-third of all Germans could be rendered homeless in 15 months by strategic bombing of cities. The Royal Air Force accordingly assigned its new Lancaster four-engine bombers to a total war on German civilians. After attacks on Lübeck and the Ruhr, Harris sent a thousand planes against Cologne on May 30–31 in an attack that battered one-third of the city. In 1943, after an interlude of bombing German submarine pens, the Lancasters launched the Battle of the Ruhr totaling 18,506 sorties and the Battle of Hamburg numbering 17,021. The fire raids in Hamburg killed 40,000 people and left a million homeless. The Royal Air Force then hit Berlin (November 1943 to March 1944) with 20,224 sorties, avenging many times over all the damage done by the Luftwaffe to London.

Area bombing of Germany

By early 1943 the U.S. 8th Air Force joined in the air campaign but eschewed terror bombing. Its B-17 Flying Fortresses and B-24 Liberators conducted daylight precision bombing of industrial targets. As a result, they suffered heavy losses that climaxed in October 1943 over the Schweinfurt ball-bearing plants, when the United States lost 148 bombers in a week. The Army Air Forces suspended daylight sorties for months until the arrival of a long-range fighter, the P-51 Mustang. Bombing then resumed and concentrated on the German oil industry, creating a serious shortage that virtually grounded the Luftwaffe by the time of the D-Day invasion. The effectiveness of strategic bombing is a subject of great debate, since German war production actually increased over the years 1942–44. German engineers became masters at shielding equipment, restoring it to operation in a matter of days, or even moving plants underground. Nor did the German people crack under British devastation of their towns and homes. But the air offensive did force the Germans to divert as many as 1,500,000 workers to the constant task of rebuilding and established the Allied mastery of the air that permitted the success of the Normandy landings.

Allied economic management. Britain was only in the early stages of rearmament when the war broke out, but after the fall of France the transition to a World-War-I-type command economy was precipitous. Churchill replaced some 60 interdepartment committees for war economics with the single Lord President's Committee under Sir John Anderson. Within 18 months Anderson organized the most centralized and complete war mobilization of any nation. It included controls on trade, foreign exchange, wages and prices, and raw materials. The National Service Act of December 1941 outdid even the U.S.S.R. by making every man under 50 and every woman under 30 liable to government assignment. Of the 2,800,000 new war workers, 79 percent were female. The state also cut consumer production to a minimum: 67 percent of the work force was employed in war-related jobs. Once again, the British exercised financial responsibility by raising taxes, deferring wages, and compelling savings.

The Lord President's Committee

Even before the war, and despite the Depression, the American gross national product (GNP) of $88,600,000,000 dwarfed that of any other country. Under the impulse of war it increased by 1944 to $135,000,000,000, of which 40 percent was directed to military purposes. About 60 percent of all the munitions used by the Allies in 1944 was made in the United States. In addition to arming its own immense air and sea forces, the United States provided $32,500,000,000 in lend-lease support, including $13,500,000,000 to Britain and $9,000,000,000 to the U.S.S.R. Total U.S. production included 300,000 aircraft, 51,400,000 tons of shipping, 8,500,000 tons of warships, and 86,700 tanks. The government financed this phenomenal buildup largely through war bonds in the early years and later through taxation.

The American war effort was also achieved without the rigid centralized control of Britain. In January 1942 the War Production Board emerged, staffed with "dollar-a-year" volunteers from business, while the Office of War Mobilization (May 1943) under James F. Byrnes served less as a dictator than an umpire in matters involving labour, business, and the military.

The War Production Board

The Soviet Union also made a stupendous economic effort in the war despite conditions as difficult as the American ones were favourable. Within a few months in 1941 the U.S.S.R. lost to the enemy over half its industrial capacity and richest farmland and countless skilled workers. Yet the Soviets rebounded quickly, relocating over 1,300 factories to the Urals region in an effort that

involved perhaps 10,000,000 people. Coal, oil, electricity, and food never regained prewar levels, but arms production boomed. The Soviets managed to turn out 136,800 aircraft and 102,500 tanks by 1945, surpassing the Germans in both. The centrally directed Gosplan and party apparatus, of course, had initiated a ruthless command economy as early as 1928, and Soviet appeals to patriotism (as opposed to Marxism), the network of forced-labour camps, and severe austerity made the effort possible. Despite punishing taxation and subsistence wages (40 percent of the 1940 level) state income covered only half the budget over 1941–45, laying the basis for the inflation that would lead to postwar devaluation. But the Soviet war economy, like that of the United States, prepared the country for postwar superpower status.

Japan's strategy was similar to Germany's Blitzkrieg in that the swift conquest of isolated territories was designed to create a self-sufficient empire capable of withstanding any blow from without. Once again, precise operational planning permitted Japan to increase weapons production steadily from the inception of a full war economy in 1942 to early 1945, when U.S. bombing intensified. By 1944 naval ordnance production was over five times that of 1941, and aviation more than four and a half times. The Japanese, like the Nazis, exploited their conquered peoples and even more than the Nazis subjected prisoners of war to slavery or death. But the fact that attacking Pearl Harbor would "awaken a sleeping giant" was lost on Japanese planners. By 1944 military expenditures absorbed 50 percent of the Japanese GNP, a degree of concentration second only to that of the Soviet Union. Yet the United States, with half its effort diverted to Europe, still overwhelmed the Co-Prosperity Sphere.

Science and technology in wartime. Of the many wartime innovations, those in macroeconomics and management techniques were among the most important, for the rapid increase achieved in labour productivity would make possible the economic miracles of many nations after the war as well. U.S. merchant vessels that took 35 weeks to build before the war were being launched in 50 days by 1943. The Soviet Ilyushin II-4 airplane absorbed 20,000 man-hours before the war and 12,500 in 1943. By the end of the war the British government was choosing contractors on the basis of management, rather than technical, experience. The industrial world was reaching a new plateau of efficiency.

Planned research and development

World War II was unprecedented in the fillip it delivered to science and technology and the maturation of planned research and development (R and D). What Churchill called "the wizard war" between scientists to devise new weapons and electronic countermeasures for air and sea combat began before 1939 in the R and D laboratories of German and British firms and institutes. The Soviet Union had since 1919 made the "scientific pursuit of science" a pillar of the regime, and the 1,650,000,000 rubles budgeted for R and D in 1941 was far and away the largest effort in the world. The Fascist regimes also made a fetish of technological progress. Mussolini established a National Council of Research in 1936 under the famed radio pioneer Guglielmo Marconi. Hitler took for granted the preeminence of German science, and showed a lively interest in new weapons technology. But the totalitarian regimes' insistence on "communist science" or "Fascist science," their secrecy, persecutions, and suppression of intellectual freedom meant that their R and D investment yielded less than that of the liberal states. Stalin's fear that technical experts might turn to political opposition led him to consign thousands of scientists and engineers to the *gulag,* where they worked under the eye of the secret police. Nazi persecution chased dozens of brilliant Jews and others (especially nuclear physicists) out of Europe, thereby enriching the brainpool of Britain and the United States. The dictators' personal interventions in matters of weapons research and deployment, while sometimes breaking bottlenecks and ending jurisdictional feuding, more often skewed the work of scientists in less productive or dead-end directions. In short, World War II made planned R and D a permanent and mighty tool of state power while demonstrating that too much state control or ideological content in research inevitably brought diminishing returns.

The liberal states, by contrast, responded quickly and effectively to the scientific challenge. Nowhere was this more evident than in cryptanalysis and espionage, in which the Allies repeatedly bested the otherwise secretive and devious Axis. As early as 1931, Captain Gustave Bertrand of French intelligence procured documents from a German traitor concerning the cryptographic rotor device Enigma. A brilliant Polish mathematician, Marian Rejewski, cracked Enigma by 1938, only to have the unsuspecting Germans add two rotors to the machine. Britain's scientists in the ULTRA project then worked on methods to generate keys for Enigma until they devised the cumbersome Colossus machines that some consider the first electronic computers. ULTRA not only compromised every German spy in Britain but also provided the British with decryptions of German directives and deployments for the whole of occupied Europe for the entire war.

The ULTRA project

Following the Battle of Britain, to which radar made such a vital contribution, Churchill established a Scientific Advisory Committee under L.A. Lindemann. He and his rival Sir Henry Tizard helped to direct the research programs that discovered various means of jamming the German bombers' radio navigation system. By autumn 1940 the Germans countered with their X-Gerät that broadcast its signal on several frequencies, but this was overcome in turn by British airborne radar that allowed fighters to home in on bombers individually. A similar contretemps occurred in the air battles over Germany and inspired the development of devices that guided night bombers to their targets despite jamming, the H_2S system that permitted crews to "see" through cloud cover, and the use of billows of aluminum strips dropped from bombers to confuse German radar. Microwave radar helped search planes locate submerged U-boats after March 1943.

Roosevelt entrusted the American effort to Vannevar Bush's Office of Scientific Research and Development (OSRD), which channeled contracts of $1,000,000 or more to over 50 universities during the war. The OSRD, the Naval Research Laboratory, and army arsenals produced such innovations as the antitank "bazooka" rocket, the proximity fuse, the DUKW amphibious vehicle, the first use of DDT to combat malaria, and mass production of the antibiotic penicillin for war wounds (1943). Soviet researchers, despite the handicaps imposed by invasion and their own regime, developed the devastating Katyusha rocket-cluster (called "Stalin organs"), the sturdy T-34 tank, and, by war's end, a prototype jet fighter. The Germans eased their shortages of vital materials through processes for coal gasification (5,700,000 tons worth in 1943) and for producing synthetic rubber. They were also first with an operational combat jet aircraft, the Me-262, but the Nazi regime instead chose to allocate steel and fuel to submarines, ending any chance that Germany might regain control of the skies.

New weapons

The four technological developments that would come to define the postwar strategic environment were radio-electronics, the electronic computer, the ballistic missile, and the atomic bomb. The medium-range ballistic missile A4 (called the Vengeance weapon, V2 by Goebbels) was the brainchild of German rocket engineers who had first come together as amateur spaceflight enthusiasts in the 1920s. The German army began funding their research in 1932 and built a large test range at Peenemünde after 1937. There, Commander Walter Dornberger and Chief Engineer Wernher von Braun developed and tested the A4 by 1942. But the program did not receive top priority until 1943, at which time a British air raid on Peenemünde forced construction of an underground factory in the Harz Mountains to construct the rockets. The V2s, of which 4,300 were fired (half of them at Antwerp) after September 1944, did considerable damage until the Allies captured the launch sites in The Netherlands.

The V2 rocket

Nuclear physics had advanced to the point by 1938 that the German physicists Otto Hahn and Fritz Strassmann were able to demonstrate nuclear fission. Scientists in Britain, France, Germany, the U.S.S.R., and the United States all speculated on the possibility of building an

atomic explosive device, and in 1939 Albert Einstein wrote to President Roosevelt personally, urging a crash program to perfect such a bomb before the Nazis. The resulting Manhattan Project absorbed $2,000,000,000 of the $3,850,000,000 spent by the United States on R and D in World War II. Churchill, too, approved a nuclear program, code-named the Directorate of Tube Alloys, in Britain's dark days of 1941. But by 1943 the Americans had built up a sizeable lead and agreed at the Quebec Conference to share results with the British. German atomic research depended on heavy water from Norway, but British commandos and the Norwegian underground sabotaged the plant in 1943. The scientists also failed to press for top priority, which went instead to the missile program. Soviet atomic research kept abreast of the West until the invasion, and in June 1942, Stalin authorized a crash program that by war's end had begun to produce fissionable uranium in quantity. In no country was much official thought apparently given to the moral and long-range consequences of this potentially devastating invention.

The Manhattan Project

A final, though lesser known, scientific breakthrough of World War II was the application of methods from the physical and social sciences to problems of production, logistics, and combat. Known as "operational research," this application of science to practical problems was a major step in the process by which military men in the 20th century lost primacy in their profession to civilian specialists. Whether in the scientific study of various antisubmarine tactics, the selection of targets for strategic bombing, or the optimal size and pattern for naval convoys, operational research completed the mobilization by governments of the world's intellectual community.

STRATEGY AND DIPLOMACY OF THE GRAND ALLIANCE

Allied strategy to the fall of Italy. In the wake of Operation "Torch," Roosevelt and Churchill met at Casablanca (January 1943) to determine strategy for the coming year. Once again Roosevelt conciliated Churchill, agreeing to put off opening a second front in France in favour of more modest operations against Sicily, Italy, and the "soft underbelly" of Europe after the liberation of North Africa. General George Marshall and Admiral Ernest King succeeded in winning approval for offensives in Burma and the southwest Pacific. The French rivals, de Gaulle and Giraud, were persuaded at least to feign unity and later to create a French Committee of National Liberation under their joint chairmanship (May 1943). But the main event was Roosevelt's parting announcement that "peace can come to the world only by the total elimination of German and Japanese military power . . . (which) means unconditional surrender." This surprise declaration was not spontaneous, as Roosevelt claimed; it was a considered signal to Stalin of Allied resolve, especially necessary after General Eisenhower's ignominious "Darlan deal." But it also rashly committed the United States to a power vacuum, rather than a balance of power, in postwar Europe, and may have discouraged Germans from attempting to oust Hitler in hopes of escaping utter defeat.

Stalin's reaction to Casablanca was predictably sour. In March he expressed great anxiety about repeated postponement of the second front in France. On the other hand, the Battle of Stalingrad had more or less assured eventual Soviet victory. Would it not have served Soviet interests more to delay the Allied presence in Europe as long as possible? It is likely that Stalin's continued pressure for a second front was a function of his perennial fears for internal Soviet security. Stalin may have wanted to recapture his lost ground, especially the Ukraine, as quickly as possible lest anti-Soviet movements take hold there or in neighbouring countries. At this time Stalin also began to denounce the London Poles as reactionaries and sponsored a new Union of Polish Patriots in Moscow as a rival government-in-exile. The final breach between the London Poles and Stalin followed in April 1943, when the Germans uncovered a mass grave in the Katyn forest containing the corpses of over 4,000 Polish officers captured by the Russians in 1939. (Another 10,000 Polish officers were killed in Soviet secret police concentration camps.) Churchill advised Władysław Sikorski, prime minister in the London government-in-exile, not to pursue the issue out of deference to Stalin, who blamed the massacre on the Germans. But the Poles invited an International Red Cross investigation that strongly suggested the Soviets had committed the crime in the spring of 1940, presumably to exterminate Poland's non-Communist leadership class. Stalin's seemingly benign dissolution of the Comintern in May 1943 was likewise inspired by postwar planning. The party purges and the assassination of Trotsky in Mexico (August 1940) placed foreign Communists so securely under Moscow's thumb that the formal apparatus of control was no longer needed, while the appearance of independence on the part of Communist parties would ease their participation in coalition governments after the war.

The rival Poles

At the Trident Conference in Washington (May 1943) Churchill and Roosevelt finally projected a 29-division invasion of France for May 1944. The long delay was the consequence of the need to build up troop strength, landing craft, and supplies, and to ensure complete command of air and sea. But Stalin again castigated Allied bad faith and initiated a series of vitriolic communications with Churchill.

Plan for the invasion of France

The final defeat of Rommel's Afrika Korps opened the way for the invasion of Sicily in July 1943. The Allies' rapid success there gradually undermined Mussolini's eroding Fascist regime. Badoglio, Ciano, and Grandi had all denounced Mussolini's leadership and had been sacked by February 1943. Other Fascist leaders insisted on convening the Grand Council in July and after violent debate voted 19 to 8 in favour of restoring "the prerogatives of the King and parliament." Mussolini resigned the next day, and Badoglio took power in the face of a complex dilemma. Italy wanted peace, but to break the alliance with Hitler might provoke a German attack and condemn Italy to prolonged fighting. Thus, while feigning continued loyalty to Germany, Badoglio made secret contact with Eisenhower in the hope of synchronizing an armistice and an Allied occupation. But the Americans insisted on August 11 that Italy give an unconditional surrender and would not promise to land as far north as Rome. With tension and German suspicions mounting—and two British corps crossing the Straits of Messina—Badoglio agreed secretly to invite Allied occupation on September 3. The armistice was announced on the 8th, and Allied landings followed that night in the Bay of Salerno south of Naples. Four days later Hitler sent a crack team of commandos under Otto Skorzeny to rescue Mussolini and set him up as a puppet dictator in the north of Italy.

The new Italian government, far from exiting the war, was obliged to do a volte-face and declare war on Germany on October 13. The Allies did not take Naples until October 1 and made no dent in the Germans' reinforced Gustav Line until 1944.

Italy's volte-face

Early war-aims agreement. The Quebec Conference (Aug. 14–24, 1943) was the first in which Roosevelt and Churchill spent more time discussing the Pacific War than the European. They gave green lights to General MacArthur to fight northward toward the Philippines and to the U.S. Navy to drive straight across the Pacific to the Ryukyu Islands. The British even reluctantly accorded the U.S. Navy program top priority. The Allies also confirmed the invasion of France for May 1944, and thenceforth the American strategy of concentration would take precedence over British peripheral strategy. Eden and Hull then journeyed to Moscow (October 19–30), where they assured Stalin of the date for a second front. They also won his approval of the arrangements made for Italy, according to which the interallied commission requested by Stalin would merely advise the Anglo-American commanders on the spot rather than govern on its own. When Soviet armies later entered eastern European states, Stalin would point to the Italian precedent to justify unilateral Soviet military control.

At the Cairo Conference (November 22–26), Roosevelt, Churchill, and Chiang discussed the Burma theatre and made the Cairo Declaration, which prescribed as terms for ending the Pacific War the Japanese surrender of Manchuria, Formosa, Korea, the Pescadores, and Pacific

The Cairo Declaration

islands acquired since 1914. It also established Chiang as one of the Great Power allies, a point that did not please Churchill.

The first Big Three summit meeting followed in Tehrān from Nov. 28 to Dec. 1, 1943. From the Soviet point of view, the results could only have been satisfactory, for Stalin saw with his own eyes the conflicts that Communist theory predicted must erupt between the "imperialist" powers. In fact, Roosevelt and Churchill displayed the inevitable divergences between a moralizing democracy recently forced out of isolation and a world empire committed for 250 years to preserving the balance of power. What was more, Churchill had no illusions about the Soviet dictator, whereas Roosevelt preferred to believe that he could reason with "Uncle Joe" if only he could allay Soviet suspicions. Roosevelt made a point of chiding Churchill in Stalin's presence and advocating an end to European colonialism after the war. For his part, Stalin again demanded his 1941 frontiers, and the Baltic coast of East Prussia as well, and the others acquiesced in the restoration of the Curzon Line frontier, provided Poland was compensated with territories taken from Germany in the west. As to Germany itself, the Western powers had discussed breaking up the country and turning the Danubian regions of Austria, Hungary, and Bavaria into a "peaceful, cowlike confederation," while Churchill spoke of similar federations for eastern Europe. Stalin viewed such notions with suspicion, since they were reminiscent of the *cordon sanitaire* idea of 1918 and in any case would interfere with the piecemeal communization of the small states. His plan was to Balkanize eastern Europe, punish France for her surrender and strip her of her colonies, and keep Poland and Italy weak. As U.S. diplomat Charles E. Bohlen recorded at Tehrān: "The result would be that the Soviet Union would be the only important military power and political force on the continent of Europe." Roosevelt did win an agreement in principle on formation of a postwar international organization to be led by the United States, the Soviet Union, Britain, and China. Whether unity among them would survive victory was a question Churchill and others brooded on in silence.

Soviet plans for eastern Europe

THE DEFEAT OF NAZI GERMANY

In 1944 the German forces in Soviet territory shrank from attrition and transfers to the west, while geography and Hitler's reluctance to authorize retreats gave his generals no prospect of shortening the front. Soviet advances were limited only by their own supply capacity. A three-pronged offensive in March squeezed the Germans out of the southern Ukraine. Only the Carpathian Mountains kept the Red Army from the Hungarian Plain, and on March 20 Hitler ordered German occupation of Hungary to prevent the regent Admiral Miklós Horthy from defecting to the Allies. The Red Army entered Bessarabia and northern Romania in April. In the south, Odessa fell on April 10, and Sevastopol on May 9. In the far north, German forces withdrew from Leningrad to Lake Peipus, relieving that city after more than two years of siege and combat that killed 632,000 civilians, mostly from starvation. A two-month pause followed in the Soviet Union, during which the western Allies finally opened the second front in France.

The Allied invasion of Europe. While preparations for D-Day reached their final stages the Allies made a fateful decision to campaign vigorously on the Italian front in hopes of drawing off German reserves from France. But German resistance was fierce, and by October autumn rains curtailed Allied attacks, ending their dream of bursting into Austria from the south.

By spring 1944 the Germans had mustered 59 divisions in France and the Low Countries, but only 10 were motorized and almost 30 were in static defense positions. As the Allied buildup in England reached huge proportions, the Germans tried to divine where the blow would come. Hitler and Rommel thought Normandy; the theatre commander, Rundstedt, believed Calais. Their deployments reflected a compromise. Meanwhile, Roosevelt and Marshall chose Eisenhower to command Supreme Headquarters Allied Expeditionary Force (SHAEF), and he managed the preparation of "Overlord," the cross-Channel invasion, with tact and skill. More than 3,000,000 men crowded into southern English bases and ports, anxiously awaiting a D-Day on which 176,475 soldiers, 20,111 vehicles, 1,500 tanks, and 12,000 planes would move by air and sea across the Channel. Eisenhower described them as being "as tense as a coiled spring." Elaborate deceptions kept the Germans guessing about the point of attack, and Normandy was chosen in part because it was not the easiest or nearest French beachhead. On June 6, American, British, and Canadian forces went ashore, but seven tense and bloody weeks passed before the Allies broke out of the Norman peninsula. The initial campaign, thanks to Allied courage and matériel and German blunders, removed more divisions from the Wehrmacht's order of battle than even the great Soviet offensive of June 1944.

Operation "Overlord"

As Allied armies raced westward and northward to liberate France, Eisenhower faced the problem of what to do with Paris. He had no desire to interrupt the drive for a difficult urban battle, nor to undertake the chore of feeding 4,000,000 inhabitants. But the Parisian police went on strike on August 19, and de Gaulle secretly ordered French forces to seize the capital. Meanwhile, Hitler had ordered that the landmarks of Paris be blown up before the Germans retreated. But garrison commander Dietrich von Choltitz refused to carry out the order and negotiated a surrender that opened the city to Allied forces on the 25th. Eisenhower gave the honour of leading the parade to de Gaulle and General Jacques-Philippe Leclerc.

Soviet advances in the east. In five months from D-Day the Western Allies liberated France and Belgium and advanced 350 miles. In the midst of the Normandy campaign, on June 22, the Red Army launched its summer offensive. Armoured spearheads chased German remnants to the East Prussian border and the banks of the Vistula by July 31, an advance of 450 miles in five weeks. By October the Baltic coast was cleared of Germans. These massive victories carried the Red Army to the borders of nine states that had been independent before 1939, making possible the sovietization of eastern Europe. The first episode in that process stemmed from an uprising by the Polish Home Army in Warsaw, underground allies of the London Poles. Expecting momentary liberation from across the Vistula, the Home Army rebelled against the German occupation and seized control of the city. But Stalin called it a "reckless venture," and the Soviets sat idly by while Hitler ordered in SS divisions to crush the resistance and flatten the ancient city. To be sure, the Red Army had just finished a huge advance that stretched its supply lines to the limit. But Stalin shed no tears over the slaughter of the non-Communist Warsaw Poles, who held out bravely for eight weeks, and even hindered U.S. and British planes from supplying Warsaw by denying them landing rights in Soviet territory. On August 22, Stalin simply dismissed the Warsaw Poles as "criminals" and set up his Moscow Poles in Lublin as the acting government of "liberated Poland." In the north, the Finns sued for peace in early September, accepting their 1940 losses and giving up in addition the Arctic port of Petsamo (Pechenga), and a $300,000,000 indemnity, terms confirmed in the treaty of peace concluded in 1947. The U.S.S.R. allowed the Finns self-rule so long as Helsinki coordinated its foreign policy with that of the U.S.S.R. Latvia, Lithuania, and Estonia, however, were reannexed.

The rape of Warsaw

The Soviets unleashed another major offensive in August through Bessarabia, even though the Balkan front was irrelevant to the quick defeat of Germany. King Michael concluded an armistice with Moscow on September 12. Citing the Italian precedent, Molotov brushed aside the Western Allies' attempts to win a share of influence over Romanian affairs. Bulgaria, which was not at war with the U.S.S.R., tried to establish its neutrality, but the Red Army occupied it anyway and set up a "Fatherland Front" in which Communists were predominant. When Soviet and Romanian troops invaded Hungary in October, Horthy tried to extract his country from the war. But the SS arranged his overthrow, and fighting continued until the fall of Budapest on Feb. 13, 1945. A foolish waste of troops for the Nazis, the battle of Budapest was equally irrational

The fall of Budapest

for Stalin unless his true goal was political. Meanwhile, Yugoslav partisans under a local Communist, Josip Broz Tito, captured Belgrade on Oct. 20, 1944, and evicted the Germans.

One by one the states of eastern Europe were falling to Communist forces in circumstances prejudicing their future independence. When Churchill arrived in Moscow on Oct. 9, 1944, he tried to contain the march of Communism into central Europe by making a deal with Stalin on spheres of influence: Romania to be 90 percent Soviet; Greece 90 percent British; Yugoslavia and Hungary 50–50; Bulgaria 75 percent Soviet, 25 British. While apparently a realistic response to Soviet ambitions—and presence—in contrast to Roosevelt's reliance on vague principles, Churchill's proposal was in fact rather silly. Stalin was unlikely to grant Western influence in countries under Soviet occupation (like Hungary), while the meaning of such numbers as "75–25" was unfathomable. Poland was not mentioned at all. On the other hand, Churchill did forestall Soviet aid to the Communist partisans in Greece and may have helped to shield the crucial Mediterranean from Soviet influence for years after the war.

The final Allied agreements. In February 1945 the Big Three held their last summit conference, at Yalta on the Crimean Peninsula. It was a last chance to forestall the disintegration of the alliance upon victory or, conversely, for the British and Americans to take firm measures against Soviet control in eastern Europe. Roosevelt was now mortally ill and exhausted by the strenuous journey. Controversy later raged over his decision to attend the conference at all, his eagerness to conciliate Stalin, and the sinister presence in his entourage of Communist agent Alger Hiss. Postwar critics would charge that Roosevelt had been duped at Yalta and had "sold out" eastern Europe to the Communists. Doubtless if Churchill's advice had been followed, the policy of trust might have given way to one of hard bargaining and clear haggling over boundaries and governments in Europe and Asia. But in fact there was little the Western powers could have done to frustrate Stalin other than threatening a new world war. Nor could Churchill and Roosevelt have openly relinquished any liberated states to Stalin without abrogating the principles on which the war had been fought and alienating the millions of U.S. voters of eastern European descent. As for Asia, the United States was yet facing a campaign that might cost hundreds of thousands of American lives. Purchasing Soviet help against Japan seemed both realistic and humane. Roosevelt could not predict that the atomic bomb would render Soviet aid superfluous.

The controversy over Yalta

Of the three great allies Britain was the weakest and most interested in restoring a balance of power in Europe. Churchill, a keen critic of Bolshevism since 1919, had lobbied all throughout the summer of 1944 for an Italian campaign in hopes that the Allies might reach the Danube before the Red Army, and in October he had made the "spheres of influence" deal with Stalin. But the war map—and Roosevelt's unwillingness to strain the alliance—defeated all these tactics. On the eve of Yalta, Churchill wondered whether "the end of this war may well prove to be more disappointing than was the last." American war aims, by contrast, were nebulous to nonexistent, except for a reprise of Wilsonian internationalism. There is little evidence of economic motives in U.S. policy and, incredibly, no contingency plans for a breakdown in relations with the U.S.S.R. While Roosevelt feared another American retreat into isolationism, he also believed in the possibility of a postwar Great Power condominium. He was prepared to show Stalin that the Anglo-Saxons were not ganging up on him and wanted Soviet participation in a United Nations Organization. But Stalin pursued the old-fashioned way of postwar security: military and political control of eastern Europe to create a buffer for the U.S.S.R. and to ensure Soviet domination over its own repressed nationalities.

At the Yalta Conference, Big Three unity seemed intact, but only because the participants resorted to vagueness or postponements on the most explosive issues. A joint European Advisory Commission, it was decided, would divide Germany into occupation zones, with the Soviet zone extending to the Elbe and a French zone carved out of the Anglo-American spheres. Berlin would likewise be placed under four-power control. The Western Allies repudiated the extreme plans broached at Quebec for the pastoralization of Germany and favoured German industrial recovery under international control. But the Soviets insisted on the right to strip Germany of $20,000,000,000 worth of machinery and raw materials. The issue was assigned to a reparations commission. As for the political future of Germany, Stalin revived earlier Big Three talk of breaking Germany into several states, but the Western Allies now perceived the danger of further Balkanization in central Europe in light of Soviet power. This matter, too, was left for study.

Soviet reparations demands

Poland was, as always, a most difficult problem. The Western Allies reiterated their Tehrān approval of the Curzon Line, now modified slightly in Poland's favour, as the Soviet–Polish border. But the assignment of 2,700,000 Germans to Poland in the West worried Churchill: "It would be a pity to stuff the Polish goose so full of German food that it died of indigestion." Hence Poland's western frontier would be left to a peace conference. As for the Polish government, the most the Western Allies achieved was a vague promise from Stalin that he would reorganize the Lublin Committee and permit free elections among "non-Fascist elements" within a month after peace. But Stalin reserved the right to decide who was "Fascist" and rejected international supervision of the elections. Roosevelt proposed a Declaration on Liberated Europe, by which the Big Three promised to help all liberated peoples "to solve by democratic means their pressing political and economic problems" through "free elections of Governments responsive to the will of the people." Stalin signed, probably considering this more high-flown American rhetoric meant for domestic consumption. In the Communist lexicon words like democratic and free implied conditions virtually the opposite of what Roosevelt intended. Since Roosevelt also announced (to Churchill's despair) that the United States would evacuate its troops from Europe within two years, Stalin may have felt that he could safely ignore the Declaration on Liberated Europe.

Stalin did prove conciliatory on the United Nations, which had already been discussed at the Dumbarton Oaks Conference between Aug. 21 and Oct. 7, 1944. The Soviets had demanded that all 16 constituent republics of the U.S.S.R. be represented (ostensibly to balance the British Empire nations that would vote with London) and that permanent members of the Security Council retain a veto on all issues, not just those involving sanctions or threats to peace. At Yalta, Stalin settled for three seats in the General Assembly and a limited veto. Like Wilson at Versailles, Roosevelt put great stock in international organization and was prompted to remark, "The Russians have given in so much at the conference that I don't think we should let them down." Finally, Stalin promised to declare war on Japan within 90 days of the German surrender in return for southern Sakhalin and the Kuril Islands, retention of Outer Mongolia, and a promise of U.S. support for Soviet rights at Dairen (Lü-ta) and Port Arthur (Lüshun)—all the old objects of Russian imperialism in east Asia. Within a month news from the various commissions established at Yalta indicated that the Soviets did not intend to meet Western expectations. When Molotov announced on March 23 that most of the London Poles were disqualified from Polish elections, Roosevelt reportedly banged his fist on his wheelchair: "Averell [Harriman, ambassador in Moscow] is right. We can't do business with Stalin. He has broken every one of the promises he made at Yalta." Roosevelt then retreated, disillusioned, to Warm Springs, Ga., where he died on April 12.

The United Nations compromises

The Allied advance from the west was stalled for six weeks by the Battle of the Bulge, Hitler's last offensive, but by February 1945 German resistance was near its end. Some Soviet and Western leaders were openly describing the last campaigns as a "land-grab" directed as much against their distrustful allies as against the Germans. But the commanders in the West still took steps to prove that they were supporting the Soviet advance. The worst product of this policy was the Allied bombing of Dresden on

Feb. 13–14, 1945, allegedly to destroy a key communications centre for Germans facing the Red Army. But the two-day incendiary raid by almost 2,400 bombers created a fire storm that consumed the medieval city and killed perhaps 135,000 civilians, to virtually no military purpose.

Another product of Western efforts to reassure Stalin was the refusal to order British and American armies to race the Soviets to Berlin. On March 7, 1945, General George Patton's tanks broke through weak German lines and the 1st Army infantry captured intact a Rhine bridge at Remagen. Churchill pleaded for a rapid thrust in order to secure Berlin and Prague: "Highly important that we shake hands with the Russians as far to the east as possible." Stalin, in turn, tried to lull his allies by saying that "Berlin has lost its former strategic importance," while in fact ordering his generals to make for it as soon as possible. But Eisenhower, backed by Marshall, confined himself to military considerations alone. The Allied armies would close the Ruhr pocket, then advance in breadth in case the rumours were true of a Nazi "Alpine redoubt" in the south. When the Western armies exceeded the limits of their occupation zones in April, Eisenhower even called them back. Soviet forces, meanwhile, captured Vienna and Königsberg on April 9 and encircled Berlin by the 25th. Five days later a despairing Hitler declared that Germany had proved unworthy of him and committed suicide in his Berlin bunker. Hitler's successor, Admiral Dönitz, opened negotiations with the Western powers, hoping to save as many troops and refugees as possible from Soviet reprisals. But the U.S.S.R. refused to recognize the surrender ceremony at Eisenhower's headquarters on May 7, necessitating a second surrender, and a separate Soviet V-E Day, in Berlin on May 8. The war in Europe was over.

Allied occupation

THE DEFEAT OF JAPAN

The encirclement of Japan. By January 1944 the American buildup in the Pacific permitted both the army and navy commands to accelerate the rollback of Japanese power. Indeed, the United States had by then deployed as many men and planes and more ships in the Pacific theatre as in the European. The army under General MacArthur aimed at the liberation of the Philippines, thereby cutting Japanese communications with the East Indies and the sea route to Southeast Asia. The navy under Admiral Chester Nimitz moved up the Marshall and Mariana chains to bring U.S. bombers within range of the Japanese home islands. In both cases the Americans employed the tactic of island-hopping and relied on superior firepower to inflict appalling casualties on fanatical Japanese defenders who preferred death to the shame of surrender.

In the central Pacific, the navy's material superiority allowed Nimitz to pierce Japan's "absolute national defense sphere" almost at will. By 1943 the United States was producing 100,000 planes per year, compared to Japan's total of 63,000 for the entire war. By the summer of 1944 the United States had nearly 100 carriers of all types in the Pacific, compared to Japan's total of 20 for the war. The Japanese also lost more than 80 percent of the 6,000,000 tons of shipping with which they began the war (half to U.S. submarines) and were forced to expose their proud navy to destruction in a vain effort to supply their far-flung garrisons. The U.S. advance was limited only by its own supply lines, which stretched 5,000 miles from Pearl Harbor and 8,000 from the continental bases of California.

Japan's shipping losses

The bombing of the Japanese home islands achieved a new plateau of horror when the U.S. Army Air Forces adopted Britain's European tactics of low-level nighttime raiding on urban areas. On the night of March 9–10, 1945, napalm area bombing of largely wooden Tokyo stoked firestorms that destroyed a quarter of the city, killed 80,000 civilians, and left 1,000,000 homeless. Similar devastating fire raids were launched against Ōsaka, Kōbe, Yokohama, and other cities.

The atomic decision. By April, Japan lay open to direct assault by land as well as air and sea. How could the United States bring Tokyo to surrender? Three means suggested themselves: invasion, inducement, and shock. The first would involve a lengthy, brutal campaign in which, it was estimated, hundreds of thousands of American and perhaps 2,000,000 Japanese lives would be lost. Yet the Joint Chiefs had no choice but to prepare for this eventuality, and by May 25 they had instructed MacArthur to plan Operation "Olympic," an invasion of Kyushu, for November 1. The second means, inducement, was clearly preferable, and on May 8, the day after the German surrender, President Harry S. Truman tried it. Unconditional surrender, he said, would mean "the termination of the influence of the military leaders who have brought Japan to the present brink of disaster," but did not mean "the extermination or enslavement of the Japanese people," who would be free to "return to their families, their farms, their jobs." Unfortunately, Truman did not include (as the State Department advised) a promise that the Japanese might retain their emperor, the god-king of their Shintō state religion. On the other hand, the Japanese government foolishly dismissed Truman's appeal as propaganda and began to mobilize the home front to resist an invasion.

The third means of achieving a surrender—by shock—had become a possibility on Dec. 30, 1944, when General Leslie Groves, head of the Manhattan Project, reported that it was "reasonably certain" that a gun-type atomic bomb equivalent to 10,000 tons of TNT and an implosion-type bomb would be ready for testing by the summer of 1945. On April 25, soon after Truman's accession to the presidency, Secretary of War Stimson impressed on him the significance of this development: "Within four months we shall in all probability have completed the most terrible weapon ever known in human history, one bomb of which could destroy a whole city." He then formed an Interim Committee of statesmen and scientists to debate how the bomb should be employed. On May 31 and June 1 the committee received scientific briefings and held discussions on whether to share the secret with the Soviets, how long it would take other nations to develop their own atomic bomb, how international control might be achieved, whether the U.S. monopoly might help Washington in its relations with Moscow, and whether the bomb would be a universal blessing or a Frankenstein's monster.

The Interim Committee

In the matter at hand, however, the committee concluded that the bomb should be used to end the war as soon as possible; that it should be dropped on a military-urban target so as to demonstrate its full force; and that a demonstration or warning should not be made beforehand, lest the bomb lose its shock value. The scientific panel under J. Robert Oppenheimer concurred on June 16. As he later said: "We didn't know beans about the military situation in Japan. . . . We did say that we did not think exploding one of these things as a firecracker over a desert was likely to be very impressive."

The first atomic test at Alamogordo, N.M., on July 16, 1945, yielded an explosion equivalent to that of 15,000 tons of TNT and stunned Oppenheimer and his colleagues with its elemental power. At that moment Truman was attending the final Big Three meeting at Potsdam, Ger., and casually mentioned to Stalin that the United States had "a new weapon of unusual destructive force." Stalin said he was glad to hear of it and hoped the United States would make good use of it against the Japanese. Though little else was agreed upon at Potsdam, the Big Three did jointly invite Japan on July 26 to surrender unconditionally or face "prompt and utter destruction." When no surrender was forthcoming, Truman gave the Army Air Forces on Tinian Island the green light. He wrote later that he never lost a moment's sleep over his decision.

A specially equipped B-29, the *Enola Gay,* dropped an atomic bomb on the military port of Hiroshima on Aug. 6, 1945. The heat and blast effaced everything in the vicinity, burned 4.4 square miles, killed upwards of 66,000 people, and injured 70,000 more. Two days later the U.S.S.R. declared war on Japan and invaded Manchuria. On August 9 the second atomic bomb fell on Nagasaki, killing 39,000 people. On that day the Voice of the Sacred Crane—the emperor's command—summoned the Cabinet to an audience. Hirohito expressed his wish that Japan accept the terms of the Potsdam Declaration on the sole condition that the emperor remain sovereign. To continue the war, he said, would be suicidal. And then,

Hiroshima and Nagasaki

perhaps realizing the irony of that remark, he turned to the military men and noted that their performance had fallen rather short of their promises. Even at that late date some fanatical officers attempted a coup on the palace grounds rather than submit. But on Sept. 2, 1945, General MacArthur received the Japanese surrender on the battleship *Missouri* in Tokyo Bay, and the greatest war in history came to a close.

The coming of the Cold War, 1945–57

The symbolic first meeting of American and Soviet soldiers occurred at Torgau, Ger., on April 25, 1945. Their handshakes and toasts in beer and vodka celebrated their common victory over Nazi Germany and marked the collapse of old Europe altogether; but their inarticulate grunts and exaggerated smiles presaged the lack of communication in their relationship to come. Grand wartime coalitions invariably break up once the common fight gives way to bickering over division of the spoils, but feuding victors after the wars of Louis XIV and Napoleon or World War I at least negotiated treaties of peace, while the rancour among them was moderated by time or the danger that the common enemy might rise again. But after 1945 no grand peace conference convened, no common fear of Germany or Japan survived, and the quarrels among the victors only grew year by year into what U.S. presidential adviser Bernard Baruch and pundit Walter Lippmann termed a Cold War.

The U.S.–Soviet conflict began in 1945 over treatment of occupied Germany and the composition of the Polish government. It grew during 1946 as the Soviets communized the lands under their occupation and the victors failed to agree on a plan for the control of atomic energy. From 1947 to 1950 the reactions of Washington and Moscow to the perceived threats of the other solidified the division of Europe and much of the world into two blocs led by what are now called the superpowers, and the Cold War became universalized, institutionalized, and militarized.

Emergence of superpowers

The settlement after World War II, therefore, was a peace without treaties, and the Cold War magnified, distorted, or otherwise played upon the other historical trends given impetus by the world wars of the 20th century: Asian nationalism, decolonization, the seeming culmination of the 37-year-old Chinese Revolution, the evolution of independent Communist parties in Yugoslavia and Asia, and western Europe's drive to end four centuries of conflict through economic integration. The early Cold War was not a decade of fear and failure alone but also a creative time that gave birth to the closest thing to a world order that had existed since 1914. For with the sole major exception of the later Sino-Soviet split, the boundaries, institutions, and relationships fashioned in the late 1940s were very nearly the same ones that shaped world politics in the 1980s.

THE COLD WAR GUILT QUESTION

Historian Charles S. Maier has said that the debate over the origins of the Cold War is an extension of the Cold War itself. Indeed, as early as 1948 American left-liberals blamed the Truman administration for the icy tone of its relations with Moscow; rightists blamed the Communists but accused Roosevelt and Truman of appeasement; while moderates of both parties shared a consensus that Truman's containment policy was, as the historian Arthur Schlesinger, Jr., wrote, "the brave and essential response of free men to communist aggression." After all, Stalin's tyranny was undeniable and his seizure of countries in eastern Europe one by one was reminiscent of Hitler's "salami tactics." To be sure, Roosevelt may have helped to foster mistrust by refusing to discuss war aims earlier and then relying on vague principles, and Truman may have blundered or initiated steps that solidified the Cold War. But those steps were taken only after substantial Soviet violation of wartime agreements and in fearful confusion over the motivations for Soviet policy. Was the U.S.S.R. implacably expansionist, or were its aims limited? Was it executing a plan based on Communist faith in world revolution, or reflecting the need of the regime for foreign enemies to justify domestic terror, or merely pursuing the traditional aims of Russian imperialism? Or was Stalin's own paranoia or ambition responsible for Soviet aggression, in which case different leaders might prove more accommodating? Since Soviet sources remain inaccessible these questions have continued to bedevil Western historians and statesmen.

Soviet motives

The fact that Western societies tend to parade their disagreements and failures in public, in contrast to the Soviet fetish for secrecy, guaranteed that historical attention would fix on American motivations and mistakes. In the late 1950s and 1960s traditional left-liberal scholars smarting from the excesses of McCarthyism and new leftists of the Vietnam era began publishing revisionist interpretations of the origins of the Cold War. The "hard revisionism" of William Appleman Williams in 1959 depicted the Cold War in Marxist fashion as an episode in American economic expansion in which the U.S. government resorted to military threats to prevent Communists from closing off eastern European markets and raw materials to U.S. corporations. Less rigidly ideological "soft revisionists" blamed the Cold War on the irascible Truman administration, which, they charged, had jettisoned the cooperative framework built up by Roosevelt at Tehrān and Yalta and had dropped the atomic bombs on Japan as a means of frightening the Russians and forcing an "American peace." These revisionist interpretations were not based so much on new evidence as on new assumptions about U.S. and Soviet motives, influenced in turn by the protest movements against the Vietnam War, nuclear weapons, and alleged domination of American society by the "military-industrial complex." Looking back to the years after 1945, they argued that Stalin was not a fanatical aggressor but a traditional Soviet statesman. After all, the Soviet Union had been brutally invaded and had lost 20,000,000 lives in the war. Stalin could thus be excused for insisting on friendly governments on his borders. But he was betrayed, said revisionists, by American militancy and Red-baiting after the death of Roosevelt.

Traditional historians countered that little evidence existed for most of the revisionist positions. To be sure, American hostility to Communism dated from 1917, but the record proved Roosevelt's commitment to good relations with Stalin, while no proof at all was forthcoming that American policymakers were anxious to penetrate eastern European markets, which were, in any case, of minor importance to the U.S. economy. Williams rebutted that policymakers so internalized their economic imperialism that they did not bother to put their thoughts on paper. But this "argument from no evidence" made a mockery of scholarship. The preponderance of evidence also indicated that the atomic decision was made for military considerations, although isolated advisers did hope that it would ease negotiations with Moscow. These and other examples led most historians to conclude that, while the revisionists brought to light new issues and exposed American aimlessness, inconsistency, and possible overreaction at the end of World War II, they failed to establish their primary theories of American guilt.

Historians with a longer perspective on the Cold War transcended the passions of Vietnam-era polarization and observed that deeper forces must have been at work for the Cold War to have persisted for so long after 1945. Indeed, it is difficult to imagine how leaders of the two countries could have sat down agreeably and settled the affairs of the world. For the new superpowers were wrenched out of isolationism and thrust into roles of world leadership, they nurtured contrary universalist ideologies, and they mounted asymmetrical military threats (one based on conventional weapons, sheer numbers, and land power; the other on nuclear might, technological superiority, and air and sea power). To these liabilities could be added the fact that both countries had been forced into World War II by sneak attacks and had resolved never again to be seduced into appeasement or to be taken by surprise.

Roots of the Cold War

Even such a balanced long-range view should not be taken uncritically. It remains the case that the Cold War grew out of specific diplomatic disputes, among them Germany, eastern Europe, and atomic weapons. Could

those disputes have been avoided or amicably resolved? Certainly some prior agreement on war aims might have softened the discord after 1945, but Roosevelt's policy of avoiding divisive issues during the war, while wise in the short run, enhanced the potential for conflict. It might, without undue exaggeration, be said that the United States entered the postwar period with only a vision of a postwar economic world and few political war aims at all, and thus had little excuse for indignation once Stalin set out methodically to realize his own aims. But this does not justify a Soviet policy bent on denying self-rule to neighbouring peoples and imposing police states as cruel as those of Hitler. For although the Soviets had lost 20,000,000 in the war, Stalin had killed at least an equal number of his own citizens through deliberate famine and purge. American hegemony, if it can be called that, was by contrast liberal, pluralistic, and generous.

The question has been posed, is it not an expression of American exclusivism, self-righteousness, or cultural imperialism to insist that the rest of the world conform to Anglo-Saxon standards of political legitimacy? Even if so, critics must take care not to indulge in a double standard: excusing the U.S.S.R. for being "realistic" and damning the United States for being insufficiently "idealistic."

WASTELAND: THE WORLD AFTER 1945

The ruin of Europe and Japan. Harry Truman had been an artilleryman in World War I and remembered well the lunar landscape of the Western Front. Yet while driving from Potsdam to Berlin in July 1945 he exclaimed: "I never saw such destruction!" Almost all the great cities of central and eastern Europe were jagged with ruined buildings, pitted roads, wrecked bridges, and choked waterways. Amid it all were the gaunt survivors, perhaps 45,000,000 of them homeless, including 25,000,000 in those lands—Poland, the Ukraine, and Russia—that had been overrun and scorched two or three times. European communications and transportation reverted to 19th-century levels: 90 percent of French trucks and 82 percent of French locomotives were out of commission, as were over half the rolling stock in Germany and two-thirds of the Balkan railroads. European coal production was at 40 percent of prewar levels, and more than half the continent's merchant marine no longer existed. Some 23 percent of Europe's farmland was out of production by war's end. Of course, people could be fed with American aid while the rubble was cleared away and utilities restored. But World War II cost Europe more in monetary terms than all its previous wars put together. The war also set in train the greatest *Völkerwanderung*—movement of peoples—since the barbarian incursions of the late Roman Empire. During the Nazi onslaught some 27,000,000 people fled or were forced out by war and persecution, and 4,500,000 more were seized for slave labour. When the Red Army advanced westward millions more fled before it to escape reprisals or Communism. All told, about 60,000,000 people of 55 ethnic groups from 27 countries were uprooted. Finally, 7,000,000 Axis prisoners of war were in Allied hands, along with 8,000,000 Allied prisoners of war liberated from the Axis and 670,000 survivors of Nazi death camps.

The survivors

The landscape in much of Japan was just as barren, its cities flattened by bombing, its industry and shipping destroyed. Large parts of China had been under foreign occupation for up to 14 years and—like Russia after World War I—still faced several years of destructive civil war. Indeed, the war had laid waste every major industrial region of the globe except North America. The result was that in 1945–46 the United States accounted for almost half the gross world product of goods and services and enjoyed a technological lead symbolized by, but by no means limited to, its atomic monopoly. On the other hand, Americans as always wanted to demobilize rapidly and return to the private lives and careers interrupted by Pearl Harbor. The Soviet Union, by contrast, was in ruin, but its mighty armies occupied half a dozen states in the heart of Europe, while local Communist parties agitated in Italy and France. The United States and the Soviet Union thus appeared to pose asymmetrical threats to each other.

U.S. economic dominance

As British ambassador Sir Archibald Clark-Kerr reported from Moscow: "There was great exultation. Russia could be made safe at last. . . . Then plump came the Atomic Bomb. At a blow the balance which had now seemed set was rudely shaken. Russia was balked by the West when everything seemed to be within her grasp. The three hundred divisions were shorn of much of their value."

U.S. vision of reconstruction. American planners envisioned postwar reconstruction in terms of Wilsonian internationalism but were determined to avoid the mistakes that resulted after 1918 in inflation, tariffs, debts, and reparations. In 1943 the United States sponsored the United Nations Relief and Rehabilitation Administration to distribute food and medicine to the stricken peoples in the war zones. At the Bretton Woods Conference (summer 1944) the United States presided over the creation of the International Monetary Fund and the World Bank. The dollar was returned to gold convertibility at $35 per ounce and would serve as the world's reserve currency, while the pound, the franc, and other currencies were pegged to the dollar. Such stability would permit the recovery of world trade, while a General Agreement on Tariffs and Trade (ratified in 1948) would ensure low tariffs and prevent a return to policies of economic nationalism. Treasury Secretary Henry Morgenthau tried to entice the Soviets to join the Bretton Woods system, but the U.S.S.R. opted out of the new economic order.

The American universalist program seemingly had more luck in the political realm. Roosevelt was convinced that the League of Nations had been doomed by the absence of the United States and the Soviet Union and thus was anxious to win Soviet participation in the compromises at Yalta. The Big Four powers accordingly drafted the Charter of the United Nations at the San Francisco Conference in April 1945. Roosevelt wisely appointed several leading Republicans to the U.S. delegation, avoiding Wilson's fatal error and securing the Senate ratification of the Charter on July 28, 1945, by a vote of 89–2. Like Wilson, Roosevelt and Truman hoped that future quarrels could be settled peacefully in the international body.

The UN charter

The end of East–West cooperation. By the time of the Potsdam Conference, Truman was already aware of Soviet unwillingness to permit representative governments and free elections in the countries under their control. The U.S.S.R. compelled the king of Romania to appoint a Communist-dominated government, Tito's Communists assumed control of a coalition with royalists in Yugoslavia, Communists dominated in Hungary and Bulgaria (where a reported 20,000 people were liquidated), and the Red Army extended an invitation to "consult" with 16 underground Polish leaders only to arrest them when they surfaced. As Stalin said to the Yugoslav Communist Milovan Djilas: "In this war each side imposes its system as far as its armies can reach. It cannot be otherwise." On April 23, 1945, Truman scolded Molotov for these violations of the Yalta Accords and, when Molotov protested such undiplomatic conduct, replied: "Carry out your agreements and you won't get talked to like that." On May 11, three days after the German surrender, Truman abruptly ordered the termination of Lend-Lease aid to the U.S.S.R. Two weeks later Stalin replied in like terms to envoy Harry Hopkins by way of protesting the suspension of Lend-Lease, Churchill's alleged plan to revive a *cordon sanitaire* on Russia's borders, and other matters. But Hopkins assured him of American good will and acquiesced in the imprisonment of the Polish leaders and the inclusion of only a few London Poles in the new government. The United States and Britain then recognized the Warsaw regime, assuring Soviet domination of Poland.

The short-lived détente was to be consummated at Potsdam, the last meeting among the Big Three. In the midst of the conference, however, the British electorate rejected Churchill at the polls, and Labour Party leader Clement Attlee replaced him in the councils of the great. Aside from the Soviet promise to enter the war against Japan and Truman's hint that the United States had developed the atomic bomb, the Potsdam Conference dealt with postwar Europe. The U.S.S.R. was authorized to seize one-third of the German fleet, extract reparations-in-kind

The Potsdam Conference

from its east German occupation zone, and benefit from a complicated formula for delivery of industrial goods from the western zones, 15 percent to be counted as payment for foodstuffs and other products sent from the Soviet zone. The conference provided for peace treaties with the defeated countries once they had "recognized democratic governments" and left their drafting to the Council of Foreign Ministers. Finally, the Potsdam nations agreed to prosecute Germans for war crimes in trials that were conducted at Nürnberg for a year after November 1945.

While Potsdam seemed to promise continued cooperation, it left the most divisive issues—the administration of Germany and the configuration of east European governments—to future discussion. At the first such meeting, in September, the new U.S. secretary of state, James F. Byrnes, asked why Western newsmen were not allowed in eastern Europe and why governments could not be formed there that were democratic yet still friendly to Russia. Molotov asked on his own account why the U.S.S.R. was excluded from the administration of Japan.

Truman enumerated the principles of American foreign policy in his Navy Day speech of October 27. Its 12 points echoed the Fourteen Points of Woodrow Wilson, including national self-determination, non-recognition of governments imposed by foreign powers, freedom of the seas, commerce, expression and religion, and support for the United Nations. But confusion reigned in Washington as to how to implement these principles in league with Moscow. As political commentator James Reston observed, two schools of thought seemed to compete for the ear of the President. According to the first, Stalin was committed to limitless expansion and would only be encouraged by concessions. According to the second, Stalin was amenable to a structure of peace but could not be expected to loosen his hold on eastern Europe so long as the United States excluded him from, for instance, Japan. Truman and the State Department drifted between these two poles, searching for a key to unlock the secrets of the Kremlin and hence the appropriate U.S. policy.

The Byrnes mission

Truman's last attempt to win the Soviets to his universalist vision was the Byrnes mission to Moscow of December 1945. There the Soviets promptly accepted an Anglo-American plan for a UN Atomic Energy Agency meant to control the development and use of nuclear power. Stalin also conceded that it might prove possible to make some changes in the Romanian and Bulgarian parliaments, though conceding nothing that might weaken his hold on the satellites. George F. Kennan of the Moscow embassy called the concessions "fig leaves of democratic procedure to hide the nakedness of Stalinist dictatorship," while Truman's own dissatisfaction with the results at Moscow, and growing domestic criticism of his "coddling" of the Russians, were pushing him toward a drastic reformulation of policy.

Why, in fact, did Stalin engage in such a hurried takeover of eastern Europe when it was bound to provoke the United States (magnifying Soviet insecurity) and waste the opportunity for access to U.S. loans and perhaps even atomic secrets? Was not Stalin's policy, in retrospect, simply unwise? Such questions cannot be answered with assurance, since less is known about the postwar Stalinist era (1945–53) than any other in Soviet history. But the most tempting clue is again to be found in Stalin's domestic calculations. If the Soviet Union were to recover from the war, not to mention compete with the mighty United States, the population would have to be spurred to even greater efforts, which meant intensifying the campaign against alleged foreign threats. What was more, the Soviets had only recently regained control of populations that had had contact with foreigners and, in some cases, collaborated with the invaders. Ukrainians in particular had tried to establish an autonomous status under the Nazis, and they persisted in guerrilla activity against the Soviets until 1947. If Soviet citizens were allowed widespread contact with foreigners through economic cooperation, international institutions, and cultural exchanges, loyalty to the Communist regime might be weakened. Firm control of his eastern European neighbours helped assure Stalin of firm control at home. Indeed, he now ordered the utter isolation of Soviet life to the point that returning prisoners of war were interned lest they "infect" their neighbours with notions of the outside world. Perhaps Stalin did not really fear an attack from the "imperialists" or consider a Soviet invasion of western Europe. But neither could he welcome the Americans and British as genuine comrades in peace without undermining the ideology and the emergency that justified his own iron rule.

Soviet isolationism

A swift return to Communist orthodoxy accompanied the clampdown on foreign contacts. During the war the U.S.S.R.'s leading economist, Evgeny Varga of the Institute of World Economy and World Politics, argued that government controls in the United States had moderated the influence of monopolies, permitting both dynamic growth and a mellower foreign policy. The U.S.S.R. might therefore benefit from East–West cooperation and prevent the division of the world into economic blocs. Stalin appeared to tolerate this nontraditionalist view as long as large loans from the United States and World Bank were a possibility. But the suspension of lend-lease, opposition to a Soviet loan in the State Department, and Stalin's renewed rejection of consumerism doomed these moderate views on the world economy. The new Five-Year Plan, announced at the start of 1946, called for continued concentration on heavy industry and military technology. The war and victory, said Stalin, had justified his harsh policies of the 1930s, and he called on Soviet scientists to overtake and surpass Western science. Varga perforce recanted and embraced the traditional view that Western economies were about to enter a new period of inflation and unemployment that would increase the imperialist pressure for war. Andrey Zhdanov, the Communist leader of Leningrad, was a bellwether. In 1945 he wanted to reward the Soviet people with consumer goods for their wartime sacrifices; in early 1947 he espoused the theory of the "two camps," the peace-loving, progressive camp led by the Soviet Union and the militaristic reactionary camp led by the United States.

American confusion came to an end after Feb. 9, 1946, when Stalin's great speech inaugurating the Five-Year Plan reiterated clearly his implacable hostility to the West. Kennan responded with his famous "Long Telegram" from Moscow (February 22), which for years to come served as a primer on Soviet behaviour for many in Washington. The Kremlin's "neurotic view of world affairs," he wrote, was the product of centuries of Russian isolation and insecurity vis-à-vis the more advanced West. The Soviets, like the tsars, viewed the influx of Western ideas as the greatest threat to their continued power, and they clung to Marxist ideology as a cover for their disregard for "every single ethical value in their methods and tactics." The U.S.S.R. was not Nazi Germany—it would not seek war and was averse to risk taking—but it would employ every means of subverting, dividing, and undermining the West through the actions of Communists and fellow travelers. Kennan's advice was to expect nothing from negotiations but to remain confident and healthy, lest the United States become like those with whom it was contending.

The Kennan analysis

Kennan's analysis implied several important conclusions: that the Wilsonian vision inherited from Roosevelt was fruitless; that the United States must take the lead in organizing the Western world; that the Truman administration must prevent a renewal of isolationism and persuade the American people to shoulder their new responsibilities. Churchill, though out of office, aided this agenda when he warned the American people (with Truman's confidential endorsement) from Fulton, Mo., on March 5, 1946, that an "iron curtain" had descended across the European continent.

THE COLD WAR IN EUROPE

Peace treaties and territorial agreements. The early spring of 1946 was a turning point when the United States gave up its hopes of cooperation in favour of what would soon be called "containment." The first manifestation occurred in March 1946, when the U.S.S.R. failed to evacuate Iran on schedule and Secretary of State Byrnes was obliged to go to the UN Security Council and even hint at hostilities to get Moscow to retreat. This incident,

together with Soviet pressure on Turkey and Yugoslav involvement in the Greek civil war, seemed to indicate that Communists were prepared to use force to expand.

The year 1946 saw many meetings of the Council of Foreign Ministers, including the peace conference envisioned at Potsdam. The council convened in Paris on July 29, 1946, and ultimately produced treaties of peace with Italy, Hungary, Romania, Finland, and Bulgaria, signed on Feb. 10, 1947. Border questions after World War II were comparatively minor—a somewhat ironic fact, given the interwar attacks on Versailles by all parties. Romania ceded northern Bukovina and Bessarabia back to the U.S.S.R., which also claimed Petsamo and the Karelian isthmus from Finland and the Carpatho-Ukraine from Czechoslovakia. Hungary returned northern Transylvania to Romania. Italy ceded the Dodecanese islands to Greece and surrendered its overseas colonies, although a Soviet demand for a trusteeship over Libya was denied. Trieste was contested by Italy and Yugoslavia and remained under Western occupation until 1954. The major change affected Poland, which was figuratively picked up and moved some 150 miles to the west. This meant that east and west Prussia and Silesia came under Polish administration, while the U.S.S.R. absorbed the entire Baltic coast as far as the venerable German port of Königsberg (Kaliningrad). The U.S.S.R. was the only power to make significant territorial gains from the war.

The new Poland

Four-power cooperation in Germany continued to deteriorate. The Americans had agreed at Potsdam to reparations-in-kind but opposed extreme efforts by the Soviets and the French to pauperize the Germans lest the burden of feeding them fall entirely on the American taxpayer. What was more, the Soviets would be unwilling (in Kennan's view) to countenance centralized German institutions unless they were in a position to use them to communize the entire country. In early May 1946, General Lucius Clay, commanding the U.S. zone, refused to authorize shipments out of western Germany until agreement was reached on treating Germany as a unit under four-power control. On September 6, Byrnes then announced a new policy: if unification of all Germany proved impossible, the United States would instead promote "maximum possible unification," (*i.e.,* in the western zones only). This ensured a permanently divided Germany.

Atomic energy. The superpowers also failed to join hands on atomic energy. Despite resistance from powerful circles in the press, Congress, and the military against any giveaway of atomic secrets, Byrnes appointed a committee in January 1946 to draft proposals for international control of atomic energy. The resulting (Dean) Acheson–(David) Lilienthal Report called for a UN authority to survey and control all uranium deposits and ensure that atomic research was conducted for peaceful purposes only. Once controls were in place, the United States would relinquish its arsenal and scientific information to the world community. Truman entrusted the diplomatic task to Baruch, who insisted that nations not be allowed to employ their Security Council veto in atomic matters. He then appealed to the UN on June 14, 1946: "We are here to make a choice between the quick and the dead." The Soviet plan, presented by Andrey Gromyko, called instead for immediate prohibition of all manufacture and use of atomic weapons. Measures to ensure compliance would follow, but there could be no tampering with the Security Council veto. Western delegates pointed out that the Soviets were asking the United States to give up its monopoly and make public all its data in return for a paper promise of compliance. Gromyko countered that the United States was asking all other countries to reveal the state of their own research before it gave up its own arsenal. At the final vote in December, the U.S.S.R. and Poland vetoed the Baruch Plan, and international control of atomic energy ceased to be a possibility. While the United States was not as forthcoming as it might have been, the Soviet refusal to allow on-site inspection would frustrate disarmament for the next 40 years.

The Baruch Plan

The economic battle with Communism. By the turn of 1947 it appeared that Truman's foreign policy was foundering. His secretary of agriculture, Henry A. Wallace, had been outspoken in criticism of the Baruch Plan and of the policy of "getting tough" with Russia. Upon resigning he became a leader of those whom Truman privately described as the "Reds, phonies and the parlour pinks" he feared were "a sabotage front for Uncle Joe Stalin." The 1946 elections then returned a Republican Congress bent on cutting costs and "bringing the boys home." Yet the United States was on the verge of the greatest reversal of its foreign policy traditions since 1917. For on Feb. 21, 1947, the British government announced that its economic difficulties would force it to suspend economic and military aid to Greece and Turkey by March 31. Greece was embroiled in civil war provoked by Communists. Turkey was under Soviet pressure for bases and naval passage through the Dardanelles. If those countries succumbed to Communist influence, the Mediterranean and the entire Middle East might follow. Truman, his new secretary of state, George C. Marshall, and Marshall's deputy, Dean Acheson, resolved at once that the United States must step in. On February 27 Acheson impressed congressional leaders with a vivid account of the Soviet strategy of expansion and its implications for American security. After a tense silence, Republican Senator Arthur Vandenberg vowed to support the new policy if Truman would explain it with equal clarity to the American people. On March 12, Truman accordingly told Congress that "at the present moment in world history nearly every nation must choose between alternative ways of life. The choice is too often not a free one. . . . It must be the policy of the United States to support free people who are resisting attempted subjugation by armed minorities or by outside pressure." He asked for $400,000,000 in aid specifically for Greece and Turkey, but the Truman Doctrine thus propounded universalized the American commitment to contain the spread of Communism.

The Truman Doctrine

The mobilization of American might for this task followed swiftly. On June 5, 1947, at Harvard University, Marshall called for a massive program of foreign aid to help the European states recover. In July, Kennan, signing himself "X," educated the public on "The Sources of Soviet Conduct" and outlined the strategy of containment in the journal *Foreign Affairs.* The National Military Establishment Act of 1947 (in the works since the war) created a permanent Joint Chiefs of Staff, a single secretary of defense, the U.S. Air Force as a separate service with its nuclear-armed Strategic Air Command, and the Central Intelligence Agency (CIA). Kennan himself soon criticized the Truman Doctrine as indiscriminate and excessively military. Drawing on classical geopolitics, he narrowed U.S. interests to the protection of those industrial regions not yet in the hands of the U.S.S.R. (North America, Britain, Germany, and Japan). In practice, however, defense of those regions seemed to require defense of contiguous areas as well. Japanese security, for instance, depended on the fate of Korea, and European security on not being outflanked in the Middle East. American responsibilities, therefore, could easily appear to be global.

The Marshall Plan was born in the State Department in response to the fact that western Europe was making little progress toward prosperity and stability. Britain was exhausted and committed to the Labour government's extensive welfare programs. In France, Charles de Gaulle's postwar government quickly gave way to a Fourth Republic paralyzed by quarreling factions that included a large, disciplined Communist party. In Italy, too, Communists threatened to gain power by parliamentary means. All suffered from underproduction, a shortage of capital, and energy shortages exacerbated by the severe winter of 1946–47. Marshall therefore put forward a plan for cash grants to a joint European economic council "to assist in the return of normal economic health, without which there can be no political stability and no assured peace."

The Marshall Plan

The British foreign secretary, Ernest Bevin, spoke for western Europe when he told Parliament, "When the Marshall proposals were announced, I grabbed them with both hands." At Kennan's insistence, Marshall aid was offered to all of Europe, including the Soviet bloc, but Stalin denounced the plan as a capitalist plot. The one eastern European state not yet communized, Czechoslo-

vakia, attempted to join the Marshall Plan, but Communist pressure forced it to back out. In February 1948, less than 10 years after Munich, the Czech Communist party subverted the republic and Czech democracy again fell to totalitarian rule, a tragedy punctuated by the suicide—or murder—of Foreign Minister Jan Masaryk. Stalin reinforced his attack on the Marshall Plan by reviving the Communist International, now called the Communist Information Bureau (Cominform), in October 1947, and by escalating ideological warfare against the West.

Communist Czechoslovakia

The new hope kindled in western Europe by the Marshall Plan helped secure the defeat of the Communists in the 1948 Italian election (the $1,000,000 of CIA funds for the Christian Democrats was hardly decisive) and stabilize politics elsewhere in western Europe. Under the Marshall Plan, the United States then transferred $13,600,000,000 to the stricken economies of western Europe in addition to $9,500,000,000 in earlier loans and $500,000,000 in private charity.

The division of Europe. The Marshall Plan's manifold effects included the hardening of the division of Europe, the movement for integration within western Europe, and the creation of the two Germanies. "Bizonia," the product of an economic merger between the U.S. and British occupation zones, was announced on May 29, 1947, and a new U.S. policy followed on July 11 that ended Germany's punitive period and aimed at making its economy self-sufficient. When in March 1948 some of the western European states responded to the coup in Czechoslovakia by signing the Brussels Treaty and pressing ahead with the establishment of a West German currency and government, the Russians walked out of the Allied Control Council. On June 24 Soviet occupation forces in the eastern zone blocked Allied road and rail access to the western zones of Berlin. This first Berlin crisis, made possible by the anomaly of a U.S.-British-French interest 100 miles inside the Soviet zone, forced Truman to define the limits of his "get tough" policy. Clay and Acheson advocated sending an armed convoy along the access routes to assert Allied rights, but neither the Joint Chiefs, nor the British and French, were prepared to risk war. Instead, the United States responded with an enormous airlift, totaling 277,264 sorties, to keep western Berlin supplied with food, fuel, and medicine. Perhaps Stalin hoped to drive the Allies from Berlin, or to prevent the setting up and possible rearmament of a West German state, or to induce the American electorate in 1948 to return to isolationism. In the event, the blockade only frightened the Western powers into stronger new measures. On April 4, 1949, the foreign ministers of the United States, Britain, France, Belgium, The Netherlands, Luxembourg, Italy, Portugal, Denmark, Iceland, Norway, and Canada founded the North Atlantic Treaty Organization (NATO) in Washington, providing for mutual aid in case of attack against any member. On May 8, the West German parliamentary council adopted a constitution, and on May 23 the Federal Republic of Germany came into being. Stalin acknowledged defeat in Berlin and lifted the blockade on May 12, but the Soviets countered by creating mirror institutions—the German Democratic Republic (Oct. 7, 1949) and the Council for Mutual Economic Assistance (Comecon) in the Soviet bloc.

The Berlin airlift and NATO

The parallel and hostile German states and regional alliances institutionalized and militarized the Cold War even as the Communist ideological offensive and the Truman Doctrine had universalized it. But before this first phase of the Cold War closed, two events called into question some root assumptions of the two sides. The first was the West's assumption that Communism was a monolithic movement controlled from the Kremlin. For in June 1948 the world became aware of a rift between Stalin and Tito that threatened to shake the Soviet empire of "people's democracies." This rift could be traced to the war, in which Tito's Communist partisans expelled the Nazis from Yugoslavia without large-scale aid from the Soviet Union. As a national hero, Tito had strong domestic support and thus was not personally dependent on Stalin. He even persevered in support for the Greek Communists while Stalin was adhering to his 1944 agreement with Churchill to keep hands off Greece. When Stalin and Molotov vetoed his plans for a Balkan confederation, Tito purged Yugoslav Communists known to be in the pay of Moscow. Stalin countered with brutal threats and a purge of Communists in the satellites accused of Titoist tendencies. But Tito held firm: Yugoslavia would "choose its own path to socialism," seek economic ties with the West, and indirectly place itself under Western protection. Tito also ceased to support the Greek Communists, and the civil war there soon ended in a victory for the royal government (October 1949).

Tito's rebellion

The second assumption of the early Cold War was shattered in August 1949 when the Soviet Union exploded its first atomic bomb. Its development might have been hastened by espionage, but Soviets had been among the leaders in nuclear physics before the war, and knowledgeable observers had known that a Soviet atomic bomb was only a matter of time.

THE COLD WAR IN THE MIDDLE EAST AND ASIA

The creation of Israel. Islāmic and south Asian nationalism, first awakened in the era of the first World War, triumphed in the wake of the second, bringing on in the years 1946–50 the first great wave of decolonization. The British and French fulfilled their wartime promises by evacuating and recognizing the sovereignty of Egypt, Jordan, Lebanon, and Syria in 1946 and Iraq in 1947. (Oman and Yemen remained under British administration until the 1960s, Kuwait and the Trucial States until 1971.) The strategic importance of the Middle East derived from its vast oil reserves, the Suez Canal, and its position on the southern rim of the U.S.S.R. While the Islāmic kingdoms and republics were not drawn to Communist ideology, the Soviets hoped to expand their influence by pressuring Turkey and Iran and involving themselves in the intramural quarrels of the region. Chief among these was the Arab-Israeli dispute.

The Zionist movement of the late 19th century had led by 1917 to the Balfour Declaration by which Britain promised an eventual homeland for Jews in Palestine. When that former Ottoman province became a British mandate under the League of Nations in 1922, it contained about 700,000 people, of whom only 58,000 were Jews. By the end of the 1920s, however, the Jewish community had tripled, and Arab resentment exploded in bloody riots in 1929 and again in 1936–39 with the encouragement of Amīn al-Ḥusaynī, grand mufti of Jerusalem and admirer of the Nazis. For self-protection the Jews formed Haganah (Defense), an underground militia that by 1939 had grown into a semiprofessional army. The Zionist cause then began to benefit from the worldwide sympathy caused by the Nazi holocaust and by Haganah co-belligerency in the British war against Germany. But the Irgun Zvai Leumi (National Military Organization), a Zionist terror organization under Menachem Begin, and the Abraham Stern Group, which found even the Irgun too mild, turned against the British occupation in 1944 despite vehement opposition from Chaim Weizmann and others promoting the Jewish cause overseas. The newly formed Arab League, in turn, pledged in March 1945 to prevent the formation of any Jewish state in Palestine.

Zionist terror campaigns

Meanwhile, Zionists concentrated on the United States, whose large Jewish voting bloc was believed likely to influence policy. In the 1944 campaign Roosevelt endorsed the founding of a "free and democratic Jewish Commonwealth," and U.S. policy subsequently clashed with Britain's, which aimed at maintaining paramountcy in the region through good relations with the Arabs. Foreign Secretary Bevin opposed and Truman endorsed a proposal in April 1946 by an Anglo-American Committee of Inquiry to allow another 100,000 Jews into Palestine, an idea dwarfed by David Ben-Gurion's demand for 1,200,000. Jewish terrorism exacerbated British hostility through such incidents as the flogging and murder of British soldiers, culminating in the bombing of the King David Hotel on July 22, 1946, in which 41 Arabs, 28 British, and 22 others died. All told, Jewish terrorists killed 127 British soldiers and wounded 331 from 1944 to 1948, as well as thousands of Arabs. On the other hand, heartrending tales of Jewish

survivors of Nazi Europe being turned back from their "promised land" also tugged at Western consciences.

On April 2, 1947, Bevin washed his hands of Palestine and placed it on the docket of the UN, which recommended partition into Jewish and Arab states. The United States and Britain feared that the Arabs would turn to the Soviets for aid, but the U.S.S.R. mystified all parties in October by agreeing with the American plan for partition. The Soviets apparently hoped to hasten British withdrawal, insinuate themselves into Middle Eastern diplomacy, and profit from the discord following partition. The General Assembly approved partition on November 29, granting to Jews some 5,500 square miles, mostly in the arid Negev. When the Arab League proclaimed a jihad (holy war) against the Jews, Truman's advisers began to reconsider partition, for the loss of Arab oil might cripple the Marshall Plan and the U.S. military in case of war. But when the British pulled out and Ben-Gurion declared the state of Israel on May 14, 1948, Stalin and Truman (whether out of sympathy or domestic politics) immediately advanced recognition.

Partition of Palestine

At the moment of partition the number of Jews had risen to some 35 percent of the total population of Palestine, and they were faced with Arab League forces totaling 40,000 men. The Haganah fielded about 30,000 volunteers armed with Czech weapons sent at the behest of the U.S.S.R. On the day after partition the Arab League launched its attack, but the desperate Jewish defense prevailed on all five fronts. The UN called for a cease-fire on May 20 and appointed Count Folke Bernadotte as mediator, but his new partition plan was unacceptable to both sides. A 10-day Israeli offensive in July destroyed the Arab armies as an offensive force, at the cost of 838 Israeli lives. Members of the Stern Group assassinated Bernadotte on September 17. A final offensive in October carried the Israelis to the Lebanese border and the edge of the Golan Heights in the north, to the Gulf of Aqaba and into the Sinai in the south. Armistice talks resumed on Rhodes on Jan. 13, 1949, with the American Ralph Bunche mediating, and a truce followed in March. But no Arab state recognized Israel's legitimacy. Over a half-million Palestinian refugees were scattered around the Arab world. Between 1948 and 1957 some 567,000 Jews were expelled from Arab states, nearly all of whom resettled in Israel. The 1948 war thus marked only the beginning of trouble in the region.

South Asia. The British faced a similar problem on a much larger scale in India, whose population included 250,000,000 Hindus, 90,000,000 Muslims, and 60,000,000 distributed among various ethnic and religious minorities. Between the wars Mohandas Gandhi's passive-resistance campaigns had crystallized Indian nationalism, which was nurtured in part by the relative leniency of British rule. Parliament set in motion the process leading to home rule in 1935, and the Attlee Cabinet rewarded India for its wartime loyalty by instructing Lord Louis Mountbatten on Feb. 20, 1947, to prepare India for independence by June 1948. He did so, too hastily, in only six months, and the partition of the subcontinent into a mainly Hindu India and a mainly Muslim but divided Pakistan (including part of Bengal in the east) on Aug. 15, 1947, was accompanied by panicky flight and riots between Hindus and Muslims that claimed between 200,000 and 600,000 lives. Perhaps a bloodbath was inevitable whatever Mountbatten did or however long he took to do it. But nothing tarnished Britain's colonial record in India so much as its termination. The Congress Party of Jawaharlal Nehru then took firm control and governed the Dominion (after 1950 the Republic) of India in parliamentary style and made India one of the first decolonized states to adopt a posture of nonalignment among the great powers. Disputes with Pakistan, especially over the contested province of Jammu and Kashmir, however, ensured continued strife on the subcontinent.

Partition of India and Pakistan

Elsewhere in South Asia the colonial powers expelled the Japanese only to confront indigenous nationalist forces. The British fought a successful counterinsurgency against Communist guerrillas in Malaya, but the French waged a protracted and ultimately unsuccessful war with Communist Viet Minh in Indochina, while the Dutch failed to subdue nationalists in Indonesia and granted independence in 1949. The United States transferred power peacefully in the Philippines in 1946.

In Japan, the American occupation under General Douglas MacArthur effected a peaceful revolution, restoring civil rights, universal suffrage, and parliamentary government, reforming education, encouraging labour unions, and emancipating women. In the 1947 constitution drafted by MacArthur's staff Japan renounced war and limited its military to a token force. During the Korean War a majority of the Allies signed a separate peace treaty and the United States entered into a mutual security pact with Japan (Sept. 8, 1951). These policies laid the foundation for a peaceful and prosperous Japan, but the United States took upon itself the burden of defending the western Pacific for the foreseeable future.

The Chinese civil war. The Asian future would be determined above all by the outcome of the civil war in China, a war that had never totally ceased even during the Japanese invasion and occupation. In 1945, Truman reaffirmed America's commitment to a "strong, united, and democratic China" and dispatched Marshall to seek a truce and a coalition government between Chiang Kai-shek's Nationalists at Chungking and Mao Zedong's Communists in Yen-an. But neither side had any intention of compromising with the other, and fighting resumed in October 1946. At first the United States imposed an arms embargo, but after May 1947 it extended aid to Chiang—a policy aptly described as "neutrality against the Communists."

Stalin, having blundered badly in China in the 1920s, kept up correct relations with the Nationalists on the assumption that Chiang was too strong to defeat but not strong enough to defy Soviet interests in Manchuria, Mongolia, and Sinkiang. The U.S.S.R. concluded a treaty of friendship with the Nationalist government on Aug. 14, 1945. Soviet policy at that time was to depict Mao as a mere agrarian reformer and to call for a coalition government. Having won Chiang's blessing, the Soviets systematically looted Manchuria of industrial equipment and reassumed their old rights on the Chinese Eastern railway. At the same time, Molotov insisted that the United States withdraw its advisers.

Chiang's forces advanced on all fronts until they captured Yen-an itself in March 1947. But the rapid occupation of north China and Manchuria, with American aid but against American advice, overextended the Nationalist army and tied it to cities and railroad lines. Corrupt officers also sold vast numbers of U.S. weapons to the enemy and siphoned off much of the $2,000,000,000 in U.S. aid into personal fortunes. When the Communists counterattacked at the end of 1947, Nationalist units were left isolated in the cities or simply melted away. The Communists took Tientsin and Peking in January 1949 and opened a southward offensive in April. By June their army had grown to 1,500,000 men and Chiang's had shrunk to 2,100,000. On August 5 a State Department White Paper announced the cessation of all aid to the Nationalists and concluded that "the ominous result of the civil war in China is beyond the control of the government of the United States." The remaining Nationalists fled to the island of Formosa (Taiwan), and the Communists officially proclaimed the People's Republic of China at Peking on Oct. 1, 1949. Only then did Stalin recognize the Maoist regime and negotiate to return Port Arthur and the Manchurian railway to Chinese control.

Isolation of the Nationalists

The fall of China to Communism, following hard on the Berlin Blockade and the first Soviet A-bomb test, was a terrific blow to the United States. The disaster gave Republicans a stick with which to beat the Truman administration, while the perjury of Alger Hiss (a high-ranking State Department officer, president of the Carnegie Endowment for World Peace, and erstwhile Communist agent) lent credence to charges that Communist sympathizers were at work in Washington. On Feb. 9, 1950, Senator Joseph R. McCarthy claimed to know the identities of 205 State Department officials tainted by Communism. Over the course of four years of congressional hearings McCarthy used innuendo and intimidation to propound charges that,

in virtually every case, proved groundless. But the tide of suspicion he incited—or exploited—ironically made him, as Truman said, "the greatest asset that the Kremlin has." For not only did his behaviour besmirch the image of the United States, but it bequeathed the charge of "McCarthyism" as an impregnable defense to be used by all manner of leftists.

"Who lost China?"

The original question—Who lost China?—had been answered by the White Paper: America was not omnipotent and China was not America's to lose. But misperception of Asian realities and the "Europe-first" bias of the East Coast establishment, most Democrats, and the Army certainly contributed to the debacle. "Asia-firsters," including the much less influential West Coast establishment, most Republicans, and the Navy, rued the equanimity with which the administration witnessed the collapse of the Nationalists. For his part, Stalin must have found it equally mysterious that the United States would go to the brink of war over Berlin and spend billions to aid western Europe, then stand aside while the world's most populous nation went Communist and shrug that it would "wait for the dust to settle" (Acheson's phrase).

The Korean War. Events in neighbouring Korea determined that the dust would not settle for another 20 years. In 1945, Soviet and American troops had occupied the peninsula, ruled by Japan since 1895, on either side of the 38th parallel. In North Korea indigenous Marxists under Kim Il-sung took control with Soviet assistance and began to organize a totalitarian state. In South Korea, General John R. Hodge, lacking firm instructions from Washington, began as early as the autumn of 1945 to establish defense forces and police and to move toward a separate administration. He also permitted the return of nationalist leader Syngman Rhee. By the time Washington and Moscow noticed Korea, the Cold War had already set in and the de-facto partition, as in Germany, became permanent. South and North Korean governments formally arose in 1948, each claiming legitimacy for the whole country and threatening to unify Korea by force. Between October 1949 and June 1950 several thousand soldiers were killed in border incidents along the parallel. The war that followed, therefore, was not so much a new departure as a denouement.

On Jan. 12, 1950, Acheson outlined his Asian policy in a speech before the Press Club in Washington, D.C. He included Japan, Okinawa, and the Philippines within the American line of defense but excluded Taiwan and Korea.

North Korean invasion

Five months later, on June 25, 1950, the North Koreans invaded across the 38th parallel. Conventional wisdom had it that Kim Il-sung was acting on Stalin's orders and that Acheson's omission had "invited" the attack. But the declassification of documents of the period has led to a reconsideration of the question of the origins of the Korean War. The United States had not ignored Korea; rather, the State Department considered South Korea vital to the defense of Japan. It is more likely that Acheson's failure to mention Korea meant that the United States did not intend to station its own forces in Korea, unlike the countries mentioned, and that the United States was purposely withholding unequivocal support from Rhee lest he take it as encouragement to invade the north. Thus, Acheson was trying to prevent a war, but probably trying also to ensure that if hostilities did occur, the Communists would be to blame. Perhaps that is why he later referred to North Korea's attack, not as an act of perfidy or aggression, but as one of stupidity.

Stalin always behaved toward his client states with similar caution and strove to keep them under control. Why then should he "unleash" Kim and expose North Korea to a U.S. counterattack that might become a precedent for pushing Communism back elsewhere? The possibility exists that Kim (like Ho Chi Minh) acted on his own in pursuit of a united national Communist state. On the other hand, Stalin may indeed have encouraged North Korea to attack in order to keep Kim—and Mao—dependent on the U.S.S.R. or to create a costly diversion for the Americans. According to Khrushchev's memoirs, Kim initiated the idea of invading and Stalin, almost casually and certainly foolishly, approved it.

The Truman administration responded with alacrity, viewing Korea as a test case for the policy of containment. The United States appealed to the Security Council (which the Soviets were boycotting for its continued seating of Nationalist China) and obtained a condemnation of North Korea and an affirmation of collective security. Once the South Korean rout was evident, Truman ordered MacArthur to transfer forces from Japan to Korea, where they barely established a perimeter around the port of Pusan. Against Senator Robert A. Taft's protest of Truman's actions as a usurpation of Congress' right to declare war, most Americans accepted Truman's analogy with the 1930s and his determination not to appease the aggressor. Ultimately 16 United Nations member states provided troops for this "police action," but U.S. and South Korean troops bore the brunt of the fighting.

In September 1950, following MacArthur's brilliant amphibious landing at Inch'ŏn, Truman approved operations north of the 38th parallel, and soon UN forces were driving through North Korea toward the Yalu river border with China. When the UN General Assembly adopted a U.S. resolution (October 7) to establish a unified, democratic Korea, it appeared that the Western alliance was going beyond containment to a "rollback" strategy: Communists who attacked others ran the risk of being attacked themselves. But in November, contrary to MacArthur's confident predictions, Chinese forces attacked across the Yalu. By the new year UN armies had retreated south of the 38th parallel and MacArthur demanded the right to expand the war. If American boys were dying, he asked, how could the government in good conscience fail to attack the enemy's home base or use every weapon at its disposal? But Prime Minister Attlee, speaking for the allies, strongly opposed a wider war or the use of nuclear weapons. By April 1951 the UN forces had recaptured Seoul and regained the 38th parallel.

UN invasion of North Korea

The effects of the Korean War reverberated around the world. Europeans feared that Korea was a diversion and that Stalin's real aim was to attack in Europe. Accordingly, Acheson agreed in September 1950 to contribute U.S. divisions to a NATO army under the command of General Eisenhower. "Asia-firsters" objected strenuously to this and kicked off what was known as "the great debate." Herbert Hoover even called for the United States to write off western Europe and make the Western Hemisphere the "Gibraltar of Western Civilization." But the Truman administration, backed by eastern Republicans and Eisenhower himself, persuaded Congress to commit four additional divisions to Europe. The Korean War also hastened implementation of NSC-68, a document drafted by Paul Nitze that called for a vigorous program of atomic and conventional rearmament to meet America's global commitments.

As American and allied publics grew increasingly impatient with the bloody deadlock in Korea, Truman determined to seek a negotiated peace. MacArthur tried to undermine this policy, issuing his own ultimatum to Peking and writing Congress that "there is no substitute for victory," whereupon Truman fired him for insubordination in April 1951. The popular warrior and proconsul went home to a hero's welcome, and the Senate held hearings on the propriety of the "limited war" strategy. Marshall defended the President, arguing that a wider war in Asia would expose Europe to attack, while General Omar Bradley insisted that MacArthur's plans would "involve us in the wrong war, at the wrong place, at the wrong time and with the wrong enemy." MacArthur retorted that limited war was a form of appeasement.

Truman and MacArthur

Truce negotiations opened at Kaesŏng on July 10 after the Chinese dropped their demands for withdrawal of all foreign troops from Korea and admission of the People's Republic to the UN in place of Nationalist China. The talks broke off in August, then resumed at P'anmunjŏm in October. Bitter fighting continued for two more years as each side sought to improve its tactical position. The talks centred on two issues: the demarcation line between North and South Korea and the repatriation of more than 150,000 Chinese and North Korean prisoners of war, many of whom did not want to return home. After hinting that

the United States might resort to use of the atomic bomb, newly elected president Dwight Eisenhower achieved an armistice signed at P'anmunjŏm on July 27, 1953, that separated the armies with a demilitarized zone and otherwise restored the *status quo ante bellum.* But Chinese torture of U.S. prisoners and anti-American propaganda, combined with U.S. refusal to recognize the Peking regime and the conclusion of a defense treaty with Nationalist China (Taiwan), ensured continued hostility between Washington and Peking. Indeed, documents declassified in the late 1980s showed that Truman and Eisenhower saw early on the potential for a Sino-Soviet split and that maximum pressure on Peking, not conciliation, was the way to bring it on.

Asian wars and the deterrence strategy. While war raged in Korea, the French were battling the nationalist and Communist Viet Minh in Indochina. When a French army became surrounded at Dien Bien Phu in 1954, Paris appealed to the United States for air support. American leaders viewed the insurgency as part of the worldwide Communist campaign and first propounded the theory that if Indochina went Communist, other Southeast Asian countries would also fall "like dominoes." But Eisenhower was reluctant to send U.S. troops to Asian jungles, to arrogate war-making powers to the executive, or to sully the anti-imperialist reputation of the United States, which he considered an asset in the Cold War. In any case both he and the American people wanted "no more Koreas." Hence the United States supported partition of Indochina as the best means of containing the Viet Minh, and after French Premier Pierre Mendès-France came to power promising peace, partition was effected at the Geneva Conference of 1954. Laos and Cambodia won independence, while two Vietnams emerged on either side of the 17th parallel: a tough Communist regime under Ho Chi Minh in the north, an unstable republic in the south. National elections intended to reunite Vietnam under a single government were scheduled for 1956 but never took place, and when the United States assumed France's former role as South Vietnam's sponsor, another potential "Korea" was created.

The two Vietnams

The Korean War and the new administration brought significant changes in U.S. strategy. Eisenhower believed that the Cold War would be a protracted struggle and that the greatest danger for the United States would be the temptation to spend itself to death. If the United States were obliged to respond to endless Communist-instigated "brushfire wars" it would soon lose the capacity and will to defend the free world. Hence Eisenhower and Secretary of State John Foster Dulles determined to solve "the great equation," balancing a healthy economy with only what was essential by way of military force. Their answer was a defense policy whereby the United States would deter future aggression with its airborne nuclear threat. As Dulles put it, the United States reserved the right to reply to aggression with "massive retaliatory power" at places of its own choosing. In implementing this policy, Eisenhower cut overall defense spending by 30 percent over four years but beefed up the Strategic Air Command. The diplomatic side of this new policy was a series of regional pacts that linked the United States to countries ringing the entire Soviet bloc. Truman had already founded the NATO alliance, the ANZUS pact with Australia and New Zealand (1951), the Pact of Rio with Latin-American nations (1947), and the defense treaty with Japan (1951). Now Dulles completed an alliance system linking the 1954 Southeast Asia Treaty Organization (SEATO), stretching from Australia to Pakistan, to the 1955 Baghdad Pact Organization (later the Central Treaty Organization [CENTO]), stretching from Pakistan to Turkey, to NATO, stretching from Turkey (after 1952) to Iceland.

Regional alliances

Dulles viewed the postwar world in the same bipolar terms as had Truman and, for that matter, Stalin. But Asian independence not only expanded the arena of Cold War but also spawned the third path of nonalignment. In April 1955 delegates from 29 nations attended the Bandung (Indonesia) Afro-Asian Conference, which was dominated by Nehru of India, Gamal Abdel Nasser of Egypt, and Sukarno of Indonesia. In theory the delegates met to celebrate neutrality and an end to "the old age of the white man"; in fact they castigated the imperialist West and praised, or tolerated, the U.S.S.R. Although most of the Bandung leaders were sloganeering despots in their own countries, the movement captivated the imagination of many guilt-ridden Western intellectuals.

The nonaligned nations

THE PACE OF EUROPEAN INTEGRATION

The nature and role of Germany. The shared horror of World War II and the decline of Europe from the seat of world powers into an arena of U.S.-Soviet competition revived the ancient dream of European unity. In modern times, Roman Catholics, liberals, and Socialists had all conceived of one means or another to transcend nationalism, and after 1945 a combination of factors made the dream plausible. First, the Soviet threat gave western Europeans an incentive to unite for defense and economic recovery. Second, the very scale of the superpowers suggested that Europeans must pool their resources if they hoped to play a major role in world affairs. Third, two world wars and the Fascist interlude had discredited nationalism and propelled moderate Christian Democrats and Social Democrats to prominence in postwar Europe. Fourth, integration was a means by which German economic and military power might be safely revived. Fifth, centralized planning, which had evolved naturally with the war economies, made economic integration seem possible and attractive. Finally, the United States used its leverage through the Marshall Plan to encourage multinational institutions, cooperation, and free trade.

In early disputes over the occupation of Germany, France often sided with the U.S.S.R. in order to keep Germany weak and obtain reparations. But the Berlin crisis of 1948 convinced the French that a way must be found to reconcile German recovery with their own security. The architects of an integrationist solution were the French technocrat Jean Monnet and Foreign Minister Robert Schuman. The Schuman Plan of May 1950 called for a merger of west European coal and steel industries to hasten recovery, forestall competition, and make future wars between France and Germany impossible. The patriarchal chancellor of the new West German republic, Konrad Adenauer, embraced the offer at once, for the primary foreign policy goal of his new state was economic and political rehabilitation. The founding of the West German state was his first success; the drafting of a sturdy democratic constitution was the second; his adoption, with Ludwig Erhard, of a dynamic free-market economic policy was the third. Once Marshall Plan aid arrived, West Germany was well on its way to the *Wirtschaftswunder,* the economic miracle of the 1950s. But it still remained for Adenauer to achieve security and full sovereign rights for West Germany. The Cold War permitted him to do both at once. By moving West Germany into the democratic free-market camp he earned protection and trust from the West. Of course, Adenauer could not ignore the emotional issue of German reunification, and thus he refused to recognize the East German regime or Polish control of the lands east of the Oder-Neisse rivers. The Hallstein Doctrine extended this non-recognition to all countries that did recognize East Germany. But Adenauer knew that to base policy on the prospect of reunification was unrealistic. The Soviets' Prague Proposals of October 1950 had envisioned a united, demilitarized German state—Kennan now endorsed such a neutral zone in central Europe to separate the Cold War rivals—but the Soviets insisted on a Constituent Council with equal representation for East and West Germany, even though the West had twice the population. At best, the East German delegation could block progress indefinitely while preventing West Germany from joining the Western bloc. At worst, the Soviets might subvert or coerce a disarmed Germany into aligning with Moscow. In the atmosphere of the Korean War, the Prague Proposals could not be taken up with confidence.

The Schuman Plan

The Prague Proposals

Instead, Adenauer endorsed the Schuman Plan and helped to found the European Coal and Steel Community among "the Six": France, West Germany, Italy, and the Benelux countries. The Korean War sparked the next initiative toward integration when the United States, bogged

down in Asia, requested a sizable increase in the European contribution to NATO. In 1951 the French and British cabinets both fell over the costly issue of rearmament before a committee managed to work out an acceptable distribution of burdens in October. But the obvious solution was German rearmament, something the nervous French refused to countenance unless the German army were merged into an international force, a European Defense Community (EDC). The implications were profound, for a common western European army would require a common defense ministry, coordinated foreign policy, joint defense budget, even a common parliament to approve spending and policy. In sum, the EDC would go far toward creating a United States of Europe. The West German parliament was first to ratify the EDC, in March 1953, but Britain, still clinging to the vestiges of empire and its "special relationship" with the United States, opted out. As Anthony Eden put it, joining a European federation "is something which we know, in our bones, we cannot do." The French, in turn, debated the issue until Stalin's death and the Korean armistice eroded the sense of emergency. French Communists, of course, opposed the EDC, while Gaullists blanched at merging France's proud services into a European potpourri. Despite Dulles' threat of an "agonizing reappraisal" of U.S. policy should the EDC fail, the French parliament voted it down on Aug. 30, 1954. An alternate solution quickly followed: West Germany was simply admitted to NATO and its *Bundeswehr* (armed forces) placed under Allied command. The Soviets responded in 1955 by creating the Warsaw Pact, a military alliance of the U.S.S.R. and its eastern European satellites.

NATO and the Warsaw Pact

Postwar European recovery. The first postwar decade was one of anxiety and crisis for Europe but one also of astounding economic recovery. Thanks to rational planning, labour–management cooperation, emphasis on production, the Marshall Plan, and the very destructiveness of the war, which made new plant construction necessary and thorough, the members of the Organisation for European Economic Co-operation all exceeded their prewar production levels by 1950 and achieved an annual average growth rate of 5 to 6 percent through 1955. The political stability wrought by the Cold War and the Western alliance and by the American military umbrella, which permitted Europeans to devote more resources to building the welfare state, made for unprecedented prosperity.

Eastern Europe also recovered from the war, but more slowly and not always to its own benefit. In the late 1940s the U.S.S.R. forced one-sided trade treaties on its satellites so that Polish and Romanian foodstuffs and Czech and East German technology flowed to the U.S.S.R. rather than to world markets. Stalin's death on March 5, 1953, sparked hopes for a thaw in the eastern bloc and in the Cold War. The ephemeral collective leadership that succeeded him executed the hated secret-police chief, Lavrenty Beria, and released thousands from prison camps. Riots in East Germany and Poland also induced Moscow to scale back its exploitation of the satellites and reduce reparations from East Germany. A Soviet delegation even visited Belgrade in 1955 to attempt a reconciliation with Tito. That same year the Austrian State Treaty provided for the first Soviet military withdrawal since the war and brought into being a neutral Austrian state.

In 1956 Nikita Khrushchev emerged as the new Soviet premier and shocked the 20th Party Congress with his midnight speech denouncing Stalin's "cult of personality" and manifold crimes against the party. But de-Stalinization, however carefully undertaken, created a crisis of legitimacy for the Soviet empire. In the summer of 1956 Władisław Gomułka rose to leadership of the Polish Communist Party on a wave of strikes and riots. When Moscow received his reassurances and allowed him to stay in power, other eastern Europeans were tempted to test the limits of de-Stalinization. The Hungarians reached them in October 1956 after the reformist premier Imre Nagy was deposed and protests spread that Soviet troops already on the scene were unable to quell. Nagy returned to power to announce the end of the one-party state and to release the Roman Catholic primate, József Cardinal Mindszenty, from his long imprisonment. Nagy also promised freedom of speech and the withdrawal of Hungary from the Warsaw Pact. While Hungary's fate hung in the balance, the Western powers had their attention diverted by a second Middle Eastern war.

The Hungarian uprising

The Suez Crisis. The Arab states, after their defeat in 1948, passed through a period of political unrest. The most critical change occurred in Egypt, where in 1952 a cabal of young army officers backed by the Muslim Brotherhood forced the dissolute King Farouk into exile. In 1954 Nasser emerged to assume control. Nasser envisioned a pan-Arab movement led by Egypt that would expel the British from the Middle East, efface Israel, and restore Islāmic grandeur. Egypt began sponsoring acts of violence against Israel from the Gaza Strip and cut off shipping through the Strait of Tiran. The British were understandably hostile to Nasser, as were the French, who were battling Islāmic nationalists in Morocco, Algeria, and Tunisia.

Israel had used the years since 1948 to good effect, developing the arid country and training a reserve force of 200,000 men and women armed primarily with French weapons. David Ben-Gurion believed that the Arabs would never accept the existence of Israel except by force. U.S. policy was to play down the Arab–Israeli dispute and alert all parties to the danger of Communist penetration. To this end, Eisenhower dispatched a futile mission in January 1956 in hopes of reconciling Cairo and Tel Aviv. In addition, the United States agreed to contribute $56,000,000, and $200,000,000 through the World Bank, to Egypt's project for an Aswān Dam on the Nile. But Nasser's flirtations with Moscow alienated Dulles. Then, on July 26, 1956, Nasser nationalized the Suez Canal.

The conservative Cabinet in London, the French, and the Israelis resolved to thwart Nasser. They could cite as precedent a CIA-backed coup d'état in Iran (August 1953) that overthrew the ascetic nationalist Mohammad Mosaddeq, who had expropriated foreign oil interests and also looked for support to the U.S.S.R. In any case, British, French, and Israeli planners met to work out a joint strike at the Sinai and Suez that might permit a far-reaching realignment in the Middle East. Eisenhower got wind of Israeli military preparations but believed that the blow would fall on Syria. He especially opposed hostilities before the U.S. election lest he lose Jewish votes by having to scold Israel. But Moshe Dayan quietly mobilized all of Israel's mobile brigades, which struck on October 29 and took the Egyptians—and the Americans—by surprise. Israeli war aims included the elimination of the Egyptian army as an offensive threat, neutralization of Palestinian bases in Gaza, and capture of the Strait of Tiran. The Anglo-French goals were to secure the Suez Canal and possibly to topple Nasser and thus strike a blow at Arab radicalism.

The CIA in Iran

An Israeli airborne assault secured the Mitla Pass in the Sinai while armoured columns penetrated the peninsula. The Anglo-French then issued an ultimatum to Cairo and proceeded to bomb Egyptian bases. The Egyptian army evacuated the Sinai. Eisenhower, preoccupied with Hungary and the election, was furious at this act of insubordination on the part of his allies and sponsored a UN resolution for a cease-fire on November 1. Egypt frustrated the Anglo-French plan by the simple expedient of scuttling ships in the canal, but the Anglo-French went ahead with a landing at Port Said. The superpowers then forced an evacuation and the insertion of UN peacekeeping forces in the Sinai and Gaza Strip. There matters stood for 10 years.

The only one who gained in the Suez muddle was the U.S.S.R. With the West in disarray and involved in a campaign that looked very much like old-fashioned imperialism, Soviet tanks returned to Budapest on November 4, crushed the Hungarians fighting with their homemade weapons, and liquidated their leaders. In 1957 the Soviets declared a new policy of "centralism" for the satellites and denounced both "dogmatism" (a code word for Stalinism) and "revisionism" (a code word for liberty).

The events of October 1956 nevertheless helped to renew momentum for European integration. Hungary reminded western Europeans of the nature and proximity of the

Soviet regime; Suez made them resentful of American tutelage. Inspired by Monnet and Belgian economist Paul-Henri Spaak, "the Six" drafted the Euratom Treaty for a joint nuclear energy agency and the Treaty of Rome to expand the coal and steel community into a full-fledged Common Market. The treaties were signed on March 25, 1957, and went into effect on Jan. 1, 1958. The European Economic Community provided for internal and external tariff coordination, free movement of labour and capital, and a common agricultural pricing policy. Integration theorists hoped that international economic institutions would sustain a momentum leading to political unity as well.

The Treaty of Rome

NUCLEAR WEAPONS AND THE BALANCE OF TERROR

The race for nuclear arms. The postwar arms race began as early as 1943, when the Soviet Union began its atomic program and placed agents in the West to steal U.S. atomic secrets. When the U.S.S.R. rejected the Baruch Plan in 1946 and U.S.–Soviet relations deteriorated, a technological race became inevitable. The years of the U.S. monopoly, however, were a time of disillusionment for American leaders, who discovered that the atomic bomb was not the absolute weapon they had first envisioned. First, the atomic monopoly was something of a bluff. As late as 1948 the U.S. arsenal consisted of a mere handful of warheads and only 32 long-range bombers converted for their delivery. Second, the military was at a loss as to how to use the bomb. Not until war plan "Half Moon" (May 1948) did the Joint Chiefs envision an air offensive "designed to exploit the destructive and psychological power of atomic weapons." Truman searched for an alternative, but balancing Soviet might in conventional forces with a buildup in kind would have meant turning the United States into a garrison state, an option far more expensive and damaging to civic values than nuclear weapons. A few critics, notably in the navy, asked how a democratic society could morally justify a strategy based on annihilation of civilian populations. The answer, which had been evolving since 1944, was that U.S. strategy aimed at deterring enemy attacks in the first place. "The only war you really win," said General Hoyt Vandenberg, "is the war that never starts."

Nuclear deterrence, however, was subject to at least three major problems. First, even a nuclear attack could not prevent the Soviet army from overrunning western Europe. Second, the nuclear threat was of no use in cases of civil war, insurgency, and other small-scale conflicts, a fact Stalin evidently relied on in several instances. Third, the U.S. monopoly was inevitably short-lived. By 1949 the Soviets had the atomic bomb, and the British joined the club in October 1952. The United States would be obliged to race indefinitely to maintain its technological superiority.

The first contest in that race was for the "superbomb," a hydrogen, or fusion, bomb a thousand times more destructive than the atomic fission variety. Many scientists opposed this escalation. The dispute polarized the political and scientific communities. On the one hand it seemed as if the Cold War had created a climate of fear that no longer permitted principled dissent even on an issue involving human survival; on the other hand, it seemed as if the dissenters, inadvertently or not, were promoting the interests of the U.S.S.R. In January 1950, Truman gave his approval to the H-bomb project, and the first fusion bomb was tested successfully at Enewetak Atoll in November 1952. No debate occurred in the Soviet Union, where scientists moved directly to fusion research and exploded their first bomb in August 1953.

The hydrogen bomb

In the meantime, Soviet agitprop agencies laboured abroad to weaken Western resolve. A prime target was NATO, which the Kremlin evidently viewed as a political threat (since its inferior order of battle was scarcely an offensive military threat). After 1950 the Soviets alternately wooed the western Europeans with assurances of goodwill and frightened them with assurances of their destruction if they continued to host American bases. Cominform parties and front organizations (such as the World Peace Council) denounced the Pentagon and U.S. "arms monopolies" and exploited fear and frustration to win over intellectuals and idealists. The Stockholm Appeal of 1950, initiated by the French Communist physicist Frédéric Joliot-Curie, gathered petitions allegedly signed by 273,470,566 persons (including the entire adult population of the U.S.S.R.). Similar movements organized marches and protests in Western countries against nuclear arms (no such manifestations occurred in the Soviet bloc).

Eisenhower's defense policy brought a sharp increase in research and development of warheads and long-range bombers and the construction of air bases on the territory of allies circling the U.S.S.R. But the H-bomb breakthrough also triggered a race to develop intercontinental ballistic missiles (ICBMs). The United States entered the postwar era with an advantage in long-range rocketry, thanks to the suspension of the Soviet program during the war and the decision by the Germans' V-2 rocket team, led by Wernher von Braun, to surrender to the U.S. Army. But in the budget-cutting of the late 1940s, the Truman administration surmised that the United States, possessed of superior air power and foreign bases, did not need long-range guided missiles. The first atomic weapons, bulky and of limited yield, also suggested that no rocket large and accurate enough to destroy a target 6,000 miles distant was then possible. But the vastly greater yield of fusion bombs and the expectation of smaller warheads changed that calculation. The U.S. ICBM project received top priority in June 1954. The Soviets, by contrast, needed to find a means of threatening the United States from Soviet soil. As early as 1947, therefore, Stalin gave priority to ICBM development.

The missile race

Arms control and defense. How could the arms race be headed off before the world became locked into what Churchill called "the balance of terror"? The UN Disarmament Commission became a tedious platform for the posturings of the superpowers, the Americans insisting on on-site inspection, the Soviets demanding "general and complete disarmament" and the elimination of foreign bases. Eisenhower hoped that Stalin's death might help to break this deadlock. Churchill had been urging a summit conference ever since 1945, and once de-Stalinization and the Austrian State Treaty gave hints of Soviet flexibility, even Dulles acquiesced in a summit, which convened at Geneva in July 1955. The Soviets again called for a unified, neutral Germany, while the West insisted that it could come about only through free elections. On arms control, Eisenhower stunned the Soviets with his "open skies" proposal. The United States and the Soviet Union, he said, should exchange blueprints of all military installations and each allow the other side to conduct unhindered aerial reconnaissance. After some hesitation, Khrushchev denounced the plan as a capitalist espionage device. The Geneva summit marginally reduced tensions but led to no substantive agreements.

The "open skies" proposal

"Open skies" reflected the American fear of surprise attack. In 1954 a high-level "Surprise Attack Study" chaired by the scientist James Killian assured the president of a growing American superiority in nuclear weapons that would hold until the 1958–60 period but warned that the U.S.S.R. was ahead in long-range rocketry and would soon achieve its own secure nuclear deterrent. The panel recommended rapid development of ICBMs, construction of a distant early warning (DEW) radar line in the Canadian Arctic, strengthened air defense, and measures to increase intelligence-gathering capabilities, both to verify arms control treaties and to avoid overreaction to Soviet advances. The Killian report gave birth to the U-2 spy plane, which began crisscrossing the U.S.S.R. above the range of Soviet air defense in 1956, and to a research program to develop reconnaissance satellites to observe the U.S.S.R. from outer space.

In 1955 both the United States and the Soviet Union announced programs to launch artificial Earth satellites during the upcoming International Geophysical Year (IGY). The Eisenhower administration, concerned that the satellite program not interfere with military missile programs or prejudice the legality of spy satellites to come, entrusted its IGY proposal to the small, nonmilitary Vanguard rocket. While Vanguard development crept ahead, the Soviet program won the first space race with Sputnik

1 on Oct. 4, 1957. The Soviet achievement shocked the Western world, challenged the strategic assumptions of every power, and thus inaugurated a new phase in the continuing Cold War.

Total Cold War and the diffusion of power, 1957–72

The concomitant arrival of the missile age and of an independent and restive Third World multiplied the senses in which politics had become global. Intercontinental rockets not only meant that the most destructive weapons known could now be propelled halfway around the world in minutes, but also, because of the imminent nuclear standoff they heralded, that a Cold War competition would now extend into other realms—science and technology, economic growth, social welfare, race relations, image making—in which the Soviets or Americans could try to prove that their system was the best. At the same time, the decolonization of dozens of underdeveloped states in Asia and Africa induced the superpowers to look beyond the original front lines of the Cold War in Europe and East Asia.

These technological and political revolutions would seem to have raised the United States and the Soviet Union to unequaled heights of power. The Soviets and Americans advanced rapidly in the high technology required for spaceflight and ballistic missiles, while techniques for the mobilization and management of intellectual and material resources reached a new level of sophistication, especially in the United States, through the application of systems analysis, computers, bureaucratic partnership with corporations and universities, and Keynesian "fine-tuning" of the economy. By the mid-1960s the vigorous response of the Kennedy and Johnson administrations to the Cold War challenge seemed to ensure American technological, economic, and military primacy for the forseeable future. A mere five to seven years later, however, it became clear that the 1960s, far from establishing an American hegemony, had in fact wrought a diffusion of world power and an erosion of the formerly rigid Cold War blocs. Western Europe and Japan, now recovered from the war, also achieved dynamic economic growth in the 1960s, reducing their relative inferiority to the United States and prompting their governments to exercise a greater independence. The Sino-Soviet split, perhaps the most important event in postwar diplomacy, shattered the unity of the Communist bloc, and Third World countries often showed themselves resistant to superpower coercion or cajoling. By 1972 the U.S.S.R., despite its achievement of relative parity in nuclear weapons, was obsessed with the prospect of a hostile China, while the United States, having squandered its wealth, prestige, and domestic tranquillity in the Vietnam War, was trying to scale back its global commitments. The Nixon Doctrine, détente with Moscow, the opening to China, and uncoupling of the dollar from gold were the symptoms of this American retreat.

The American watershed

THE WORLD AFTER SPUTNIK

Soviet progress and American reaction. Premier Khrushchev anticipated the new correlation of forces in his foreign policy address to the 20th Party Congress in 1956. Soviet H-bombs and missiles, he said, had rendered the imperialists' nuclear threat ineffective, the U.S.S.R. an equal, the socialist camp invincible, and war no longer inevitable, and thus "peaceful coexistence" inescapable. In Leninist doctrine this last phrase implied a state of continued competition and socialist advance without war. The immediate opportunities for socialism, according to Khrushchev, derived from the struggle of the colonial peoples, which the U.S.S.R. would assist through foreign aid, propaganda, subversion, and support for "wars of national liberation."

The Soviet successes in outer space just 40 years after the Bolshevik Revolution were powerful evidence for Khrushchev's claims that the U.S.S.R. had achieved strategic equality and that Communism was the best system for overcoming backwardness. Sputnik restored Soviet prestige after the 1956 embarrassment in Hungary, shook European confidence in the U.S. nuclear deterrent, magnified the militancy of Maoist China, and provoked an orgy of self-doubt in the United States itself. The two Sputnik satellites of 1957 were themselves of little military significance, and the test missile that launched them was too primitive for military deployment, but Khrushchev claimed that long-range missiles were rolling off the assembly line "like sausages," a bluff that allowed President Eisenhower's opponents—and nervous Europeans—to perceive a "missile gap." Khrushchev in turn tried to capitalize on the apparent gap in a series of crises. But his adventurous policy only provoked perverse reactions in China, the United States, and Europe that undermined his own political support at home.

Soviet missile bluff

Eisenhower was apprised in advance of Soviet missile progress thanks in part to overflights of the U-2 spyplane. By the time of Sputnik the Pentagon already had several parallel programs for ballistic missiles of various types, including the advanced, solid-fueled Polaris and Minuteman. The great fleet of B-47 and B-52 intercontinental bombers already deployed also assured continued American strategic superiority through the early 1960s. The frugal Eisenhower thus tried to play down the importance of Sputnik and to discourage a race for arms or prestige, but he was frustrated by a coalition of Democrats, journalists, academics, and hawks of both parties who insisted that the United States not only leapfrog the Soviets in space and missiles but also increase federal support to education, extend more military and economic aid to the Third World, and expand social programs at home intended in part to polish the American image abroad—in short, pursue the Cold War more vigorously. Eisenhower conceded to this mood in 1958 by sponsoring creation of the National Aeronautics and Space Administration and passage of the National Defense Education Act, accelerating weapons programs, and deploying intermediate-range missiles in England, Italy, and Turkey. He also acknowledged the expanded Soviet threat in his State of the Union Address in 1958: "Trade, economic development, military power, arts, science, education, the whole world of ideas—all are harnessed to this same chariot of expansion. The Soviets are, in short, waging total cold war." But a similarly total American response to this challenge, requiring virtually wartime levels of national mobilization to outdo a totalitarian system in whatever field of endeavour it chose to emphasize would, in Eisenhower's mind, have undermined the free market and fiscal soundness that were the foundation of American strength in the first place. Liberal economists argued in response that a sharply expanded role for the federal government was a matter of survival in the "space age" and would even stimulate economic growth, military prowess, and social progress.

Creation of NASA

The Sino-Soviet split. A still more energetic U.S. riposte would await the end of Eisenhower's term, but "Mr. Khrushchev's boomerang" (as John Foster Dulles termed Sputnik) had an immediate and disastrous impact on Soviet relations with the other Communist giant, China. Under their 1950 treaty of friendship, solidarity, and mutual assistance, Soviet technical aid flowed to Peking during the Korean War and helped support China's successful Five-Year Plan after 1953. Western observers looked in vain for ways to split the Communist bloc. But as early as 1956 Chinese leaders showed displeasure over Khrushchev's denunciation of Stalin, the Kremlin's tendency to treat the Chinese party as it did those of the lesser satellites, and the new Soviet leaders themselves, whom Mao evidently considered mediocrities. Mao also denounced "peaceful coexistence" as decadent and revisionist, a position shared by the tiny Stalinist dictatorship of Albania. Russian leadership in the world Communist movement was thus challenged for the first time.

Mao was a romantic revolutionary with an unquestionable bent for cruel or irrational theatrics on a gigantic scale. In the mid-1950s he paraded the slogan "Let a Hundred Flowers Bloom," ostensibly to encourage the voicing of new ideas on national development but perhaps rather to entice potential dissenters into revealing themselves. In 1958 this campaign was suddenly replaced by the "Great Leap Forward," by which all 700,000,000 Chinese were to form self-sufficient communes devoted to local industrial-

The Great Leap Forward

ization. Large-scale industries and infrastructure collapsed, much to the disgust of Soviet guest engineers. By 1960–61 the economic chaos had become so severe that famine claimed 6,000,000–7,000,000 lives. Nevertheless, the Chinese leadership seized upon Sputnik as proof that the "East wind" was prevailing over the "West wind" and insisted that the Soviets use their new superiority to press the revolution worldwide and, to the same end, provide China with atomic bombs and rockets. If the imperialists insisted on unleashing nuclear war, lectured Mao, and "half of mankind died, the other half would remain, while imperialism would be razed to the ground and the whole world become socialist." The Soviets were appalled, especially since their superiority was, for the time being, a sham. At a November 1958 summit Mao learned that the Soviets would insist on retaining control over any warheads sent to China and would not share missile technology. When the Soviets also failed to back the Chinese in their 1958–59 conflicts with Taiwan and India, Sino-Soviet tensions increased. In the end Khrushchev refused to deliver a prototype nuclear warhead, whereupon the Chinese angrily repudiated "slavish dependence" on others and pledged to create their own nuclear arsenal. On July 16, 1960, the U.S.S.R. recalled all its specialists from China.

The Sino-Soviet split shattered the strict bipolarity of the Cold War world (though the United States would not take advantage of that fact for more than a decade) and turned the U.S.S.R. and China into bitter rivals for leadership in the Communist and Third worlds. The fundamental causes of the split must be traced to contradictions in the Soviet role as both the leader of the Communist movement and as a great power with its own national interests. Before 1949 the U.S.S.R. had been able to subordinate the interests of foreign Communists to its own, but the Communist triumph in China, paradoxically, was a potential disaster for the U.S.S.R., for Mao and the Chinese would inevitably refuse to play the role of pupil. Once the Korean War was over and Stalin dead, the Chinese asserted themselves, learned the limits of "socialist internationalism," and angrily began to plot their own course. While the ideological rift served, in the short run, to invigorate both Communist rivals as they competed for prestige and influence among the world's revolutionaries, it destroyed the myth that Communism transcended nationalism and power politics. It also meant that the U.S.S.R. was delicately situated between the nuclear-armed NATO powers and the fanatical (and numerous) Chinese, and to appease either meant to alienate the other. Accordingly, Khrushchev played a risky double game from 1958 to 1962, alternately holding out hope for arms control to the NATO powers and leveling demands backed by rocket-rattling. The historian Adam Ulam has seen in this a "grand design" by which Khrushchev hoped to ingratiate himself with the West (for instance, through a nuclear test-ban treaty) in return for the evacuation of West Berlin, recognition of the East German government, and permanent denial of nuclear weapons to West Germany—all of which might demonstrate Soviet commitment to the Communist cause while providing a pretext for denial of nuclear weapons to China. But whether a grand design or an improvisation, Soviet diplomacy had to reckon at every turn with Peking's reactions and their likely effect on the rest of the Communist bloc.

Soviet diplomatic offensive. The Polish foreign minister, Adam Rapacki, was chosen to open Moscow's post-Sputnik campaign with a proposal to the UN General Assembly in October 1957 for a ban on nuclear weapons in Poland, Czechoslovakia, and the two Germanies. This initiative, like others of the sort before and after, was a no-lose strategem for the U.S.S.R. Given the Warsaw Pact's superiority in conventional weapons, any reduction of the West's nuclear deterrent in Europe stood to weaken NATO, even as the burden of seeming to oppose arms control would fall on the West if it refused. At the same time, the U.S.S.R. combined open and covert support for Western antinuclear movements with loud reminders of its ability to destroy any nation that foolishly hosted American bases. NATO leaders resisted the Rapacki Plan but had immediately to deal with a March 1958 Soviet offer to suspend all nuclear testing provided the West did the same. Throughout the 1950s growing data on the harmful effects of nuclear fallout had been increasing pressure on the nuclear powers to take such a step. The United States and Britain were caught in the midst of testing warheads for the many new missiles under development, but a one-year test ban did go into effect in November 1958. With the Chinese making noises about a Soviet sellout to the West, however, Khrushchev immediately provoked a new crisis in Berlin, demanding that the Allies withdraw from West Berlin within six months. Khrushchev also indicated that the best way to solve the Berlin question would be to neutralize and disarm the two German states. In January 1959 the Soviets expanded their proposed nuclear-free zone to include the Far East and the whole Pacific Ocean area—a clear hint of their desire to prevent China from going nuclear.

The Rapacki Plan

The Berlin deadline passed without incident as Khrushchev accepted an invitation to become the first Soviet premier to visit the United States. The increased recognition by the United States and the U.S.S.R. that each had interests in coexistence which outweighed their ideological loyalties was made manifest in August 1958, when Chinese artillery began an intense bombardment of the Nationalist-held offshore islets of Quemoy and Matsu. Peking may have hoped to force Moscow to support its claim to sovereignty over Taiwan, while Chiang may have hoped to drag the United States into supporting an invasion of the mainland. But neither superpower was willing to risk war. The U.S. 7th Fleet resupplied Chiang's forces, while the Soviets pledged to defend mainland China, but both discouraged offensive action.

Quemoy and Matsu

By September 1959, when Khrushchev arrived in the United States, Dulles had died, and Eisenhower was intent to use personal diplomacy in an attempt to put a cap on the arms race. The tour itself—from New York City to Iowa to Hollywood—was a sensation, though Khrushchev professed distaste for American consumerism and predicted "your grandchildren will live under Communism." His talks with Eisenhower produced an ephemeral "spirit of Camp David" and the scheduling of a followup summit conference for May 1960 in Paris. Meanwhile, Khrushchev's last-ditch efforts to mend relations with Peking exploded in the spring of 1960. Mao himself reportedly authored an article cryptically condemning Khrushchev's détente policy as vile revisionism and reiterating Chinese willingness to confront nuclear war. The Chinese observer at a Warsaw Pact meeting in February 1960 declared in advance that any arms agreements reached at the U.S.–Soviet summit would not be binding on Peking. On the eve of the Paris summit an American U-2 spy plane was shot down over the U.S.S.R. When Eisenhower refused to apologize for the incident and assumed personal responsibility, Khrushchev had little choice but to walk out.

Decolonization and development. Events in the other new arena of the post-Sputnik era—the Third World—likewise antagonized relations among the U.S.S.R., the United States, and China. All three assumed that the new nations would naturally opt for the democratic institutions of their mother countries, or on the other hand would naturally gravitate toward the "anti-imperialist" Soviet or Maoist camps. The United States had urged Britain and France to dismantle their empires in the aftermath of World War II, but once those countries became Washington's most potent allies in the Cold War, the United States offered grudging support for Anglo-French resistance to nationalist and Communist forces in their colonies. President Truman's Point Four Program mandated U.S. foreign aid and loans to new nations lest they "drift toward poverty, despair, fear, and the other miseries of mankind which breed unending wars." When the Eisenhower administration cut back on foreign aid, a great debate about its efficacy ensued among American experts. Critics insisted that the Marshall Plan was not a valid analogy for Third World aid since the former had been a case of helping industrial populations rebuild their societies, while the latter was a case of sparking industrial or even merely agricultural development in primitive economies. Foreign

The Point Four Program

aid did not necessarily serve U.S. interests, since many Third World rulers chose neutralism or socialism, nor did it promote economic growth, since most new nations lacked the necessary social and physical infrastructure for a modern economy. Proponents of aid replied that U.S. capital and technology were needed precisely to build infrastructure, to assist "nation building," and to fortify recipients against Communists and others who might subvert the development process in its early stages. In the late 1950s, U.S. economic aid averaged about $1,600,000,000 per year, compared to about $2,100,000,000 in military aid to friendly regimes. The Soviet line, by contrast, held that new nations would not be truly independent until they freed themselves from economic dependence on their former masters. But the Soviets invariably expected a political return for their own assistance. The claim of the People's Republic of China to be the natural leader of Third World revolt also obliged Khrushchev to make bolder endorsements of wars of national liberation. By 1960 it was already clear, however, that local politics and culture made every Third World situation unique.

The Middle East had reached an unstable deadlock based precariously on the UN-administered cease-fire of 1956. But the eclipse of British and French influence after the Suez debacle made the United States fearful of growing Soviet influence in the region, symbolized by the Soviet offer to take over construction of the Aswān Dam in Egypt. In January 1957 the U.S. Congress authorized the president to deploy U.S. troops in the region if necessary and to dispense $500,000,000 in aid to friendly states. This Eisenhower Doctrine appeared to polarize the region, with Middle East Treaty Organization members in support and Egypt, Syria, and Yemen (which formed a short-lived United Arab Republic in 1958) in opposition. When, in July 1958, nationalist generals backed by a variety of factions, prominent among which were Communists, overthrew the pro-Western Hāshimite monarchy in Iraq, and unrest spread to Jordan and Lebanon, Eisenhower responded at once. The 14,000 U.S. troops that landed in Beirut allowed the Lebanese president to restore order on the basis of a delicate compromise among radical, Muslim, and Christian factions. Khrushchev denounced the intervention, demanded that the U.S.S.R. be consulted, and tried without success to convene an international conference on the Middle East. His extension of an invitation to India, and not China, needlessly alienated Peking and signaled a new Soviet interest in relations with New Delhi.

The climactic year of African decolonization was 1960, and the first Cold War crisis on that continent occurred when, in that year, Belgium hastily pulled out of the vast Congo. Tribal antagonisms and rival personalities made even the independence ceremonies a catastrophe, as the Congolese nationalist leader and first prime minister, Patrice Lumumba, supported an insurrection by Congolese army units that involved the murder of whites and blacks alike. No sooner had Belgian troops returned to restore order than Moise Tshombe declared the secession of the iron-rich Katanga province. UN Secretary General Dag Hammarskjöld intervened against the Belgians and Katangese (thereby setting an ominous precedent of UN toleration for black violence against blacks or other races), while the Soviets accused Tshombe of being a dupe for imperialist mining interests and threatened to send arms and Soviet "volunteers" to the leftist Lumumba. Hammarskjöld then organized a UN armed force to subdue Katanga and save the Congo—and Africa—from Cold War involvement. The clumsy UN efforts did not prevent, and may have incited, the spread of civil war. Lumumba tried to establish his own secessionist state, but he then fell into the hands of the Congolese army headed by Joseph Mobutu, a former sergeant, and was murdered by the Katangans in January 1961. Hammarskjöld himself died in a plane crash in the Congo in September 1961. UN troops remained until 1964, but as soon as they were withdrawn rebellion returned and Mobutu seized control in a military coup d'état in 1965. The Katangan revolt was not quelled until 1967.

The Congo crisis

In Southeast Asia the Geneva Accords disintegrated rapidly after 1954. The planned elections to reunify Vietnam were never held, since South Vietnam's leader, Ngo Dinh Diem, both feared the results and denied the possibility of free elections in the Communist north. Ho Chi Minh's regime in Hanoi then trained 100,000 native southerners for guerrilla war and launched a campaign of assassination and kidnapping of South Vietnamese officials. In December 1960 the Viet Cong (as Diem dubbed them) proclaimed the formation of a National Liberation Front (NLF), with the avowed aim of reuniting the two Vietnams under a Hanoi regime. American advisers tried vainly to arrest the disintegration of South Vietnam with advice on counterinsurgency and state-building techniques.

Guerrilla war in Vietnam

In neighbouring Laos the Communist Pathet Lao took control of the two northernmost provinces of the country in defiance of the neutral government under Prince Souvanna Phouma agreed upon after Geneva. Those provinces sheltered the Ho Chi Minh Trail supply route bypassing the demilitarized zone between the two Vietnams. When a new, assertive Laotian government sent troops to enforce its authority over the provinces in 1958–59, civil war appeared inevitable. A military coup d'état led by Kong Le briefly returned Souvanna Phouma to power, but when Kong Le was in turn driven out in December 1960, he joined forces with the Pathet Lao in their strategic stronghold in the Plain of Jars. Having secured the Laotian territory needed for infiltration and assault on South Vietnam, North Vietnam persuaded China and the U.S.S.R. in December 1960 to approve Ho's plan for a "nonpeaceful transition to socialism" in Vietnam.

Latin-American problems. Finally, Cold War rivalry and Third World problems intersected devastatingly in the Americans' own backyard. Before the era of Roosevelt's Good Neighbor Policy, the United States had frequently been accused of meddling too much in the affairs of other states in the hemisphere. By the 1950s the contradictory charge was leveled that the United States was not involving itself enough, as evidenced by the fact that the United States spent $12,600,000,000 on aid to Asia and the Middle East in the period 1953–57 compared with $1,900,000,000 on Latin America. But resentment over the CIA's role in thwarting a Communist-backed coup in Guatemala in 1954 and violent protests against Vice President Richard M. Nixon during his trip to Caracas and Lima in 1958 alerted Washington to the dangers inherent in neglecting the genuine needs of the region. The United States agreed to fund an Inter-American Development Bank, while the State Department sought to avoid too close an association with unpopular, authoritarian regimes. Whatever the overall merits of such a policy, it had immediate and disastrous effects in Cuba.

In 1952 Fulgencio Batista established a corrupt dictatorship in Cuba, and four years later a young revolutionary named Fidel Castro took to the Sierra Maestra with 150 comrades and made pretensions of fighting a guerrilla war. In fact, Castro's campaign was largely propaganda (the insurgents lost only 40 men in the largest engagement), and the real struggle for Cuba was fought out in the arenas of Cuban and American public opinion. After Nixon's tour, liberal opinion and the State Department deserted Batista, and the new ambassador to Havana was ordered to preside over his fall. In March 1958 the United States suspended arms sales to Cuba, and on Jan. 1, 1959, a triumphant Castro entered Havana without the necessity of fighting a battle. Contrary to his image as populist and a democrat, Castro made himself the new dictator, nationalized hundreds of millions of dollars worth of American property, and declared that he was and had always been a Marxist. His actions gradually alienated whatever sympathy he had in the United States. Castro invited Soviet aid and came to rely on it heavily after the United States curtailed Cuba's sugar import quota in July 1960. Eisenhower instructed the CIA to explore means of removing Castro, who made Cuba into an immensely valuable Soviet satellite 90 miles from the United States.

Castro's victory

By 1960, therefore, the post-Sputnik world posed new challenges for the Western alliance stretching from outer space to Third World jungles. Polls showed that a majority of western Europeans believed Khrushchev's propaganda about Soviet superiority and that a majority of Americans

no longer believed in Eisenhower's low-key approach to Cold War issues.

SUPERPOWER RELATIONS IN THE 1960S

Policies of the Kennedy administration. The inauguration of John F. Kennedy as president of the United States infused American foreign policy with new style and vigour. He had promised to "get America moving again," and he appointed a Cabinet and staff who shared his belief that the United States could be doing far more to prove its technological and moral superiority over the U.S.S.R., win the "hearts and minds" of Third World peoples, and accelerate social progress at home. His administration also overturned Eisenhower's policy on economy and defense and held that Keynesian fiscal policy and large programs for research, education, and human resources would foster the rapid growth needed to pay for the new federal activism. Kennedy's inaugural address was thus an exhortation and warning: "Let every nation know, whether it wishes us well or ill, that we shall pay any price, bear any burden, meet any hardship, support any friend, oppose any foe to assure the survival and the success of liberty." He and Secretary of Defense Robert McNamara accordingly increased the U.S. defense budget by 30 percent in their first year in office and approved deployment of a strategic triad of weapons—the land-based Minuteman ICBMs, submarine-launched Polaris missiles, and B-52 bombers. The Kennedy advisers had also been highly critical of the policy of reliance on massive retaliation and determined to make the United States capable of flexible response by expanding conventional armed forces as well. Kennedy paid special attention to the training of counterinsurgency "special forces."

U.S. strategic triad

On May 25, 1961, Kennedy told a joint session of Congress that "the great battlefield for the defense and expansion of freedom today is the whole southern half of the globe—Asia, Latin America, Africa, and the Middle East." The enemies of freedom were seeking to capture these rising peoples "in a battle of minds and souls as well as lives and territories." Expanded aid programs, the Peace Corps, active promotion of democracy through the U.S. Information Agency, and military support against guerrilla warfare would, he declared, all help in cases "where the local population is too caught up in its own misery to be concerned about the advance of Communism." Kennedy also underscored the impact of the Soviet space program on world opinion (Yuri Gagarin had become the first man to orbit the Earth on April 12) and asked that Congress commit the United States to a program to land a man on the Moon by 1970. Kennedy's call for the creation of an International Telecommunications Satellite Consortium bespoke his desire to associate the United States with the peaceful uses of outer space.

The new attitude toward the Third World was perhaps the clearest break in American diplomacy. Basing its policy on W.W. Rostow's "non-Communist manifesto" describing stages of economic development, the Kennedy administration increased foreign aid for Third World nations whether or not they were politically aligned with the United States. The Alliance for Progress, created in March 1961, especially targeted Latin America. By 1965 U.S. foreign aid reached $4,100,000,000 as compared with $2,300,000,000 contributed by all other developed countries. The validity of Rostow's investment model for economic "takeoff" was debated for two decades, but perhaps the greatest weakness in U.S. aid programs was the assumption that local rulers could be persuaded to put their own people's welfare first. Instead, aid money often fed corruption, bolstered power-hungry leaders or socialist bureaucracies, or helped to finance local conflicts. What was more, the Soviets had some natural advantages in dealing with such leaders, since they offered no moralistic advice about democracy and human rights, while their own police-state methods served the needs of local despots. On the other hand, sustained world economic growth and measures to stabilize commodity prices helped the developing countries to achieve an average annual growth rate of 5 percent during the 1960s (compared with 5.1 percent for industrial countries). But the crushing rate of Third World population growth (2.6 percent annually) meant that even in the best of times foreign aid only just offset the effects of Third World fertility.

The development models debate

Kennedy's first crisis stemmed from his endorsement of the CIA plan to unseat Castro. The CIA had trained Cuban exiles in Guatemala and flown them to Florida, whence they were to stage an invasion of Cuba in expectation of a popular revolt there. Instead, the landing at the Bay of Pigs on April 17, 1961, was a fiasco. No coordination had been achieved with dissidents inside Cuba, while the failure to provide U.S. air cover (pehaps for fear of retaliation in Berlin) doomed the invasion. Castro's army killed or captured most of the 1,500-man force in two days. The U.S.S.R. reaped a propaganda harvest and pledged to defend Cuba in the future. Kennedy had to content himself with a promise to resist any efforts by Castro and guerrilla leader Che Guevara to export revolution elsewhere in Latin America.

The Bay of Pigs

Kennedy and Khrushchev held a summit meeting in Vienna in June 1961. With Berlin and the Third World uppermost in his mind, Kennedy proposed that neither superpower attempt to upset the existing balance of power in any region where the other was already involved. Khrushchev evidently considered the young president to be weak and on the defensive and tried to intimidate him with a new ultimatum, threatening to turn over control of western access to West Berlin to the East German government. (Khrushchev was being pressured by East German leader Walter Ulbricht to stem the tide of thousands of skilled workers who were fleeing across the zonal boundary into West Berlin.) Kennedy responded by pledging to defend West Berlin and calling up 250,000 reservists. On Aug. 13, 1961, Soviet and East German troops closed down interallied checkpoints and proceeded to build the Berlin Wall, sealing off the western city. Just as in 1948, the U.S. leadership debated whether to respond with force to this violation of the Potsdam Accords, but the hesitancy of the NATO allies and the timidity—or prudence—of Kennedy limited the West to a reassertion of access rights to West Berlin.

The Cuban missile crisis. In the midst of this crisis the Soviets unilaterally broke the moratorium on nuclear testing, staging a series of explosions yielding up to 50 megatons. Soviet technology had also perfected a smaller warhead for the new Soviet missiles now ready to be deployed, like the Minuteman, in hardened silos. Khrushchev, his nation still behind in strategic nuclear firepower, tried to redress the balance by insinuating 42 medium-range missiles into Cuba, whence they could reach most of the continental United States. He apparently hoped that these missiles, once in place, could then serve as a bargaining chip in negotiations leading to a neutralized Germany, which in turn might help Moscow persuade the Chinese to cease their own nuclear program. Instead, the ploy brought the world to the brink of war. U-2 spy planes photographed the missile sites under construction in Cuba on Oct. 14, 1962. Two days later Kennedy convened a secret crisis-management committee that leaned at first toward a surgical air strike to destroy the sites. But the President opted for a less risky response: a naval quarantine to prevent Soviet freighters from reaching Cuba and an ultimatum demanding that the bases be dismantled and the missiles removed. On October 18, Soviet Ambassador Andrey Gromyko met with Kennedy and denied that the U.S.S.R. had any offensive intentions with respect to Cuba. On October 22 the President informed the nation of the crisis and called on Khrushchev to pull back from "this clandestine, reckless, and provocative threat to world peace." For two days the world waited anxiously, and on the 24th Soviet ships in transit abruptly changed course away from Cuba. On the 26th Khrushchev sent Kennedy a message offering to withdraw the missiles in exchange for a U.S. pledge never to invade Cuba. The next day a harsher message arrived with a new demand that the United States withdraw its own missiles from Turkey. Those antiquated Jupiters, deployed in the early post-Sputnik scare, were already due for removal, but Kennedy would not do so under Soviet threat. Hence Attorney General Robert Kennedy suggested a ploy: simply reply to Khrushchev's

U.S. blockade of Cuba

first note as if the second had never been sent. On the 28th the Soviets agreed to dismantle the Cuban bases in return for a no-invasion pledge. Several months later the United States quietly removed its missiles from Turkey.

The Cuban missile crisis seemed at the time a clear victory for Kennedy and the United States and was widely attributed to American superiority in nuclear weapons. In fact, neither side showed the slightest willingness even to bluff a nuclear strike, and it was probably the overwhelming U.S. superiority in conventional naval and air power in its home waters that left the U.S.S.R. no option but retreat. Nor was the crisis an unmitigated American victory. Kennedy's pledge never to overthrow Castro by force meant that the United States would have to tolerate whatever mischief he, backed by $300,000,000 a year in Soviet aid, might contrive in the future. To be sure, Kennedy warned that the United States would never tolerate any expansion of Communism in the hemisphere. (This pledge was underwritten by Lyndon Johnson in 1965 when he sent U.S. troops into the Dominican Republic to prevent a leftist takeover, but such interventionism only reminded Latin Americans of past "Yankee imperialism" and gave credence to Castro's anti-American propaganda.) The existence of a Communist base in the Caribbean, therefore, was to be a source of unending vexation for future American presidents. What is more, the Cuban missile crisis hardened Soviet determination never again to be humiliated by military inferiority. Khrushchev and his successors accordingly began the largest peacetime military buildup in history, which, by the 1970s, accorded the Soviet Union parity with the United States in nuclear forces and the ability to project naval power into every ocean of the world.

On the other hand, the Cuban missile crisis marked the final frustration of Khrushchev's efforts to force a German peace treaty and prevent the deployment of nuclear weapons on German or Chinese soil. Peking, of course, had supported the Soviets' bid to place missiles in Cuba and had taken the opportunity to attack India (see below). But the precipitous Soviet retreat prompted Chinese charges of "capitulationism." The Chinese nuclear program proceeded apace, with the People's Republic exploding its first atomic device in 1964. Never again would the Soviet leadership hope to control the foreign policy of the other Communist giant.

Renewed U.S.–Soviet cooperation. U.S.–Soviet relations, by contrast, markedly improved after the sobering visit to the brink of war. Hopes for a comprehensive nuclear test-ban treaty ran afoul of the U.S.S.R.'s customary refusal to permit on-site inspection to monitor underground tests, but a partial Test-Ban Treaty was signed by the United States, Britain, and the U.S.S.R. on Aug. 5, 1963, prohibiting nuclear explosions in the air, under the sea, and in outer space. The superpowers also established a direct communication link between Washington and Moscow for use in crisis situations. But other powers anxious to join the nuclear club, notably China and France, refused to adhere to the Test-Ban Treaty. Instead, the Chinese denounced Soviet collaboration with "the leader of world imperialism." Mao resurrected all of China's territorial claims against the Soviet Union dating from tsarist Russian imperialism and advocated partition of the Soviet empire. The Soviets, in turn, branded Mao with their most hateful current epithet: he was "another Stalin."

Nuclear Test-Ban Treaty

President Kennedy was assassinated on Nov. 22, 1963, and Khrushchev was removed from power by the Politburo in October 1964, a victim of his own failures in foreign policy and agriculture, and of the Communist Party's resistance to his attempted reforms. But the bilateral effort to pursue arms control survived under President Johnson and under Leonid Brezhnev and Alexey Kosygin. The Outer Space Treaty ratified in 1967 banned nuclear weapons and other weapons of mass destruction in Earth's orbit and on the Moon. A U.S.–Soviet draft Non-Proliferation Treaty was also adopted by the UN in June 1968. (Once again, France, China, India, Pakistan, and Israel refused to sign.) But none of the arms-control instruments of the 1960s put a cap on the arms race or restrained the signatories from doing anything in the strategic field that they had a desire to do anyway. The superpowers were able to modernize their arsenals through underground nuclear testing; outer space was an awkward and vulnerable place to deploy warheads in any case; and neither superpower had an interest in seeing nuclear weapons spread to more countries. Rather, American nuclear policy aimed, at least in the short run, at ensuring the continued stability of U.S.–Soviet deterrence, lately dubbed "mutual assured destruction." Adopting the views of strategist Bernard Brodie, McNamara concluded early on that the Soviets must eventually catch up and that a state of parity was the best that could be achieved in the nuclear age. Soon each side would be capable of obliterating the other in a retaliatory strike, even after a sneak attack. At that point, any attempt by either side to achieve an illusory superiority would only destabilize the balance and tempt one or the other into launching a first strike. Whether the Soviets ever shared this doctrine of deterrence is dubious. Marshal Sokolovsky's volumes on military strategy in the 1960s, while granting that nuclear war would be an unprecedented disaster for all, still committed the U.S.S.R. to a war-winning capability.

Nuclear Non-Proliferation Treaty

China, meanwhile, succumbed to another series of Maoist actions that completed that country's drift into chaos and isolation. In February 1966, Mao gave the nod to the young and fanatical Red Guards to make, by force, a Cultural Revolution. Violence swallowed up schools, factories, bureaucracies, cultural institutions, and everything that smacked of foreign or traditional Chinese influence. Countless victims suffered internal exile, public humiliation, forced "self-criticism," or death, while attacks on foreign embassies and denunciations of the superpower "condominium" persuaded Americans and Soviets alike that the Chinese were, for the moment at least, the major threat to world peace.

By the late 1960s, therefore, relations between the United States and the Soviet Union underwent a marked thawing. But at the same time, the Soviets and Americans alike had to acknowledge a growing lack of control over their once coherent Cold War camps.

THE EUROPE OF THE FATHERLANDS

Great Britain and decolonization. The Suez crisis of 1956, followed by Soviet space successes and rocket-rattling after 1957, was a serious blow to the morale of western Europe. Given the potential of the war scares over Berlin to fracture NATO, the United States had to reassure its allies and try to satisfy their demands for greater influence in alliance policy. American efforts largely succeeded in the case of Britain, an ally much depleted in power and will. American policy largely failed in the case of France, an ally stronger and more stable than at any time since 1940.

Since World War II, Britain had tried to maintain the appearance of a global power, developing its own nuclear weapons, deploying conventional forces around the world, and keeping hold of its African colonies. Churchill, returned to office in the early 1950s, had vowed never to "preside over the liquidation of the British Empire." Likewise, the British held aloof from the continental experiments with integration and saw their role rather as the vertex of three great world systems: the English-speaking peoples, the British Commonwealth, and the old European Great Powers. All this came to a sudden end when a combination of factors—sluggish economic performance by the world's oldest industrial power, growing pressure to decolonize, demands for greater social expenditures at home, and the superpowers' leap into the missile age—convinced London that it could no longer afford to keep up appearances in foreign policy. A defense White Paper of 1957 signaled a shift away from conventional armed forces toward reliance on a cheap, national nuclear deterrent. Then Sputnik convinced the British government to cancel its own ballistic-missile program and rely on its special relationship with the United States to procure modern weapons. Eisenhower agreed to sell the Skybolt air-launched missile to Britain by way of healing the wounds inflicted by Suez and shoring up NATO after Sputnik. But when McNamara subsequently cut the Sky-

Britain's retreat

bolt program in his campaign to streamline the Pentagon, the British government was acutely embarrassed. Kennedy met with Prime Minister Harold Macmillan at Nassau in December 1962 and offered Polaris submarines instead. It was hoped at the time that the British deterrent would be subsumed in a multilateral NATO force. The Conservative government also made the hard decision in 1963 to seek admission to the Common Market, only to be vetoed by the French. Not until 1971 was Britain's application, together with those of Ireland and Denmark, approved and the European community broadened.

The period 1957–62 was also the climax of decolonization. As early as 1946–47, when Britain was granting independence to India and states of the Middle East, the Attlee government sponsored the Cohen–Caine plan for a new approach to West Africa as well. It aimed at preparing tropical Africa for self-rule by gradually transferring local authority from tribal chiefs to members of the Western-educated elite. Accordingly, the Colonial Office drafted elaborate constitutions, most of which had little relevance to real conditions in primitive countries that had no natural boundaries, no ethnic unity or sense of nationalism, and no civic tradition. But when the Gold Coast (Ghana) elected radical leader Kwame Nkrumah, who then demanded immediate independence and got it in 1957, the British felt unable to deny similar grants to neighbouring colonies. Britain had, in fact, when the matter was faced squarely, little desire to hang on, given the exorbitant financial and political costs of late imperialism. In 1959 the Cabinet quietly decided to withdraw from Africa as soon as it won reelection. Macmillan then announced the new policy in Cape Town on Feb. 3, 1960, when he spoke of "the winds of change" sweeping across the continent. Nigeria, Togo, and Dahomey (Benin) became sovereign states in 1960, Tanganyika (Tanzania), Uganda, and Kenya in East Africa between 1961 and 1963, and Malawi and Northern Rhodesia (Zambia) in the south in 1964. White residents of Southern Rhodesia, however, declared their own independence in defiance of London and the UN. The Republic of South Africa and the surviving Portuguese colonies of Angola and Mozambique made those portions of southern Africa the last refuges of white rule on the continent.

African decolonization

Most new African states had little more to support their pretensions to nationhood than a paper constitution, a flag, and a London-backed currency. The leaderships blamed African underdevelopment on past exploitation rather than on objective conditions, thus rejecting the American and European development theories that saw political stability as possible only within the context of economic growth. Nkrumah lectured to his Pan-African Congress in 1963 that "the social and economic development of Africa will come only within the political kingdom, not the other way around." Indeed, Africa's politicians invariably styled themselves as charismatic leaders whose political and even spiritual guidance was the prerequisite for progress. Nkrumah himself seized all power in Ghana and made himself a quasi-divine figure until the army overthrew him in 1966. Togo's government fell to a military coup in 1963, and mutinies broke out in Kenya, Uganda, and Tanganyika. In the latter country, Julius Nyerere, much admired in Europe and the United States, declared a one-party dictatorship based on his ideology of *ujamaa* (familyhood), and courted aid from Communist China. Other leaders contrived similar ideologies to justify personal rule. By 1967 black Africa had suffered 64 attempted coups d'état, many born of tribal hatreds, and most Africans had fewer political rights than under colonial rule.

With the exception of the Congo, Cold War rivalries were absent from Africa in the 1960s, while the African regimes themselves wisely declared the inviolability of their boundaries lest the artificial lines drawn by the colonial powers provoke endless warfare. When Igbo tribesmen seceded from Nigeria in 1967 and formed the rebel state of Biafra, only four African nations supported their cause. Nigeria suppressed the secession in a bloody civil war. Decolonization nonetheless had a profound effect on international relations through the medium of the UN. The three dozen or so new African states combined with those of Asia and the Soviet bloc to form a permanent majority made up mostly of one-party dictatorships nevertheless claiming moral superiority over the Western "imperialists." Thus the founders' dreams that the UN might become a "parliament of the world" and bulwark of democracy and human rights were undermined by the very process of what, with one or another degree of irony, was called liberation. Instead, the UN degenerated into a forum for polemics and a playground for intrigue.

The Biafra revolt

France's independent course. Where Britain was enervated by the advent of the missile age and the Third World, France was invigorated. The weak Fourth Republic had suffered defeat in Indochina and was embroiled in a civil war between French settlers and native Muslims in Algeria. But Charles de Gaulle was called back to power 8 months after Sputnik 1, and he set about to forestall a threatened coup d'état by the French army, stabilize French politics, end the Algerian debacle (independence was granted in 1962 in the Treaty of Évian), and restore French power and prestige in the world. His constitution for a Fifth Republic established presidential leadership and ended France's chronic political instability, itself an achievement of great value to the West. But de Gaulle's vision of France involved neither *la plus grande France* of the colonial empire nor the Atlanticist France of NATO, nor the European France of the Common Market (EEC). Rather, de Gaulle proclaimed that a France without grandeur was not France at all and set out to reestablish French military, technological, and diplomatic independence.

France's decolonization proceeded as rapidly as Britain's, culminating in 1960 with the partition and independence of French West Africa. But de Gaulle refused to exhibit any guilt or doubt about France's *mission civilisatrice* and offered the populations a choice between going it alone or joining a linguistic, monetary, and development community with the former metropole. Only Guinea elected to follow a Marxist leader who sought ties with the U.S.S.R.

In defense matters, de Gaulle bristled at NATO's reliance on the United States and publicly doubted whether the U.S. nuclear umbrella over Europe was still reliable after Sputnik. Would the Americans really risk a nuclear attack on New York City or Washington, D.C., to defend Berlin or Paris? Therefore de Gaulle accelerated the quiet development of a nuclear capacity begun under the Fourth Republic, and France exploded its first atomic bomb in 1960. He also quintupled French spending on research and development, built independent bomber, missile, and submarine forces—the nuclear *force de frappe*—and made France the third space power with the launch of an Earth satellite in 1965. Gaullist France's rebellion against the tutelage of a superpower unwilling to accord it diplomatic equality or help it develop nuclear weapons bore genuine comparison to Maoist China. And the United States, like the U.S.S.R., tried various means to rein in its obstreperous ally, first trying to dissuade France from developing nuclear weapons, then inviting it to join a multilateral nuclear force (MLF) under NATO command. First suggested in December 1960, the MLF was pushed by Kennedy and Johnson, but de Gaulle responded with contempt, while Adenauer feared to join lest he damage West German relations with France. The idea of an MLF died in 1965, and in July 1966 de Gaulle took the final step of withdrawing French armed forces from NATO (though France remained a political member of the alliance). NATO headquarters were then moved from Paris to Brussels.

The *force de frappe*

De Gaulle similarly distrusted the movement for European integration, preferring what he termed "the Europe of the fatherlands" stretching "from the Atlantic to the Urals"—the latter phrase provocatively including the European portion of the Soviet Union. He tolerated European institutions such as the EEC, but only on terms of strict French leadership in partnership with West Germany, hence his veto of Britain's application in 1963. Moreover, de Gaulle viewed European cooperative programs in atomic and space research as ways to tap foreign contributions for the improvement of French national competitiveness, not as ways for France to contribute to European unity. Adenauer eagerly accepted de Gaulle's

leadership in order to complete Germany's postwar rehabilitation and retain the EEC market for Germany's booming industry. But de Gaulle crushed any lingering hopes for European political integration by boycotting the EEC in 1965–66 rather than allow the federalist commissioner Walter Hallstein to enhance the decision-making power of the EEC parliament. Finally, de Gaulle delighted in open criticisms of American foreign policy and courted closer relations with Moscow (which in return seized upon what appeared to be an opportunity to split the alliance), culminating in the pomp of a state visit in 1966. In all these ways Gaullist policy was a constant vexation to Washington, but in the long run it was probably a boon to the Western alliance for the technological dynamism, political stability, and military might it restored to France.

De Gaulle's diplomacy

ASIA BENEATH THE SUPERPOWERS

The first rebellions against the European imperial system had occurred on the rimlands of Asia at the start of the 20th century: the Russo-Japanese War, the Indian home-rule movement, and the Chinese and Young Turk revolutions. By the 1960s the southern tier of Asian states had given birth to local systems of power and rivalry beyond the control of the Great Powers. Several factors set these nations and their conflicts apart. First, the Middle East, the Indian subcontinent, and Indochina all seethed with ethnic conflicts that had little to do with the Cold War. Second, eastern and southern Asia continued to undergo a demographic explosion that made China and India by far the most populous states in the world and non-Soviet Asia the home of 55 percent of the human race. Third, the politics of these societies, involved as they were in the awakening of vast peasant masses, the breakdown of traditional village agriculture, religious and dynastic structures, and programs for rapid modernization, did not easily fall into categories familiar to Soviet and American planners of the 1950s. Fourth, most of the Asian rim was remote from the European Soviet Union and North America, making direct intervention there expensive and risky. Nevertheless, continued Soviet efforts to win influence in the Middle East, Chinese claims to natural leadership of the poor southern half of the globe, and American attempts to preserve a structure of containment of the Communist world necessarily involved the Great Powers in Asian diplomacy. The fate of half of mankind could not, it seemed, be a matter of indifference to countries that claimed universal missions.

The Six-Day War. In the Middle East, Nasser's star began to decline in the 1960s from its post-Suez peak. The Syrian Ba'th party, though socialist, resented Nasser's assumption of Arab leadership and took the country out of the United Arab Republic in 1961. Likewise, the presence of 50,000 Egyptian troops in Yemen failed to overcome the forces supporting the Yemeni imam, who was backed in turn by Saudi Arabia. On the other hand, the Cairo Conference of 1964 succeeded in rallying pan-Arab unity around resistance to Israel's plans to divert the waters of the Jordan. Also with both eyes on Israel, the conference restored an Arab High Command and elevated the Palestinian refugees (scattered among several Arab states since 1948) to a status approaching sovereignty, with their own army and headquarters in the Gaza Strip. Syria likewise sponsored a terrorist organization, al-Fatah, whose raids against Jewish settlements provoked Israeli military reprisals inside Jordan and Lebanon. Syria was divided principally between the socialist Ba'th, led by the minority 'Alawite community that dominated the army, and pro-Nasser pan-Arabists. In 1966 a military coup established a radical Ba'thist regime, but the army itself then split into rival factions. Nasser took the initiative to prevent a rightist reversal in Syria and reassert his leadership of the Arab cause.

Armed with Soviet tanks and planes, Nasser claimed his option under the 1956 accord to demand withdrawal of UN peace-keeping forces from the Sinai. Secretary-General U Thant complied on May 19, 1967. Four days later Nasser closed the Gulf of Aqaba to Israeli shipping. The Soviets apparently urged Nasser to show moderation, while President Johnson told Israeli Foreign Minister Abba Eban to remain calm: "Israel will not be alone unless it decides to go alone." But neither superpower was able to restrain its client. When Egyptian and Iraqi troops arrived in Jordan, giving every sign of an imminent pan-Arab attack, the Israeli Cabinet decided on a preemptive strike. The Israeli air force destroyed Nasser's planes on the ground, and in six days of fighting (June 5–10) the Israeli army overran the Sinai, the West Bank of the Jordan including the Old City of Jerusalem, and the strategic Golan Heights in Syria. The UN Security Council arranged a cease-fire and passed Resolution 242, calling for a withdrawal from all occupied regions. The Israelis were willing to view their conquests (except Jerusalem) as bargaining chips but insisted on Arab recognition of the right of Israel to exist and firm guarantees against future attack. The so-called frontline Arab states were neither able (for domestic reasons) nor willing to give such guarantees and instead courted Soviet and Third World support against "U.S.–Israeli imperialism." Hence Israel remained both greatly enlarged and possessed of shorter, more defensible borders, although it did acquire the problem of administering more than a million Arabs in Gaza and the West Bank.

Israel's first strike

China, India, and Pakistan. The Indian subcontinent comprised another system of conflict focused on border disputes among India, Pakistan, and China. Nehru's Congress Party had stabilized the political life of the teeming and disparate peoples of India. The United States looked to India as a laboratory of democracy and development in the Third World and a critical foil to Communist China, and in consequence had contributed substantial amounts of aid. But the U.S.S.R. also began an effective aid program in 1955, and Nehru looked to the U.S.S.R. for support against China once the Sino-Soviet split became evident. The Peking regime had brutally suppressed the buffer state of Tibet in 1950 and disputed the border with India at several points between the tiny Himalayan states of Nepal, Bhutan, and Sikkim. American military aid to Pakistan (a member of CENTO) also gave the Indians and Soviets reason to cooperate. In 1961, when President Ayub Khan of Pakistan earnestly sought Kennedy's mediation in the dispute over Kashmir, U.S. pressure proved inadequate to bring Nehru to the bargaining table.

Nehru was humbled, however, when the Chinese suddenly attacked in force across the disputed boundaries, choosing as their moment the height of the Cuban missile crisis. Indian forces were soundly defeated, 7,000 men having been killed or captured, and the lowlands of Assam lay open to the invaders. Apparently the Chinese leadership expected a Soviet triumph in Cuba or at least a drawn-out crisis that would prevent superpower intervention in India. But the swift resolution in Cuba in favour of the United States permitted Washington to respond to Nehru's request for help. The Chinese then halted the offensive and soon afterward withdrew.

The Kennedy administration used its newly won leverage to urge Nehru to settle his quarrel with Pakistan, but the negotiations failed to overcome Hindu–Muslim antipathy and the fact that the conflict was a unifying element in the domestic politics of both countries. Pakistani troops crossed the cease-fire line in Kashmir in August 1965, and India responded by invading Pakistan proper. Both superpowers backed U Thant's personal quest for a cease-fire, and the Indians withdrew. But the U.S.S.R. was able to regain influence with New Delhi, especially after the accession to power of Nehru's daughter, Indira Gandhi. In 1971 India and the U.S.S.R. concluded a 20-year Treaty of Peace and Friendship and Cooperation, an indication of how much the United States (not to mention Britain) had lost touch with the once model Third World democracy. Pakistan, meanwhile, was in ferment. President Ayub Khan was forced to step down in 1969 in favour of Yahya Khan, while elections in 1970 polarized the geographically divided country. West Pakistan chose Zulfikar Ali Bhutto as prime minister, but densely populated East Pakistan (Bengal) voted almost unanimously for a separatist party under Mujibur Rahman. When talks between the two leaders broke down, Bhutto gambled on sending in troops and jailing the secessionists. Vicious fighting broke out in Bengal, flooding India with some 10,000,000 refugees

India–Pakistan war

and provoking Indian intervention. The Soviets cautioned restraint but clearly favoured India, while U.S. President Nixon sent a carrier task force into the Bay of Bengal and openly favoured Pakistan, influenced by Pakistan's role as intermediary between Washington and Peking. In two weeks of fighting (Dec. 3–16, 1971) the Indians defeated the Pakistanis on all fronts, and East Pakistan became the new state of Bangladesh, comprising the delta of the Indus river. Pakistan thus lost well over half its population. Once Nixon's opening to China bore fruit, the subcontinent seemed to be polarized around a U.S.S.R.–India axis and a U.S.–Pakistan–China axis, though the United States resumed aid and food shipments during the Indian famine of 1972.

Creation of Bangladesh

To the south and east of the Asian mainland lay the vast, populous archipelago of Indonesia, where another romantic revolutionary, Sukarno, had played host to the Bandung Conference of 1955. Like Nasser, Nehru, and Mao, he ruled his 100,000,000 people by vague, hortatory slogans that added up to a personal ideology with nationalist and Communist overtones. The Kennedy administration had tried to appease Sukarno with development aid and even obliged the Dutch to cede Irian Barat (Irian Jaya) in the face of Sukarno's threats in 1963. Sukarno still turned to Moscow for support and gave himself over to profligate personal behaviour and foreign adventures, most notably an attempted attack on Malaysia in 1963. By 1965 Indonesia was $2,400,000,000 in debt and suffering widespread famine. In January of that year Sukarno withdrew his country from the UN over a dispute with Malaysia. The Soviets were clearly disgusted with Sukarno's regime, while the rival Chinese persuaded (perhaps blackmailed) him into approving a savage pro-Communist putsch in October 1965. General Suharto, however, put down the uprising and exacted a violent revenge in which as many as 300,000 Communists and their supporters were killed. Indonesia subsequently concerned itself with its internal problems, frustrating Soviet, Chinese, and American hopes for a strong ally.

Revolt and repression in Indonesia

The destruction of Indonesian Communism, achieved without the slightest American effort, was a source of great comfort for the United States, for a diametrically opposite course of events had, by 1965, begun to unfold in the last theatre of Asian conflict, Vietnam.

THE WAR IN SOUTHEAST ASIA

Cold War assumptions and the quagmire. As the Vietnam war began to recede into the past, the entire episode, from a neutral perspective, increasingly came to seem incredible. That the most powerful and wealthy nation on Earth should undertake 15 years of wasting conflict against a tiny state 10,000 miles from its shores—and lose—almost justifies historian Paul Johnson's phrase, "America's suicide attempt." Yet the destructive and futile U.S. engagement in Southeast Asia was a product of a series of trends that had been maturing since World War II. The early Cold War gave rise to U.S. leadership in the containment of Communism. Decolonization then thrust the United States into a role described by advocate and critic alike as "the world's policeman"—protector and benefactor of the weak new governments of the Third World. The potential of guerrilla insurgency, demonstrated in Tito's resistance to the Nazis and especially in the postwar victories of Mao, the Viet Minh, and Castro, made it the preferred mode for revolutionary action around the world. The emerging nuclear stalemate alerted Washington to the need to prepare for fighting limited (sometimes called "brushfire") wars sponsored by the Soviet Union or China through proxies in the Third World. In this era of Khrushchevian and Maoist assertiveness the United States could not allow any of its client states to fall to a Communist "war of national liberation" lest it lose prestige and credibility to Moscow and Peking. Finally, the "domino theory," to the effect that the fall of one country would inexorably lead to the communization of its neighbours, magnified the importance of even the smallest state and guaranteed that sooner or later the United States would become entangled under the worst possible conditions. One or even all of the assumptions under which the United States became involved in Vietnam may have been faulty, but very few in the government and the public questioned them until long after the country was committed.

The domino theory

By 1961, Diem's fledgling government in South Vietnam was receiving more U.S. aid per capita than any other country except Laos and South Korea. Authoritative reports detailed both the Viet Cong's campaign of terror against government officials in the south and widespread discontent over Diem's corrupt and imperious rule. In the face of both Khrushchev's renewed vow to support wars of national liberation and de Gaulle's warning ("I predict you will sink step by step into a bottomless military and political quagmire"), Kennedy chose Vietnam as a test case for American theories of state building and counterinsurgency. He approved a proposal by W.W. Rostow and General Maxwell Taylor to assign advisers to every level of Saigon's government and military, and the number of Americans in Vietnam grew from 800 to 11,000 by the end of 1962.

Ho Chi Minh's North Vietnamese considered the struggle against Diem and his American sponsors merely the next phase of a war that had begun against the Japanese and had continued against the French. Their determination to unify Vietnam and conquer all of Indochina was the principal dynamic behind the conflict. The total number of Communist troops in the South grew by recruitment and infiltration from some 7,000 in 1960 to more than 100,000 by 1964. Most were guerrilla militiamen who served also as local party cadres. Above them were the Viet Cong (formally the National Liberation Front, or NLF), deployed in regional military units, and units of the People's Army of North Vietnam (PAVN) entering the south along the Ho Chi Minh Trail. U.S. Special Forces tried to counter Communist control of the countryside with a "strategic hamlet" program, a tactic used with success by the British in Malaya. Diem instituted a policy of relocating the rural population of South Vietnam in order to isolate the Communists. The program caused widespread resentment, while Diem's persecution of local Buddhist sects provided a rallying point for protests. When Buddhist monks resorted to dramatic self-immolation in front of Western news cameras, Kennedy secretly instructed Ambassador Henry Cabot Lodge to approve a military coup. On Nov. 1, 1963, Diem was overthrown and murdered.

South Vietnam then underwent a succession of coups d'état that undermined all pretense that the United States was defending democracy. The struggle was thenceforth viewed in Washington as a military effort to buy time for state building and the training of the South Vietnamese army (Army of the Republic of Vietnam; ARVN). When two American destroyers exchanged fire with a North Vietnamese torpedo boat eight miles off the North's coast in August 1964 (an event whose occurrence was later disputed), Congress passed the Gulf of Tonkin Resolution authorizing the president to take whatever measures he deemed necessary to protect American lives in Southeast Asia. Johnson held off escalating the war during the 1964 electoral campaign but in February 1965 ordered sustained bombing of North Vietnam and sent the first U.S. combat units to the South. By June, U.S. troops in Vietnam numbered 74,000.

Gulf of Tonkin incident

The Soviet Union reacted to American escalation by trying to reconvene the Geneva Conference and bring pressure to bear on the United States to submit to the peaceful reunification of Vietnam. But China bluntly refused to encourage a negotiated settlement and insisted that the U.S.S.R. help North Vietnam by pressuring the United States elsewhere. The Soviets, in turn, resented Peking's assertion of leadership in the Communist world and had no desire to provoke new crises with Washington. The North Vietnamese were caught in the middle; Ho's ties were to Moscow, but geography obliged him to favour Peking. Hence North Vietnam joined in boycotting the March 1965 Communist conference in Moscow. The Soviets, however, dared not ignore the Vietnam War lest they confirm Chinese accusations of Soviet "revisionism."

The conduct and cost of the war. Meanwhile, the United States slid ineluctably into the quagmire predicted by de Gaulle. U.S. forces reached a peak of 543,000 men in 1969.

(Australia, New Zealand, Thailand, and the Philippines also sent small contingents, and South Korea contributed 50,000 men). The U.S. strategy was to employ mobility, based on helicopters, and firepower to wear down the enemy by attrition at minimal cost in U.S. lives.

The war of attrition on the ground, like the bombing in the North, was designed less to destroy the enemy's ability to wage war than to demonstrate to the enemy that he could not win and bring him to the bargaining table. But stalemate suited Hanoi, which could afford to wait, while it was anathema to the Americans. Johnson's popularity fell steadily. Most Americans favoured more vigorous prosecution to end the war, but a growing number advocated withdrawal. Antiwar dissent grew and spread and overlapped with sweeping and violent demands for social change. The American foreign policy consensus that had sustained containment since the 1940s was shattered by Vietnam.

In retrospect, Johnson's attempt to prevent the war from disturbing his own domestic program was vain, and his strategic conception was grounded in folly and hubris. For he and his advisers had no clear notion of what the application of American force was supposed to achieve. It was merely assumed to be invincible.

Hanoi understood that the classic Maoist strategy of isolating cities by revolutionizing the countryside was inapplicable to Vietnam because the cities could still hold out with foreign support. Accordingly, the North Vietnamese Politburo approved in mid-1967 a plan for urban attacks throughout South Vietnam. General Vo Nguyen Giap insisted, however, that NLF guerrillas, not PAVN units, be risked. The expectation was that direct attacks on cities would undercut American claims of pacification and magnify domestic American dissent. On Jan. 31, 1968 (the Tet holiday, during which many ARVN troops were home on leave), an estimated 84,000 Communist troops infiltrated South Vietnamese cities, attacked government installations, and even penetrated the American embassy in Saigon. The Tet offensive was carried out at terrible cost to Communist strength, but American press reports turned the offensive into a psychological defeat for the United States. Instead of ordering a counterattack, Johnson removed himself from the 1968 presidential campaign, ordered a bombing halt, and pledged to devote the rest of his administration to the quest for peace. Negotiations began in Paris, but the rest of the year was spent bickering over procedural issues.

The Tet offensive

For more than 25 years after 1941 the United States maintained an unprecedented depth of involvement in world affairs. In 1968 Vietnam finally forced Americans to face the limits of their resources and will. Whoever succeeded Johnson would have little choice but to find a way to escape from Vietnam and reduce American global responsibilities.

NIXON, KISSINGER, AND THE DÉTENTE EXPERIMENT

Détente as realism. After eight years in the shadow of Eisenhower and eight more years out of office, Richard Nixon brought to the presidency in 1969 rich experience as an observer of foreign affairs and shrewd notions about how to prevent the American retreat from global commitments from turning into a rout. In broad outlines, the Nixon strategy included a phased withdrawal of ground forces from Vietnam, a negotiated settlement saving the Saigon regime, détente with the U.S.S.R., resumption of relations with mainland China, and military support for selected regional powers permitting them to take over as local "policemen" in lieu of direct American involvement. In a period of just four years, 1969–72, the United States abandoned once-unshakable Cold War attitudes toward the Communist nations, while scaling back its own exposure in response to the Sino-Soviet split, imminent Soviet strategic parity, and the economic and psychological constraints on U.S. action stemming from the new American imperative of "no more Vietnams." Nixon believed that his own record as an anti-Communist and tough negotiator would quiet conservative opposition to détente, while liberals would find themselves outflanked on their own peace issue. In both ends and means American foreign policy evinced a new realism in stark contrast to the "pay any price, bear any burden" mentality of the Kennedy-Johnson years. In his inaugural address Nixon spoke instead of an "era of negotiation."

Détente, however, was not meant to replace the abiding postwar American strategy of containment. Rather, it was meant to be a less confrontational method of containing Communist power through diplomatic accords and a flexible system of rewards and punishments by which Washington might moderate Soviet behaviour. Journalists dubbed this tactic "linkage" insofar as the United States would link positive inducements (*e.g.*, arms control, technology transfers, grain sales) to expected Soviet reciprocity in other areas (*e.g.*, restraint in promoting revolutionary movements). Nixon had no illusions that U.S.–Soviet competition would disappear, but he expected that this carrot-and-stick approach would establish rules of the game and recognized spheres of influence. Pulling the Soviets into a network of agreements, and thus giving them a stake in the status quo, would create a stable structure of peace. Finally, expanding economic and cultural ties might even serve to open up Soviet society.

By 1971, Leonid Brezhnev, now established as the new Soviet leader, was ready to welcome American overtures for a variety of reasons. In 1968 relations with the eastern European satellites had flared up again when leaders of the Czech Communist party under Alexander Dubček initiated reforms promoting democratization and free speech. A wave of popular demonstrations added momentum to liberalization during this "Prague Spring" until, on August 20, the U.S.S.R. led neighbouring Warsaw Pact armies in a military invasion of Czechoslovakia. Dubček was ousted and the reforms undone. The ostensible justification for this latest Soviet repression of freedom in its empire came to be known as the Brezhnev Doctrine: "Each of our parties is responsible not only to its working class and its people, but also to the international working class, the world Communist movement." The U.S.S.R. asserted its right to intervene in any Communist state to prevent the success of "counterrevolutionary" elements. Needless to say, the Chinese were fearful that the Brezhnev Doctrine might be applied to them. In 1969 they accused the U.S.S.R. of "social imperialism" and provoked hundreds of armed clashes on the borders of Sinkiang and Manchuria. Soviet forces arrayed against China, already raised from 12 weak divisions in 1961 to 25 full ones, now grew to 55 divisions backed by 120 SS-11 nuclear missiles. In August 1969 a Soviet diplomat had carefully inquired about the likely American reaction to a Soviet nuclear strike against China. In sum, the need to repair the Soviet image in the wake of the Prague Spring, the fear of dangerous relations with Peking and Washington at the same time, as well as the chronic Soviet need for agricultural imports and access to superior Western technology, were all powerful incentives for seeking détente.

The Brezhnev Doctrine

From a longer perspective, however, détente had been the strategy of the U.S.S.R. ever since 1956 under the rubric "peaceful coexistence." Brezhnev repeated Khrushchev's assertion that Soviet nuclear parity took the military leverage from the hands of the bourgeois world, forcing it to accept the legitimate interests of other states, to treat the U.S.S.R. as an equal, and to acquiesce in the success of "progressive" and revolutionary struggle. Détente was thus for the Soviets a natural expression of the new correlation of forces, a means of guiding the weakened Americans through the transition to a new phase of history—and was certainly not meant to preserve the status quo or liberalize the U.S.S.R. One Western proponent of détente described the Soviet conception of it as a way "to make the world safe for historical change," and pointed out the implicit double standard—*i.e.*, that it was admissible for the U.S.S.R. to continue the struggle against the capitalist world during détente, but a contradiction for the Western powers to struggle against Communism. From the Marxist point of view, however, this was merely another reflection of objective reality: Now that nuclear balance was a fact, greater weight accrued to conventional military strength and popular political action, each of which strongly favoured the socialist bloc.

Soviet concept of détente

The contrasting U.S. and Soviet conceptions of détente would eventually scotch the hopes placed in it on both sides. But from 1969 to 1972 those differences were not yet apparent, while the immediate incentives for a relaxation of tensions were irresistible.

Scaling back U.S. commitments. The first indications of a new American sense of limits in foreign policy were in the economic sphere. Since World War II the global market economy had rested on the Bretton Woods monetary system, based on a strong American dollar tied to gold. Beginning in 1958 the United States began to run annual foreign-exchange deficits, resulting partly from the costs of maintaining U.S. forces overseas. For this reason, and because their own exports benefited from an artificially strong dollar, the Europeans and Japanese tolerated the U.S. gold drain and used their growing fund of "Eurodollars" to back loans and commerce. By the mid-1960s de Gaulle began to criticize the United States for exploiting its leadership role to "export its inflation" to foreign holders of dollars. The Johnson administration's Vietnam deficits then added the prospect of internal American inflation. By 1971 the American economic situation warranted emergency measures. Nixon imposed wage and price controls to stem inflation, and Secretary of the Treasury John Connally abruptly suspended the convertibility of dollars to gold. The dollar was allowed to float against undervalued currencies like the deutsche mark and yen, in consequence of which foreign holders of dollars took sharp losses and foreign exporters faced stiffer competition from American goods. New agreements in December 1971 stabilized the dollar at a rate 12 percent below Bretton Woods, but the United States had sorely tried allied loyalty.

The American retreat from an overextended financial position and insistence that its allies share the burden of stabilizing the U.S. balance of payments was the economic analog to the Nixon Doctrine in military affairs. The new president enunciated this doctrine in an impromptu news conference on Guam during his July 1969 trip to welcome home the Apollo 11 astronauts from the Moon. Nixon announced that the United States would no longer send Americans to fight for Asian nations but would confine itself to logistical and economic support: "Asian hands must shape the Asian future." In accord with this effort to shift more of the burden of containment to threatened peoples themselves, Nixon planned to assist regional pro-Western powers like Iran in becoming bulwarks of stability by providing them with sophisticated American weapons.

The Nixon Doctrine

Before the Nixon Doctrine could be credible, however, the President had to extricate the United States from Vietnam. In March 1969 he outlined a policy of Vietnamization, comprising a phased withdrawal of American ground troops and additional material and advisory support to make the ARVN self-sufficient. Nixon also hoped to enlist the Soviets in the cause of peace, but Moscow had less influence over Hanoi than he imagined and could not afford to be seen as appeasing the United States. Nixon then shifted to a subtler approach—long-term pressure on Hanoi combined with better relations with both Communist giants. Late in 1969 secret talks began in Paris between Henry Kissinger, Nixon's adviser for national security, and North Vietnamese Politburo member Le Duc Tho. At the same time, however, Nixon stepped up pressure on the North. When the anti-Communist general Lon Nol overthrew Prince Sihanouk in Cambodia in March 1970, Nixon acceded to the U.S. army's long-standing desire to destroy Communist sanctuaries inside that country. The U.S.-ARVN operation fell short of its promise and provoked protests at home and abroad. Despite public disfavour and congressional attempts to limit such actions, Nixon ordered continued secret American bombing inside Cambodia and also supported an ARVN operation into Laos to cut the Ho Chi Minh Trail.

The opening to China and "Ostpolitik." The linchpin of Nixon's strategy for a settlement in Vietnam was détente with Moscow and Peking. He was known as a firm supporter of the Nationalist regime on Taiwan, but he had softened his stance against mainland China before taking office. In 1969 he moved to signal Peking through the good offices of de Gaulle and Yahya Khan of Pakistan. Direct contacts, conducted through the Chinese embassy in Warsaw, were broken off after the 1970 U.S.-ARVN attacks on Cambodia, but Nixon and Kissinger remained hopeful. The Cultural Revolution ended in a serious power struggle in the Chinese leadership. Army commander Lin Biao opposed relations with the United States but died when his plane crashed in unclear circumstances. Zhou Enlai and Mao (presumably) contemplated the value of an American counterweight to the Soviets, concessions on the status of Taiwan, and technology transfers. The Nixon Doctrine also promised to remove the obnoxious U.S. military presence in Asia.

The Pakistani channel bore fruit in December 1970, when Yahya Khan returned from Peking with an invitation for an American envoy to discuss Taiwan. The following April the Chinese made the surprising public gesture of inviting an American table tennis team to the championship tournament in Peking. This episode of "ping-pong diplomacy" was followed by a secret trip to Peking by Kissinger. Kissinger's talks with Zhou and Mao yielded an American promise to remove U.S. forces from Taiwan in return for Chinese support of a negotiated settlement in Vietnam. The Chinese also agreed to a presidential visit in February 1972. The American people's long-latent fascination with China immediately revived, and Nixon's trip was a sensation.

Ping-pong diplomacy

The Soviets watched with palpable discomfort as Richard Nixon and Mao Zedong embraced and saluted each other's flags, and they quickly raised the premium on improving relations with Washington. Efforts to this end had been frustrated by a series of crises: a buildup of Soviet jets in Egypt and Jordan, the discovery of a Soviet submarine base under construction in Cuba in 1970, and Nixon's escalations of the war in Southeast Asia. But substantial moves toward East–West détente had already been made in Europe. Following de Gaulle's lead, West German foreign minister Willy Brandt, a Socialist and former mayor of West Berlin, had made overtures toward Moscow. After becoming chancellor in 1969 he pursued a thorough *Ostpolitik* (eastern policy), culminating in treaties with the U.S.S.R. (August 1970), renouncing the use of force in their relations, and with Poland (December 1970), recognizing Germany's 1945 losses east of the Oder–Neisse Line. Brandt also recognized the East German government (December 1972) and expanded commercial relations with other eastern European regimes. Both German states were admitted to the UN in 1973. Support for *Ostpolitik* among West Germans reflected the growing belief that German reunification would more likely be achieved through détente, rather than confrontation, with the Soviet bloc.

The United States, Britain, and France seconded Brandt's efforts by concluding a new Four Power accord with the U.S.S.R. on Berlin in September 1971. The Soviets made what they considered a major concession by agreeing to retain their responsibility under the Potsdam Accords for access to West Berlin and achieved in return Western recognition of the status quo in eastern Europe and access to West German technology and credits.

Arms-limitation negotiations. The centrepiece of a bilateral U.S.–Soviet détente, however, had to be the Strategic Arms Limitation Talks (SALT), which began in 1969. After a decade of determined research and deployment the Soviet Union had pulled ahead of the United States in long-range missiles and was catching up in submarine-launched missiles and in long-range bombers. Indeed, it had been American policy since the mid-1960s to permit the Soviets to achieve parity in order to stabilize the regime of mutual deterrence. But stability was threatened from the technological quarter with the development of multiple independently targeted reentry vehicles (MIRVs), by which several warheads, each aimed at a different target, could be carried on one missile, and antiballistic missiles (ABMs), which might allow one side to strike first while shielding itself from retaliation. In the arcane province of strategic theory, therefore, offense (long-range missiles) became defense, and defense (ABM) offense. Johnson had favoured a thin ABM system to protect the United States from a Chinese attack, and in 1969 Nixon won Senate approval of ABM deployment by a single vote. He intended,

SALT, MIRVs, and ABMs

however, to use the program as a bargaining chip. The Soviets had actually deployed a rudimentary ABM system but were anxious to halt the U.S. program before superior American technology left theirs behind. The public SALT talks stalled, but back-channel negotiations between Kissinger and Ambassador Anatoly Dobrynin produced agreement in principle in May 1971 to limit long-range missiles and ABM deployment. The American opening to China made the Soviets increasingly eager for a prompt agreement and summit meeting, while the Americans hoped that Moscow would encourage North Vietnam to be forthcoming in the peace talks.

Since 1968 North Vietnamese negotiators had demanded satisfaction of Premier Pham Van Dong's "four points" of 1965, including cessation of all U.S. military activity in Indochina, termination of foreign military alliances with Saigon, a coalition government in the South including the NLF, and reunification of Vietnam. The United States demanded withdrawal of all foreign troops from the South, including the PAVN. This deadlock, plus Hanoi's anxiety over the possible effects of détente, prompted another North Vietnamese bid for victory on the battlefield. In March 1972 they committed 10 of their 13 divisions to a massive offensive. Nixon responded by ordering the resumption of bombing of the North for the first time since 1969 and the mining of the harbour at Haiphong, North Vietnam's major port. The offensive stalled.

Nixon's retaliation against North Vietnam prompted speculation that the U.S.S.R. would cancel the planned summit meeting, but Soviet desire for détente prevailed. Kissinger visited Moscow in April 1972 to work out details on SALT and draft a charter for détente. Nixon instructed him "to emphasize the need for a single standard; we could not accept the proposition that the Soviet Union had the right to support liberation movements throughout the world while insisting on the Brezhnev Doctrine inside the satellite orbit." The Soviets, however, refused to make explicit concessions and defined détente as a means of preventing the inevitable struggle between "progressive" and "reactionary" forces from escalating into war. The result was a vague statement of 12 "basic principles of mutual relations" committing the two parties to peaceful coexistence and normal relations based on "sovereignty, equality, non-interference in internal affairs, and mutual advantage." Nixon then proceeded to Moscow in May 1972 and signed 10 documents providing for cooperation in economics, science and technology, outer space, medicine, health, and the environment. Most important were the SALT accords: an Interim Agreement limiting ballistic-missile deployment for five years and the ABM Treaty limiting each side to two ABM sites, one protecting the national capital, the other a long-range missile site. The treaty also enjoined the signatories not to interfere with each other's "national technical means of verification," a de facto recognition of each side's space-based reconnaissance satellites.

The SALT accords

The preliminary SALT agreement appeared to be a significant achievement, but there was in some ways less to it than met the eye. The treaty mandated controlled increases, not decreases, in the Soviet arsenal, while failing to ban development of cruise missiles, space-based weapons, or the MIRVing of existing launchers by the United States or the U.S.S.R. Thus the superpowers sacrificed the right to defend their attack missiles with ABMs while failing to ensure the stability of mutual deterrence. In sum, the limitation of one sort of nuclear launcher (long-range missiles) did not preclude a continuing arms race in other sorts of launchers or in technological upgrades. To be sure, the mere fact of a U.S.–Soviet agreement seemed of psychological value, but only if both sides were genuinely seeking to reduce arsenals and not simply to maneuver diplomatically for a future advantage. Hence the practical value, or danger, of SALT would only be revealed by superpower behaviour in years to come.

End of the Vietnam War. The American achievement of détente with both Moscow and Peking and the failure of North Vietnam's spring 1972 offensive moved both protagonists in that conflict to bargain as well. In October the secret talks in Paris between Kissinger and Le Duc Tho finally produced an agreement on a cease-fire, the release of prisoners of war, evacuation of remaining U.S. forces within 60 days, and political negotiations among all Vietnamese parties. But South Vietnam's president, Nguyen Van Thieu, balked: The plan might indeed allow the Americans to claim "peace with honour" and go home, but it would leave Thieu to deal with the Communists while 100,000 PAVN troops remained in his country. When North Vietnam sought to prevent any last-minute changes by releasing in public the Paris terms, Kissinger was obliged to announce on October 26 that "peace is at hand." After his landslide reelection a week later—a victory aided by the prospect of peace—Nixon determined to force compliance with the terms on both Vietnamese states. Nixon ordered 11 days of intensive bombing over Hanoi itself (December 18–28) while sending Thieu an ultimatum threatening a separate peace and cessation of U.S. aid if Saigon did not accept the peace terms. The United States was castigated worldwide for the "Christmas bombing," but when talks resumed in January, Hanoi and Saigon quickly came to terms. A Vietnam cease-fire went into effect on Jan. 27, 1973, and the last American soldiers departed on March 29.

The Christmas bombing

Vietnam had been America's longest and most divisive war, and public and congressional opinion flatly opposed any resumption of the agony. The 1973 accords, therefore, were a fig leaf hiding the fact that the United States had just lost its first war despite an estimated expenditure of $155,000,000,000, 7,800,000 tons of bombs (more than all countries dropped in all of World War II), and 58,718 American lives. Estimates of Vietnamese dead (North and South) totaled 950,000. In its proportional impact on Vietnamese society, the Vietnam War, 1965–75, was the fourth most severe in the world since 1816.

The end of U.S. involvement in Southeast Asia also brought to a close 15 years of astounding change in world politics featuring the arrival of the space and missile age, the climax of decolonization, the assertions of Maoist China and Gaullist France, the shattering of the myth (fostered by Washington and Moscow alike) of a monolithic Communist world, and the relative decline of American power. In 1969, the very moment when astronauts were setting foot on the Moon to fulfill Kennedy's pledge to prove American superiority, Nixon and Kissinger were struggling to adjust to the new realities and manage a limited American retreat. They succeeded brilliantly in establishing a triangular relationship with Moscow and Peking and appeared to have replaced Cold War with détente. Likewise, they appeared to have escaped from Vietnam and implemented the Nixon Doctrine. But new crises and reversals were in the offing that would prove the American decline had not yet been arrested. Given these reversals, détente might be judged as much an exercise in American presumption as the Vietnam War. For the U.S.S.R. could not be expected to cease its quest for real values in world competition just because the United States was prepared to acknowledge it as a military equal. Rather, that very equality opened up new opportunities for Soviet expansion with which the United States would be less able to cope than before. Khrushchev's boast about the new correlation of forces in the world may have brought the Soviets a series of embarrassments from 1957 to 1962, but a decade later it seemed perversely justified.

Dependence and disintegration in the global village

Events after the 1960s seemed to suggest that the world was entering an era both of complex interdependence among states and of disintegration of the normative values and institutions by which international behaviour had to a reliable extent been made predictable. Perhaps this was not an anomaly, for if modern weapons, communications satellites, and global finance and commerce really had created a "global village" in which the security and well-being of all peoples were interdependent, then by the same token the opportunities had never been greater for ethnic, religious, ideological, or economic differences to spark resentment and conflict among the villagers.

In a world so seemingly out of control, it is perhaps a wonder that politics have not been even more violent and anarchic. But the liberal dreams of progress nurtured in the 19th century had surely proved false. For the spread of modern technology and economic growth around the world had not necessarily increased the number of societies based on human rights and the rule of law, nor had multilateral institutions like the United Nations or financial and economic interdependence created a higher unity and common purpose among nations, except within the durable and democratic North Atlantic alliance.

Instead, the world after the 1960s saw a proliferation of violence at every level except war among developed nations, a world financial structure under tremendous strain, the worst economic downturn since the 1930s and reduced growth rates thereafter, recurrent fears of an energy crisis, the depletion of resources and concurrent global pollution, famine and genocidal dictators in parts of Africa and Asia, the rise of an aggressive religious fundamentalism in the Muslim world, and widespread political terrorism in the Middle East and Europe. The superpowers never ceased to compete in the realms of strategic weapons and influence in the Third World and thus failed to sustain their brief experiment with détente. As President Jimmy Carter's national security adviser, Zbigniew Brzezinski, concluded: "The factors that make for international instability are gaining the historical upper hand over the forces that work for more organized cooperation. The unavoidable conclusion of any detached analysis of global trends is that social turmoil, political unrest, economic crisis, and international friction are likely to become more widespread during the remainder of this century."

Global concerns after the 1960s

THE DECLINE OF DÉTENTE

General Secretary Brezhnev and President Nixon were understandably optimistic in the wake of the endorsement by the 24th Party Congress of the Soviet peace program in 1971 and Nixon's landslide reelection in 1972. Both expected their new relationship to mature over the course of Nixon's second term. But détente had fragile foundations in foreign as well as domestic policy. The Soviets viewed it as a form of mere peaceful coexistence in which revolutionary forces could be expected to take advantage of the new American restraint, while the U.S. administration implicitly sold détente as a means of restraining Communist activity around the world. American conservatives were bound to lose faith in détente with each new incident of Soviet assertiveness, while liberals remained hostile to Nixon himself, his *Realpolitik,* and his predilection for the use of force. Between 1973 and 1976 Soviet advances in the Third World, the destruction of Nixon's presidency in the Watergate scandal, and congressional actions to limit the foreign policy prerogatives of the White House undermined the domestic foundations of détente. After 1977 the U.S.S.R. seemed to take advantage of the Carter administration's vacillations in Third World conflicts and in arms-control talks until the Democrats themselves reluctantly announced the demise of détente following the Soviet invasion of Afghanistan in 1979.

The distraction of Watergate. Analysts with a sufficiently historical point of view tended to see in the Watergate affair and Nixon's 1974 resignation the culmination of a 30-year trend by which war and Cold War had greatly expanded, and ultimately corrupted, executive power. Liberals who, in Eisenhower's time, had called for strong presidential leadership now bemoaned "the imperial presidency." With what were widely understood to be the lessons of Vietnam fresh in the nation's mind, and a majority in Congress and the press hostile to the sitting president, the moment arrived for a legislative counterattack on the executive. This interpretation is borne out by the subsequent congressional acts designed to limit executive freedom in foreign policy. The War Powers Act of 1973 restrained the president's ability to commit U.S. forces overseas. The Stevenson and Jackson–Vanik amendments imposed conditions (regarding Soviet policy on Jewish emigration) on administration plans to expand trade with the U.S.S.R. In 1974–75 Congress prevented the president from involving the United States in a crisis on Cyprus or aiding anti-Communist forces in Angola and passed the Arms Export Control Act, removing the president's discretion in supplying arms overseas. New financial controls limited the president's ability to conclude executive agreements with foreign powers, a form of which 6,300 had been signed between 1946 and 1974 as compared with only 411 treaties requiring the Senate's advice and consent. Finally, revelations of past CIA covert operations, including schemes to assassinate Fidel Castro, inspired complicated congressional oversight procedures for U.S. intelligence agencies. These assaults on executive prerogative were meant to prevent future Vietnams, prevent unelected presidential aides from engaging in secret diplomacy, and restore to Congress an "appropriate" role in foreign policy. Critics of the limitations held that no great power could conduct a coherent or effective foreign policy under such a combination of openness and restrictions, especially in a world populated increasingly by totalitarian regimes, guerrilla movements, and terrorists.

The War Powers Act

The Nixon–Brezhnev summits of 1973–74 produced only minor follow-ons in the area of arms control—the uncontroversial Agreement on the Prevention of Nuclear War and an agreement to reduce the number of ABM sites from the two permitted in 1972 to one. Gerald Ford, president from August 1974, and Kissinger, who remained as secretary of state, attempted to restore the momentum of détente through a new SALT agreement regulating the dangerous race in MIRVed missiles, which SALT I had not prevented. The United States proposed strict equality in nuclear delivery systems and total "throw-weight." This meant that the United States would be allowed to MIRV more of its missiles to offset the greater size of Soviet missiles. But since the United States had no plans for a unilateral buildup in any case, the Soviets had no incentive to make such a concession. Instead, Ford and Brezhnev signed an Interim Agreement at Vladivostok in November 1974 that limited each side to 2,400 delivery vehicles, of which 1,320 could be MIRVed. While the Soviets claimed this was a concession, since they declined to count the 90 British and French missiles aimed at them, the Soviets' giant SS-18s, able to deliver up to 10 MIRVs, ensured the U.S.S.R. an advantage in ICBM warheads. The repeated failure to restrain the growth of Soviet offensive systems soon sparked fears that the United States might become vulnerable to preemptive attack.

The Vladivostok Agreement

Meanwhile, the mid-1970s brought to a logical conclusion the process of détente in Europe. Nixon and Kissinger, aware that the United States had seemed to ignore its European allies during the 10 years of Vietnam, declared 1973 "the year of Europe" and hoped to forestall NATO governments from bargaining with Moscow on their own. But Watergate and the Arab–Israeli War of that year turned this initiative into a public-relations failure. Instead the United States was obliged to follow the European lead in the ongoing Conference on Security and Cooperation in Europe and negotiations toward a "mutual and balanced force reduction" treaty covering NATO and Warsaw Pact forces in central Europe. The climax of the security talks was the Helsinki summit of 35 nations in the summer of 1975 and an agglomeration of proposals divided into three "baskets." (A fourth basket dealt with the question of a follow-up conference.) In Basket I the signatories accepted the inviolability of Europe's existing borders and the principle of noninterference in the internal affairs of other states—thereby recognizing formally the Soviet gains in World War II and the Soviet-bloc states. Basket II promoted exchanges in science, technology, and commerce, expanding Soviet access to Western technology and opening the Soviet market to western European industry. Basket III, the apparent Soviet concession, aimed at expanding cultural and humanitarian cooperation among all states on the basis of respect for human rights. Not surprisingly, Western opinion of the Helsinki Accords, and of détente in general, came to rest heavily on whether the U.S.S.R. would voluntarily comply with Basket III. American leaders of both parties considered Helsinki misguided and empty, especially after Moscow stepped up persecution of dissidents and jailed those of their citizens engaged in a "Helsinki watch" on

The Helsinki Accords

Soviet compliance. In sum, Helsinki (and U.S. demands on behalf of Soviet Jews) pointed up another contradiction in détente, this time between American insistence on Soviet liberalization and Soviet insistence on noninterference in domestic politics of other states.

During final negotiations at Helsinki, events in Southeast Asia compounded the American sense of humiliation and growing discontent with détente. The North Vietnamese had never viewed the 1973 peace accords as anything other than an interlude permitting the final withdrawal of American forces. In the year following they built up their strength in South Vietnam to more than 150,000 regulars armed with Soviet tanks, artillery, and anti-aircraft weapons. The ARVN was poorly trained, suffered from low morale after the Americans were gone, and faced an enemy able to attack at times and places of its own choosing. The American withdrawal also removed at a blow some 300,000 jobs from the local economy, and President Thieu made matters worse by trying to establish one-party bureaucratic rule without the charisma or prestige to sustain it. By October 1974 the Politburo in Hanoi concluded that the Saigon regime was ripe for collapse. Large-scale probes of ARVN defenses in January 1975 confirmed their optimism. By the end of the month 12 provinces and 8,000,000 people had fallen to the Communists. On April 10, unable to obtain congressional approval of $422,000,000 in further military aid, President Ford declared that the Vietnam War was over "as far as America is concerned." The final North Vietnamese offensive reached Saigon on April 30, 1975, as the last remaining Americans fled to helicopters atop the U.S. embassy. Hanoi triumphantly reunified Vietnam politically in July 1976 and confined thousands of South Vietnamese to "re-education camps," while thousands of "boat people" risked death in the South China Sea to escape reprisals and Communism.

The fall of Saigon

The end in Cambodia had already occurred. The Communist Khmer Rouge cut off the capital, Phnom Penh, in January 1975. When the U.S. Congress denied further aid to Cambodia, Lon Nol fled, and in mid-April the Khmer Rouge took control. Its leader, Pol Pot, was a French-educated disciple of Maoist "total revolution" to whom everything traditional was anathema. The Khmer Rouge reign of terror became one of the worst holocausts of the 20th century. All urban dwellers, including hospital patients, were forced into the countryside in order to build a new society of rural communes. Sexual intercourse was forbidden and the family abolished. More than 100,000 Cambodians, including all "bourgeois," or educated people, were killed outright, 400,000 succumbed in the death marches, and 1,200,000 people (a fifth of the Cambodian nation) perished in all. The Khmer Rouge, however, were not allied with Hanoi, and in 1979 PAVN forces invaded Cambodia to oust the Khmer Rouge and install a puppet regime. This completed the conquest of Indochina by North Vietnam, for Laos, too, became Communist after the fall of Saigon. Thus the domino theory was at last put to the test and to a large extent borne out.

Events in Africa as well seemed to bear out the Soviet expectation that "progressive forces" would gain ground rapidly during the new era of superpower parity. Angola and Mozambique, coastal states facing the oil-tanker routes around the Cape of Good Hope, were finally slated to achieve independence from Portugal following a leftist military coup in Lisbon in April 1974. Three indigenous groups, each linked to tribal factions, vied for predominance in Angola. The MPLA (Popular Movement for the Liberation of Angola) of Agostinho Neto was Marxist and received aid from the U.S.S.R. and Cuba. The FNLA (National Front for the Liberation of Angola) in northern Angola was backed by Zaire's Mobutu and initially by a token contribution from the CIA. In the south the UNITA (National Union for the Total Independence of Angola) of Jonas Savimbi had ties to China but came to rely increasingly on white South Africa. In the Alvor agreement of January 1975 all three agreed to form a coalition, but civil war resumed in July. By the end of the year the MPLA had been reinforced by 10,000 Cuban soldiers airlifted to Luanda by the U.S.S.R. But in the United States the imperative of "no more Vietnams" and congressional ire over CIA covert operations frustrated Ford's desire to help non-Communist Angolans. Neto accordingly proclaimed a People's Republic of Angola in November 1975 and signed a Treaty of Friendship with the U.S.S.R. the following October. The rebel factions, however, remained in control of much of the country, and Cuban troop levels eventually reached 19,000. A Marxist government also assumed power in Mozambique.

The war in Angola

American uncertainty. In winning the presidential election of 1976, Jimmy Carter capitalized on the American people's disgust with Vietnam and Watergate by promising little more than an open and honest administration. Though intelligent and earnest, he lacked the experience and acumen necessary to provide strong leadership in foreign policy. This was especially unfortunate since his major advisers had sharply divergent views on the proper American posture toward the Soviet Union.

Carter's inaugural address showed how much he diverged from the *Realpolitik* of Nixon and Kissinger. Such a sentiment as "Because we are free we can never be indifferent to the fate of freedom elsewhere," recalled Kennedy's 1961 call to arms. But Carter made clear that his emphasis on human rights applied at least as much to authoritarian governments friendly to the United States as to Communist states, and that idealism was in fact, as he put it on another occasion, the most "practical and realistic approach" to foreign policy. He hoped to divert American energies away from preoccupation with relations with the U.S.S.R. toward global problems such as energy, population control, hunger, curbing of arms sales, and nuclear proliferation. Carter's first initiative in the perilous field of arms control was an embarrassing failure. Rejecting his own secretary of state's advice to take a gradual approach, he startled the Soviets with a deep-cut proposal for immediate elimination of as much as 25 percent of the U.S. and Soviet strategic missiles and a freeze on new long-range missile deployment. Brezhnev rejected it out of hand, and Foreign Minister Andrey Gromyko called this attempt to scrap the Vladivostok formula a "cheap and shady maneuver."

Carter was to gain one stunning success during his term, a peace treaty between Egypt and Israel (see below), but he was unable to stem the growth of Soviet influence in Africa. Somalia, on the strategic Horn of Africa astride the Red Sea and Indian Ocean shipping lanes, had been friendly to Moscow since 1969. In September 1974 a pro-Marxist military junta overthrew the government of neighbouring Ethiopia, had Emperor Haile Selassie confined in his palace (where he was later suffocated in his bed), and invited Soviet and Cuban advisers into the country. The Somalis then took advantage of the turmoil—perversely, from Moscow's point of view—to reassert old claims to the Ogaden region of Ethiopia and invade, while Eritrean rebels also took up arms against Addis Ababa. The Soviets and Cubans stepped up support for Ethiopia, while Castro vainly urged all parties to form a "Marxist federation." Carter at first cut off aid to Ethiopia on the ground of human-rights abuses and promised weapons for the Somalis. By August he realized the arms would only be used in the Ogaden campaign and reversed himself, making the United States appear ignorant and indecisive. Somalia broke with the U.S.S.R. anyway, but 17,000 Cuban troops and $1,000,000,000 in Soviet aid allowed Ethiopia to clear the Ogaden of invaders and to suppress the Eritrean revolt in 1978. Ethiopia signed its own treaty of friendship and cooperation with the U.S.S.R. in November. The failure of the Carter administration either to consult with the Soviets or to resist Soviet-Cuban military intervention set a bad precedent and weakened both détente and U.S. prestige in the Third World.

Revolt in Ethiopia

The events in the Horn of Africa, which Brzezinski interpreted as part of a Soviet strategy to outflank the oil-rich Persian Gulf so vital to Western economies, encouraged the United States to seek help in balancing Soviet power in the world. The obvious means of doing so was to complete the rapprochement with China begun under Nixon. Some advisers opposed "playing the China card" for fear that the Soviets would retaliate by calling off the continuing SALT negotiations, but Brzezinski persuaded the President that

closer ties between the United States and China would oblige the U.S.S.R. to court the United States, as had occurred in 1972. Brzezinski went to Peking in May 1978 to initiate discussions leading toward full diplomatic recognition. His cause was aided by important changes in the Chinese leadership. Zhou Enlai and Mao Zedong had died in 1976. Hua Guofeng won the initial power struggle and ordered the arrest and trial of the radical Gang of Four led by Mao's wife, Jiang Qing. Both superpowers hoped that the suppression of radicals in favour of pragmatists in the Chinese government might portend better relations with Peking. But the rehabilitation of the formerly condemned "capitalist roader" Deng Xiaoping led to a resumption of Soviet–Chinese border clashes, and the clear shift of Vietnam into the Soviet camp strengthened Washington's hand in Peking. Hua and Carter announced in December 1978 that full diplomatic relations would be established on Jan. 1, 1979. The United States downgraded its representation on Taiwan and renounced its 1954 mutual defense treaty with the Nationalist Chinese.

The "China card"

The spectre of a possible Sino-American alliance may have alarmed the Soviets (Brezhnev warned Carter not to sell arms to China) but was never a real possibility. The Chinese remained Communist and distrustful of the United States. They made clear that China was no card to be played at will by one or the other of the superpowers. Nor could China's underdeveloped economy sustain a large conventional war or the projection of force overseas (which the United States would not want in any case), while in nuclear systems China was as weak vis-à-vis the Soviet Union as the Soviet Union had been vis-à-vis the United States in the 1950s. Ties to the United States might provide China with high technology, but the United States was no more willing to place nuclear or missile systems in Chinese hands than Khrushchev had been. To be sure, the United States had an interest in preventing a Sino-Soviet rapprochement (an estimated 11 percent of the Soviet military effort was devoted to the Chinese front), but any pause given the U.S.S.R. by Sino-American cooperation was probably more useful to China than to the United States. Indeed, Peking was quite capable of playing its U.S. card to carry out adventures of its own.

After their 1975 victory the North Vietnamese showed a natural strategic preference for the distant U.S.S.R. and fell out with their historic enemy, neighbouring China. In quick succession Vietnam expelled Chinese merchants, opened Cam Ranh Bay to the Soviet Navy, and signed a treaty of friendship with Moscow. Vietnamese troops had also invaded Kampuchea (Cambodia) to oust the pro-Peking Khmer Rouge. Soon after Deng Xiaoping's celebrated visit to the United States, Peking announced its intention to punish the Vietnamese, and, in February 1979, its forces invaded Vietnam in strength. The Carter administration felt obliged to favour China (especially given residual American hostility to North Vietnam) and supported Peking's offer to evacuate Vietnam only when Vietnam evacuated Cambodia. The Soviets reacted with threats against China, but Chinese forces performed abysmally even against Vietnam's frontier militia, and after three weeks of hard fighting, in which Vietnam claimed to have inflicted 45,000 casualties, the Chinese withdrew. The results for United States policy were all negative: Chinese military prestige was shattered, Cambodia remained in the Soviet-Vietnamese camp, and the tactic of playing the China card was rendered ridiculous.

The SALT II treaty

To the chagrin of Peking, the Sino-Vietnamese War failed to forestall a planned U.S.–Soviet summit meeting and the signing of a second arms agreement, SALT II. After Carter's first deep-cut proposal, negotiations had resumed on the basis of the Vladivostok Agreement and had finally produced a draft treaty. The summit was held in Vienna in June 1979, and Carter returned to seek congressional approval for SALT II as well as most-favoured-nation trade status for both the U.S.S.R. and China. The treaty inspired widespread suspicion in the United States Senate on its own merits. The modest limits on nuclear forces and allowances for upgrading existing missiles did not seem sufficient to prevent the Soviets' superior long-range missile forces from threatening the survival of United States land-based missiles. The American will to upgrade its own deterrent, meanwhile, seemed to be sapped by the SALT process itself. Confusion reigned over how the MX missile might be deployed so as to survive a Soviet first strike, and Carter canceled programs to deploy the B-1 strategic bomber and an antitank neutron bomb designed for Europe. There was widespread doubt whether Soviet compliance with SALT II could be adequately monitored. The treaty foundered as well on growing American impatience with Communist expansion in the Third World.

Soviet invasion of Afghanistan

Any chance of Senate ratification of SALT II disappeared on Dec. 25, 1979, when the U.S.S.R. launched an invasion of Afghanistan to prop up a friendly regime. Even after a decade of détente the American public still thought viscerally in terms of containment, and this latest and most brazen Soviet advance pushed the President over the fence. "This action of the Soviets," said Carter, "has made a more dramatic change in my own opinion of what the Soviets' ultimate goals are than anything they've done." Calling the Afghan invasion "a clear threat to peace," Carter ordered an embargo on sales of grain and high-technology equipment to the U.S.S.R., canceled U.S. participation in the 1980 Moscow Olympic Games, reinstated registration for the draft, withdrew the SALT II treaty from the Senate, and proclaimed the Carter Doctrine, pledging the United States to the defense of the Persian Gulf. It was clear to all that détente was dead.

Postmortem. Was détente a failure because the Soviets refused to play by the rules, because the United States was unwilling to accord the U.S.S.R. genuine equality, or because détente was never really tried at all? Or did the differing U.S. and Soviet conceptions of détente ensure that, sooner or later, American patience would wear thin? The last explanation is, in foreshortened perspective, at least, the most convincing. From the Soviet point of view the United States had been a hegemonic power from 1945 to 1972, secure in its nuclear dominance and free to undertake military and political intervention around the world. But the correlation of forces had gradually shifted to the point where the U.S.S.R. could rightly claim global equality and respect for "peaceful coexistence." Under détente, therefore, the United States was obliged to recognize Soviet interests in all regions of the world and to understand that the U.S.S.R. was now as free as the United States to defend those interests with diplomacy and arms. And those interests included, above all, fraternal aid for "progressive" movements in the Third World. Détente certainly could never mean the freezing of the status quo or the trends of history as understood in Marxist theory. Instead, in the Soviet view, the United States continued to resent Soviet equality in armaments, to shut the U.S.S.R. out of regional diplomacy (as in the Middle East), to interfere in Soviet domestic policy, to support counterrevolutionary movements, and, in violation of the spirit of détente, to attempt to organize the encirclement of the U.S.S.R. in league with NATO and China.

Divergent views of détente

From the American perspective, Soviet policy from 1945 to 1972 was characterized by a Marxist-Leninist drive to export revolution and achieve world dominion by dividing and bullying the West and exploiting the struggles of Third World nations. But given the growing maturity of the U.S.S.R. itself, the split in world Communism, and the realization that the Western world was not about to collapse (either from "the contradictions of capitalism" or Soviet subversion), had made Cold War obsolete. Under détente, therefore, the U.S.S.R. was obliged to accept the responsibilities as well as the benefits of membership in the comity of civilized states, to reduce its exorbitant military spending and subversive activity, and to cease trying to turn the domestic problems of other countries to unilateral benefit. Instead, in the American view, the U.S.S.R. continued to exploit Western restraint, to build up its nuclear and conventional forces far beyond the needs of deterrence, and to exploit Communist proxy forces to take over developing nations.

Each view had a basis in reality, and, given the differing assumptions of the two governments, each was persuasive. But the burden of compromise or dissolution of the relationship fell inevitably on the democratic, status quo

power. In time American opinion would cease to tolerate Soviet advances made under the guise of détente. The notion of détente was flawed from the start in two crucial points. First, with the exception of preventing nuclear war, the United States and the U.S.S.R. still shared no major interests in the world; and second, the specific agreements on respect for spheres of influence included Europe and isolated regions elsewhere but not the bulk of the Third World. The Americans inevitably viewed any Soviet assertiveness in such undefined regions as evidence of the same old Soviet drive for world domination, while the Soviets inevitably viewed any American protestations as evidence of the same old American strategy of containment. Within a decade, the hopes raised by Nixon and Brezhnev stood exposed as illusory.

THE "ARC OF CRISIS"

Among the manifestations of the diffusion of political power in the world after 1957 was the rise of regional powers and conflicts with only distant or secondary connections to the rivalries of the Cold War blocs, of multilateral political and economic pressure groups, and of revolutionary, terrorist, or religious movements operating across national boundaries ("nonstate actors"). The politics of the Middle East after 1972 comprised all three and so frustrated attempts by the industrial states to control events in the region that by 1978 Brzezinski was describing the old southern tier of states reaching beneath the U.S.S.R. from Egypt to Pakistan as the "arc of crisis."

Palestinian terrorism and diplomacy. The sweeping Israeli victory in the Six-Day War of 1967 had forced every Arab state to rethink its own foreign policy and the extent of its commitment to the cause of Arab unity. Egypt, having lost the Sinai, faced Israelis entrenched in the Bar-Lev line directly across the Suez Canal. Jordan, having lost the West Bank, faced Israeli troops directly across the Jordan River. Syria, having lost the Golan Heights, faced Israeli forces within easy striking distance of Damascus itself. The notion of united Arab armies sweeping the Jews into the sea had clearly proved to be romantic, while political unity among the Arabs suffered from the abiding division between nationalist and socialist states like Egypt, Syria, and Iraq, and traditional Arab monarchies like Saudi Arabia and Jordan.

The Palestine Liberation Organization (PLO), organized in 1964 to represent some 2,000,000 refugees from the Palestine mandate who were scattered around the Arab world and from 1969 led by Yasir Arafat, was also divided between old families of notables, whose authority dated back to Ottoman times, and young middle-class or fedayeen factions anxious to exert pressure on Israel and the West through terrorism. The latter included the Popular Front for the Liberation of Palestine (PFLP), formed three months after the 1967 war. Over the next year the PFLP hijacked 14 foreign airliners, culminating in its spectacular destruction of four planes at once in Jordan. In 1970–71 the moderate King Hussein of Jordan lost patience with the autonomous PLO formations in his territory and expelled them, provoking a sharp military exchange with Syria. But the PLO moved its central offices to Lebanon, whence terrorists could cross the frontier to commit atrocities against civilians inside Israel. The PFLP and other Palestinian groups also linked up with extreme leftist and rightist (because anti-Semitic) conspiracies in Italy, Austria, and Germany to form a terrorist network that left no European or Mediterranean state free from the fear of random violence. In September 1972 terrorists from an organization calling itself Black September took nine Israeli athletes hostage at the Munich Olympic Games; all of the hostages and five terrorists died in the ensuing gun battle with police.

The PLO and the PFLP

The terrorist network benefited mightily from the financial support, training, or refuge provided by established pro-Soviet states like Cuba, East Germany, Bulgaria, Algeria, Syria, Yemen (Aden), and especially Libya. In 1969 the Libyan monarchy was overthrown in a military coup led by Colonel Muammar al-Qaddafi, a fanatical adherent of Nasser's pan-Arabism. Following Nasser's death in 1970 and the development of rich oil deposits in Libya, Qaddafi styled himself as the new leader and financier of the radical Arab cause. In imitation of Mao, he issued a little *Green Book* describing his "new gospel. . . . One of its words can destroy the world." The ideology was a mixture of Third World-ism, Socialism, and Muslim fundamentalism, and it called forth a "heroic politics." In the eyes of the West, the rhetoric masked a crazed cruelty, and even in Arab eyes it seemed at best antiquated in the wake of the 1967 war.

Another new feature of Middle Eastern politics was the assertiveness of the Organization of Petroleum Exporting Countries (OPEC), composed of oil-producing countries in the Persian Gulf and Arabian Peninsula as well as Libya, Nigeria, and Venezuela. The members of this producers' cartel accounted for a large percentage of the world's oil reserves and wielded tremendous potential power over the Europeans and Japanese, who relied on imports for more than 80 percent of their energy needs. In the past, oil prices had been kept artificially low by the Western oil companies through bilateral agreements with producer states. By 1970, however, most host governments had taken over ownership of the production facilities, and they saw in a drastic rise of oil prices a means of accumulating capital for development and purchases of arms, as well as a way to pressure the Western states into respecting their grievances against Israel.

The OPEC cartel

The most populous frontline (*i.e.*, bordering Israel) Arab state, but one without oil revenues, was Egypt. Since 1955 Egypt had undergone a demographic explosion. Population was growing at a rate of 1,000,000 per year, and 35,000,000 people were crowded into the Nile Valley and Delta. The numbers and youth of the Egyptians (over half were under 25 in 1980), and the country's economic weakness, meant that frustrated and unemployed youth posed the constant threat of political instability. Certainly Egypt could no longer afford an endless crusade against Israel. These considerations dominated the thinking of Nasser's successor as president, Anwar el-Sādāt. But neither could he abandon Nasser's legacy, especially with the Sinai under Israeli occupation, without losing his legitimacy at home. Accordingly, Sādāt laid a risky and courageous plan to extricate his country from its foreign and domestic stalemates. Husbanding the arms provided by the U.S.S.R. after 1967, he abruptly expelled his 20,000 Soviet advisers in July 1972 and opened a secret channel to Washington, hinting that Egypt and the United States together could eliminate Soviet involvement in the Middle East. For only the Americans, he reasoned, might influence the Israelis to return the occupied regions. Then, on Oct. 6, 1973, during the Jewish holiday of Yom Kippur, he launched the fourth Arab–Israeli war.

The Yom Kippur War

The Egyptian army moved across the Suez Canal in force and engaged the Bar-Lev line. For the first time it made substantial progress and inflicted a level of casualties especially damaging for the outmanned Israelis. Syrian forces also stormed the Golan Heights. The United States and the Soviet Union reacted with subtle attempts to fine-tune the outcome by alternately withholding or providing arms to the belligerents and by urging or discouraging a UN cease-fire. Nixon denied Israel an airlift of arms until October 13, preventing Israel from launching a prompt counterattack and thereby signaling Sādāt of American sympathy. But once assured of United States aid, the Israelis struck on both fronts, regained the Golan Heights, and crossed the Suez Canal. Henry Kissinger, alarmed that the Israeli victory might be so complete as to hinder a lasting settlement, quickly agreed to call, with the Soviet Union, for a UN cease-fire. It broke down at once, and Israeli forces encircled a 20,000-man Egyptian army corps. Brezhnev curtly warned Nixon of possible Soviet military intervention, which the United States moved to deter, perhaps recklessly, with a worldwide alert of its military forces. Finally, Kissinger threatened a cutoff of arms deliveries unless Israel halted its offensive, and peace was restored.

The 1973 war saved Egyptian honour and solidified Sādāt's prestige to the point where he could afford to be conciliatory. The United States emerged as the "honest broker" between Egypt and Israel. As Kissinger put it,

"The Arabs can get guns from the Russians, but they can get their territory back only from us." Kissinger's "shuttle diplomacy" between Tel Aviv and Cairo secured an Israeli withdrawal beyond the Suez in January 1974, the reopening of the canal, the insertion of a UN force between the antagonists, and, in September 1975, an Israeli retreat from the crucial Mitla and Gidi passes in the Sinai. The United States flooded both countries with economic and military aid, and Sādāt repudiated Nasser's socialism in favour of policies stimulating domestic private enterprise.

The limited rapprochement that emerged from the 1973 war was purchased at great economic cost, for the Arab OPEC nations, led by Saudi Arabia, seized the opportunity to enact an embargo of oil exports to all nations aiding Israel, which lasted for five months. More telling still was the price revolution that preceded and followed. OPEC had already engineered a doubling of the posted price of oil to $3.07 per barrel by the eve of the war. In January 1974 it nearly quadrupled the price again, to $11.56 per barrel. The importance of this sudden rise cannot be exaggerated. The resulting shortages and exorbitant costs accelerated the growing inflation in the Western world, exposed the energy-dependency of the industrial nations, created a vast balance-of-payments deficit in many industrial nations, wiped out the hard-won economic progress of many developing nations, and placed massive sums of petrodollars in the hands of a few underpopulated Middle Eastern states. The political upshot was that the United States and Europe would have to pay close attention to the desires of those Arab states in foreign policy as long as OPEC unity survived.

The oil embargo

In November 1977, Sādāt shocked the Arab world by announcing his willingness to go to Jerusalem personally to seek peace. When his talks with the new Israeli prime minister, Menachem Begin, broke down, President Carter invited them both to Camp David in September 1978. During 11 days of intensive discussion, Carter succeeded in bringing the rivals together. The Camp David Accords provided for complete Israeli evacuation of the Sinai, gradual progress toward self-rule for West Bank Palestinians over a five-year period, and a peace treaty signed by Begin and Sādāt at the White House in March 1979. This historic settlement dismayed other Arab states and split the PLO asunder, the so-called rejectionists refusing to recognize the settlement. Qaddafi purchased huge amounts of Soviet arms and expanded Libya's training and supply of terrorists. In December 1979, 300 Muslim fundamentalists trained in Yemen seized the holiest of all Islāmic shrines in Mecca. Sādāt himself was assassinated by Arab extremists in 1981.

Camp David Accords

The Iranian revolution. Carter's success in Middle Eastern diplomacy was likewise undercut by the collapse of the strongest and staunchest American ally in the Muslim world, the Shah of Iran. Since the monarchy had been restored by a CIA-aided coup in 1953, Reza Shah Pahlavi had used Iran's oil revenues to finance rapid modernization of his country and the purchase of American arms. Nixon had chosen Iran to be a United States surrogate in the vital Persian Gulf, and as late as 1977 Carter praised the Shah for making Iran "an island of stability." Clearly, American intelligence services failed to detect the widespread Iranian resentment of modernization (meaning, in this context, materialism, emancipation of women, and secularization), middle-class opposition to the autocracy, and the rising tide of Shīʿite fundamentalism that were undermining the Shah's legitimacy. Fundamentalist movements and conflicts between Sunnite and Shīʿite Muslims have arisen periodically in the course of Islāmic history, but the outbreaks of the late 20th century were especially notable in light of the Western assumption that less developed countries would naturally secularize their politics and culture as they modernized their society and economy. Instead, rapidly developing Iran succumbed to a religious revolution led by Ayatollah Ruhollah Khomeini. By November 1978 the beleaguered Shah saw his options reduced to democratization, military repression, or abdication. Despite the importance of Iran for U.S. interests, including the presence there of critical electronic listening posts used to monitor missile tests inside the U.S.S.R., Carter was unable to choose between personal loyalty toward an old ally and the moral argument on behalf of reform or abdication. In January 1979 the Shah left Iran; the next month, when he requested asylum in the United States, Carter refused lest he give offense to the new Iranian regime. But the gesture did not help the United States. An interim government in Tehrān quickly gave way to a theocracy under Khomeini, who denounced the United States as a "great Satan" and approved the seizure in November 1979 of the American embassy in Tehrān and the holding of 52 hostages there. The hostage drama dragged on for nearly 15 months, and most Americans were infuriated by the unfathomable Khomeini and frustrated by Carter's apparent ineffectiveness.

The hostage crisis

Carter reacted to the crisis by adopting Brzezinski's formula that the Middle East and South Asia constituted an arc of crisis susceptible to Soviet adventurism. In his State of the Union Address of January 1980 he enunciated the Carter Doctrine, declaring that any attempt by an outside force to gain control of the Persian Gulf would be viewed as an attack on the vital interests of the United States, and he pledged to form a Rapid Deployment Force to defend the region. Whether the U.S. military was truly capable of sustained combat in that remote region was doubtful. When diplomacy failed to free the hostages in Tehrān, Carter resorted in April 1980 to a military rescue mission, hoping to repeat the success of a brilliant Israeli commando raid that had freed 103 airline passengers at Entebbe, Uganda, in 1976. But the operation was a humiliating failure. Only in January 1981, after the overwhelming defeat of his reelection bid, did Carter achieve the release of the hostages.

The Soviets in Afghanistan. Brzezinski's fears that the U.S.S.R. would take advantage of the arc of crisis seemed justified when the Soviet army invaded Afghanistan in 1979. It is likely, however, that the Soviets were responding to a crisis of their own, rather than trying to exploit another's. Remote and rugged Afghanistan had been an object of imperialist intrigue throughout the 19th and 20th centuries because of its vulnerable location between the Russian and British Indian empires. After 1955, with India and Pakistan independent, the Afghan government of Mohammad Daud Khan forged economic and military ties to the U.S.S.R. The monarchy was overthrown by Daud Khan in 1973 and was succeeded by a one-party state. The small Afghan Communist party, meanwhile, broke into factions, while a fundamentalist Muslim group began an armed insurrection in 1975. Daud Khan worked to lessen Afghanistan's dependence on Soviet and U.S. aid, and he reportedly had a heated disagreement with Brezhnev himself during a visit to Moscow in April 1977. Leftists in the Afghan officer corps, perhaps fearing a blow against themselves, murdered Daud Khan in April 1978 and pledged to pursue friendly relations with the U.S.S.R. Thus Afghanistan, under the rule of Nur Mohammad Taraki, was virtually in the Soviet camp. But when Taraki objected to a purge of the Afghan Cabinet, the leader of a rival faction, Hafizullah Amin, had him arrested and killed. These intramural Communist quarrels both embarrassed the Soviets and threatened to destabilize the Afghan regime in the face of growing Muslim resistance. In the fall of 1979 the Soviets built up their military strength across the border and hinted to American diplomats that they might feel obliged to intervene. On Dec. 25, 1979, the Soviet army began its occupation, and two days later a coup d'état led to the murder of Amin and the installation of Babrak Karmal, a creature of the KGB, who had been brought into the country by Soviet paratroops.

The Soviets would probably have preferred to work through a pliant native regime rather than invade Afghanistan, but Amin's behaviour and Moscow's unwillingness to risk a domestic overthrow of a Communist regime forced their hand. The invasion, therefore, appeared to be an application of the Brezhnev Doctrine and was all the more pressing given that the Central Asian provinces of the Soviet Union were also vulnerable to the rise of Islāmic fundamentalism. The United States was tardy in responding to the 1978 coup despite Carter's concern over the arc of crisis and the murder of the United

States ambassador in Kabul in February 1979. But the Soviet invasion aroused American suspicions of a Russian grand strategy aimed at seizing a warm-water port on the Indian Ocean and the oil of the Persian Gulf. Over the course of the next decade the puppet Afghan regime lost all authority with the people, Afghan soldiers defected in large numbers, and the Muslim and largely tribal resistance, armed with U.S. and Chinese weapons, held out in the mountains against more than 100,000 Soviet troops and Soviet terror bombing of their villages. More than 2,000,000 Afghans became refugees in Pakistan and Iran. Soon Western observers began to speak of Afghanistan as the Soviets' Vietnam.

The Shī'ite revolution in Iran, meanwhile, provoked and tempted neighbouring Iraq into starting yet another war in the arc of crisis. The secular, leftist Iraqi regime was nervous about the impact Iranian events might have on its own large Shī'ite population, while its Kurdish minority had resorted to terrorism in pursuit of its goal of a Kurdish state to be carved out of Turkey, Iraq, and Iran. Finally, the Iraqi government of Saddam at-Takrītī Hussein hoped to use the opportunity of Iran's apparent near-anarchy to seize the long-disputed Shatt al-Arab waterway at the mouth of the Tigris-Euphrates river system.

The Iranian–Iraqi War

Bolstered by arms purchased with oil revenues, Hussein unilaterally abrogated a 1975 accord on the waterway and launched a full-scale invasion of Iran in September 1980. After initial victories the Iraqis were surprisingly thrown back, and a war of attrition commenced. The Iraqis employed poison gas and were building a nuclear reactor capable of producing weapons-grade plutonium until the Israeli air force destroyed the reactor in a surprise raid in June 1981. The Iranians relied on human-wave assaults by revolutionary youths assured of a place in paradise for dying in battle. Both sides employed imported planes and missiles to attack each other's oil facilities, tanker ships, and, occasionally, cities. Attacks then spread to neutral shipping as well, and oil production in the entire Gulf region was placed in jeopardy. Neither superpower had any direct interest in the war except a common opposition to any overthrow of the local balance of power, but the Soviets tended to benefit from a prolongation of the conflict. In 1987 the United States sharply increased its presence in the Gulf by permitting Kuwaiti oil tankers to fly the U.S. flag and by deploying a naval task force to protect them in passage through the Gulf. Compared to the situation of the 1950s, when John Foster Dulles' CENTO seemed to ensure a ring of stable, pro-Western governments in the South Asian region, that of the 1980s was almost totally unpredictable.

RHETORICAL COLD WAR REVIVED

Rudderless diplomacy. After the collapse in 1979 of any pretense of détente, relations between the superpowers drifted. The 1980 U.S. election brought to the White House a new president, Ronald Reagan, seemingly the most conservative president on domestic policy since the 1920s and a Cold Warrior reminiscent of the 1950s. He bemoaned an arms-control process that, he said, always favoured the U.S.S.R. and sapped the will of the Western alliance, and a détente that only duped gullible Americans into acquiescing in Soviet gains. Reagan sounded like John Foster Dulles when he denounced the U.S.S.R. as "an evil empire" and held it accountable for the unrest in the world, and he echoed John Kennedy in calling for America to "stand tall in the world" again. But such rhetoric did not necessarily imply confrontational policies. Instead, Reagan's deeds recalled the prudent containment policies of the Eisenhower era. In part this may have been because the Reagan administration had to acknowledge the constraints on American overseas involvement that had arisen over the previous 20 years, including Soviet military parity or superiority, declining American economic leverage over its own allies, the "lessons of Vietnam" as they pertained to Third World conflicts, American moral doubts about supporting anti-Communist authoritarian governments, and congressional limits placed on executive action.

With a change in State Department leadership in June 1982, the pragmatic attitude of the administration became even more pronounced. American foreign policy in the 1980s, therefore, reflected the country's widespread disaffection with the results of détente but failed to establish a clear alternative. The one means the administration found for standing against growing Soviet power and influence was to extend aid and comfort to irregular forces resisting pro-Soviet governments in the Third World. Such "freedom fighters," as Reagan termed them, in Afghanistan, Angola, and Nicaragua, seemed to offer hope that the United States could pressure, and perhaps overthrow, totalitarian regimes without risking involvement in more Vietnams. Hence this Reagan Doctrine was a natural corollary of the Nixon Doctrine.

The Reagan Doctrine

Soviet foreign policy in the early 1980s also drifted, if only because of the advanced age of Brezhnev and the frequent changes in leadership after his death in November 1982. Early in the decade a recurrence of serious unrest in eastern Europe, this time in Poland, also kept the attention of the Kremlin close to home. During détente the Polish government expanded an ambitious development plan financed largely by western European credits. But economic performance foundered, foreign debt mounted to $28,000,000,000, and the state imposed successive price hikes on staples. By 1979–80 a popular protest movement grew up around the officially unsanctioned Solidarity labour union and its charismatic leader, Lech Wałęsa. The strong Catholic roots of Polish popular nationalism were evident in the movement, especially in light of the accession in 1978 of Karol Cardinal Wojtyła as Pope John Paul II, the first non-Italian pope in 456 years, who in 1981 survived an assassination plot probably hatched in Bulgaria, a Soviet satellite. As unrest mounted in Poland, NATO countries warned against a Soviet military intervention, holding in reserve the threat of declaring Warsaw in default on its debts. In December 1981, General Wojciech Jaruzelski declared martial law, sparing Poland a Soviet invasion at the price of military rule and the suppression of Solidarity. The United States responded by suspending Poland's most-favoured-nation trade status and blocking further loans from the International Monetary Fund. Reagan held the Soviet Union responsible for martial law; his attempts to extend the sanctions to an embargo on high-technology exports to the U.S.S.R., however, angered western Europeans, who feared losing access to eastern European markets and who were in the process of completing a huge pipeline from Siberia that would make western Europe dependent on the U.S.S.R. for 25 percent of its natural gas. In both the debt and pipeline issues, it seemed that the web of interdependence woven during détente served to constrain Western countries more than it did the U.S.S.R.

Martial law in Poland

Brezhnev's successor as General Secretary of the Communist Party, the former KGB chief Yury Andropov, declared that there was no alternative to détente as the Soviets understood it. He denounced Reagan's "militaristic course" as a new bid for U.S. hegemony. But it was Reagan's image of the U.S.S.R. that seemed confirmed when a Soviet jet fighter plane shot down a civilian South Korean airliner in Soviet air space in September 1983, killing 269 people. Some in the West supported the Soviet claim that the plane was on a spy mission but produced no persuasive evidence to that effect. Andropov's demise after a year and a half elevated Konstantin Chernenko, another member of the older generation of the Politburo who would himself survive only until March 1985. Given these frequent changes in leadership and the drain on Soviet resources caused by the ongoing war in Afghanistan, the Kremlin was even less able than the White House to mount new initiatives in foreign policy until late in the 1980s.

Renewal of arms control. The most serious consequence of the collapse of détente and the failure of the SALT II Treaty (judged by Reagan as "seriously flawed") appeared to be an acceleration of the arms race between the superpowers. Liberal critics feared that Reagan would unleash a new arms race; his supporters asserted that the Soviets had never stopped racing even during the era of SALT. But Reagan waffled on arms policy because of stiff domestic and European opposition to the abandonment

of arms control. Programs to upgrade the three elements of strategic deterrence were approved only after being cut back; yet they drew complaints from the Soviet Union that the highly accurate MX missile, the new Poseidon nuclear submarines, and air-launched cruise missiles for the B-52 force were first-strike weapons. A serious NATO worry stemmed from Soviet deployment of the new SS-20 theatre ballistic missile in Europe. In 1979 the Carter administration had acceded to the request by NATO governments that the United States introduce 572 Pershing II and cruise missiles into Europe to balance the 900 SS-20s. But the European antinuclear movement, now officially patronized by the British Labour Party, the Green Party in West Germany, and the Dutch and Belgian Social Democrats, forced Reagan to link Pershing deployment with intermediate nuclear forces (INF) talks with the U.S.S.R. Reagan tried to seize the moral high ground with his "zero-option" proposal for complete elimination of all such missiles from Europe and a call for new Strategic Arms Reduction Talks (START) to negotiate real reductions in the superpower arsenals. But the Soviets refused to scrap any of their long-range missiles or to trade existing SS-20s for Pershings yet to be deployed.

'Star Wars"

In March 1983, Reagan announced a major new research program to develop antiballistic missile defenses based in outer space. This Strategic Defense Initiative (SDI, dubbed "Star Wars" by opponents) was inspired by the emergence of new laser and particle-beam technology that seemed to have the potential to devise an accurate, instantaneous, and nonnuclear means of shooting down long-range missiles in their boost phase, before their multiple reentry vehicles had a chance to separate. The President thus challenged his country to exploit its technological edge to counter the threat of Soviet offensive missiles and perhaps liberate the world from fear of a nuclear holocaust. Scientific and political critics ridiculed SDI as naive (because it would not work or could be easily countered), expensive beyond reckoning, counter-productive (since it implied repudiation of the 1972 ABM Treaty), and dangerous (since the Soviets might stage a preemptive attack to prevent its deployment). But the alarmed Soviets weakened the case of American critics by launching their own propaganda campaign against SDI, implying that they took seriously its prospects for success. Evidence also mounted that the U.S.S.R. had been engaged in similar research since the mid-1970s. A $26,000,000,000, five-year American program was approved, although Congress limited future funding and arms-control advocates pressured the President to use SDI as a bargaining chip in the START talks. The Soviets broke off the INF and START talks at the end of 1983 but resumed talks two years later, apparently with hopes of stalling SDI research.

Regional crises. U.S.–Soviet competition in the Third World also continued through the 1980s, although Reagan's assertion that the Soviets were to blame for all unrest was less convincing than the less feverish, if cynical, view that the Soviets simply sought to benefit from indigenous sources of unrest. The campaign of the Communist-led African National Congress against apartheid in southern Africa, for instance, might serve Soviet strategic aims, but the black rebellion against white rule was surely indigenous. White-supremacist governments in southern Africa might argue, correctly, that the standard of living and everyday security of blacks were better in their countries than in most black-ruled African states, but the fact remained that African blacks, like all human beings, preferred to be ruled by their own tyrant rather than one of some other nationality or race. What was more, the respect shown by African governments for international boundaries began to break down after 1970. Spain's departure from the Spanish Sahara was the signal for a guerrilla struggle among Moroccan and Mauritanian claimants and the Polisario movement backed by Algeria. The Somali invasion of the Ogaden, Libyan intrusions into Chad and The Sudan, and Uganda's 1978 invasion of Tanzania exemplified a new volatility. Uganda had fallen under a brutal regime headed by Idi Amin, whom most African leaders tolerated (even electing him president of the Organization of African Unity) until Julius Nyerere spoke out, following Uganda's invasion of his country, about the African tendency to reserve condemnation for white regimes only.

Libya and Uganda

The black revolt against white rule in southern Africa was a timely consequence of the decolonization of Angola and Mozambique and of the Lancaster House accord under which white Southern Rhodesians accepted majority rule, resulting in 1980 in the full independence of Zimbabwe under Robert Mugabe, who in 1984 declared his intention to create a one-party Marxist state. South Africa tried to deflect global disgust with its apartheid system by setting up autonomous "tribal homelands" for blacks, but no other government recognized them. United States diplomacy sought quietly to promote a comprehensive settlement of southern Africa's problems by pressuring Pretoria to release South West Africa/Namibia and gradually dismantle apartheid in return for a Cuban evacuation of Angola and Mozambique. This policy of "constructive engagement," by which the U.S. State Department hoped to retain leverage over Pretoria, came under criticism every time a new black riot or act of white repression occurred. Critics demanded economic divestment from, and stringent sanctions against, South Africa, but supporters of the policy argued that sanctions would inflict disproportionate economic harm on South African blacks, drive the whites to desperation, and encourage violence that would strengthen the hand of Communist factions. Congressional pressure finally forced the administration to compromise on a package of sanctions in 1986, and U.S. firms began to pull out of South Africa.

The Middle East remained crisis-prone despite the Egyptian–Israeli peace. In 1978 an Arab summit in Baghdad pledged $400,000,000 to the PLO over the next 10 years. A comprehensive Middle East peace was stymied by the unwillingness of rejectionist Arab states to negotiate without the PLO, and the U.S.-Israeli refusal to negotiate with the PLO. In June 1982 the Begin government determined to put an end to terrorist raids by forcibly clearing out PLO strongholds inside Lebanon. In fact the Israeli army advanced all the way to Beirut in a bitter campaign that entrenched Syrian occupation of the strategic al-Biqāʿ Valley and intensified what already amounted to a Lebanese civil war among Palestinians, Muslims of various sects and allegiances, and Christian militiamen. The United States sent marines to Beirut to facilitate the evacuation of the PLO, while it tried without success to piece together a coalition Lebanese government and induce the Israelis and Syrians to withdraw. But in October 1983, terrorists blew up the marine barracks, killing more than 200 Americans. The Middle East peace process begun by Kissinger and continued by Carter seemed to have unraveled by the late 1980s. Western governments tried to coordinate policies on terrorism, including a firm refusal to bargain with kidnappers, but concern for the lives of hostages and fear of future retaliation insidiously weakened their resolve. In October 1985, however, the Israeli air force dispatched planes to bomb the PLO headquarters in Tunis. When Libyan-supported terrorists planted bombs in airports in Rome and Vienna in December 1985 and in a discotheque in Berlin in April 1986, Reagan ordered U.S. jets to attack terrorist training camps and air-defense sites in Libya. The raid was applauded by the American public, and terrorist incidents did seem to decline in number over the following year. Qaddafi suffered another reverse in the spring of 1987 when French-supported Chadian troops drove Libyan invaders from their country.

Israeli invasion of Lebanon

In the Persian Gulf the Reagan administration held publicly aloof from the war between Iraq and Iran. But intelligence that Shīʿite terrorists were behind the kidnapping of Americans in Beirut prompted the administration secretly to supply arms to Iran in return for help, never forthcoming, in securing the release of hostages. There was also a notion that such a deal might forge links to moderate Iranians in hopes of better relations in the event of the aged Khomeini's death. While the motives were humanitarian and strategic, this action directly contradicted the policy of shunning negotiations with terrorists that the United States had been urging on its allies. When the operation was exposed, the Reagan administration lost credibility with Congress and foreign governments alike.

Iran arms deal

LATIN-AMERICAN UPHEAVALS

Marxism and the Cuban role. After a tour of Latin America in 1950, American diplomat George Kennan wrote a memo despairing that the region would ever achieve a modest degree of economic dynamism, social mobility, or liberal politics. The culture itself was, in his view, inhospitable to middle-class values. As late as 1945 almost all the Latin-American republics were governed by landowning oligarchies allied with the church and army, while illiterate, apolitical masses produced the mineral and agricultural goods to be exported in exchange for manufactures from Europe and North America. But to Fidel Castro and other radical intellectuals, a stagnant Latin America without strong middle classes was precisely suited for a Marxist, not a democratic, revolution. Before 1958 the United States—the "colossus to the north"—had used its influence to quell revolutionary disturbances, whether out of fear of Communism, to preserve economic interests, or to shelter strategic assets such as the Panama Canal. But after Castro's triumph of 1959, the United States undertook to improve its own image through the Alliance for Progress and to distance itself from especially obnoxious authoritarian regimes. But Latin-American development programs largely failed to keep pace with population growth and inflation, and frequently they were brought to naught by overly ambitious schemes or official corruption. By the 1980s the wealthiest and largest states like Brazil and Mexico faced a crushing burden of foreign debt. Neo-Marxist economists of the 1960s and 1970s argued that even the more enlightened policies of the Kennedy and Johnson administrations kept Latin America in a condition of stifling dependence on American capital and markets and world commodity prices. Some endorsed the demands of the Third World bloc in the UN for a "New World economic order," involving a massive shift of resources from the rich countries to the poor or the "empowerment" of the developing countries to control the terms of trade along the lines of OPEC. Others advocated social revolution to transform Latin states from within. But the example of Cuba's slide into the status of a Communist satellite fully dependent on the U.S.S.R. revived the fear and suspicion with which Americans habitually regarded Third World revolutions.

Even after the Bay of Pigs invasion and the 1962 missile crisis, Cuba retained a certain autonomy in foreign policy, while the Soviets exhibited caution about employing their Cuban clients. Castro preferred to place himself among the ranks of Third World revolutionaries like Nasser, Nyerere, or Ghana's Kwame Nkrumah rather than follow slavishly the Moscow party line. He also elevated himself to leadership of the nonaligned nations. When relations between Havana and Moscow cooled temporarily in 1967–68, Brezhnev applied pressure, holding back on oil shipments and delaying a new trade agreement. Castro tried to resist the pressure by exhorting and mobilizing his countrymen to produce a record 10,000,000-ton sugar harvest in 1970. When the effort failed, Castro moved Cuba fully into the Soviet camp. The U.S.S.R. agreed to purchase 3,000,000 to 4,000,000 tons of sugar per year at four times the world price, provide cheap oil, and otherwise subsidize the island's economy at a rate of some $3,000,000,000 per year; thenceforward, 60 percent of Cuba's trade was with countries in the Soviet-bloc. Brezhnev himself visited Cuba in 1974 and declared the country "a strong constituent part of the world system of socialism." Castro, in turn, voiced the Soviet line on world issues, played host to Latin-American Communist party conventions, used the forum of the nonaligned nations movement to promote his distinctly aligned program, and made tens of thousands of Cuban troops available to support pro-Soviet regimes in Africa.

Cuba becomes a satellite

Soviet domination of Cuba, however, may have harmed their chances elsewhere in Latin America, since it alerted other leftists to the dangers of seeking Soviet support. Moreover, the Soviets simply could not afford such massive aid to other clients. This limitation appeared to be crucial even when Communists had a chance of prevailing in one of the largest, most-developed South American states, Chile. The Communist party there was a charter member of the 1921 Comintern and had strong ties to the Chilean labour movement. The party was outlawed until 1956, whereupon it formed an electoral popular front with the Socialists, and it narrowly missed electing Socialist Salvador Allende Gosseus to the presidency in 1964. The Christian Democratic opponent, Eduardo Frei Montalva, had warned that an Allende victory would make Chile "another Cuba." From 1964 to 1970, when Cuba was plying an autonomous course, the Chilean Castroites staged violent strikes, bombings, and bank robberies in defiance of the regular Communist party directed from Moscow. The latter's strategy was subtler. Hinting that it might support the Christian Democratic candidate rather than their rival leftists, the Communist party provoked the extreme right to run its own candidate in protest, thus splitting the conservative vote. The Nixon administration tried clumsily to influence the nominating process or foment a military coup, but Allende won an electoral victory in 1970. Once in office, he seized U.S. property and forged close ties to Cuba at the very time when Castro was being reined in by Brezhnev. The U.S.S.R., however, held back from extending large-scale aid, even after a fall in copper prices, radical union activity, and Allende's radical policies plunged Chile into economic chaos. In September 1973, General Augusto Pinochet Ugarte and the army overthrew Allende and established an authoritarian state. The Soviets and Allende sympathizers in North and South America depicted the denouement in Chile as the work of fascists in league with U.S. imperialists.

The United States' poor image in Latin America was of special concern to Jimmy Carter because of his dedication to the promotion of human rights. During his first year in office Carter sought to counter the traditional notion of "Yankee imperialism" by meeting the demands of the Panamanian leader, General Omar Torrijos Herrera, for a transfer of sovereignty over the Panama Canal. The U.S. Senate ratified the treaty (which called for a staged transfer, to be completed in 1999) by a bare majority, but most Americans opposed transfer of the canal. Conservatives also held Carter's human rights concerns to be naive, since the linking of U.S. government loans, for instance, to a regime's performance on human rights damaged American relations with otherwise friendly states while exercising no influence on human rights practices in Communist states. Supporters of Carter retorted that the pattern of U.S. support for cruel oligarchies on the excuse of anti-Communism was what drove oppressed Latins toward Communism in the first place.

The Panama Canal Treaty

The first hemispheric explosion in the 1980s, however, occurred in the southern cone of South America when the Argentine military ruler, Lieutenant General Leopoldo Galtieri—apparently to distract attention from the abuses of his dictatorship and an ailing economy at home—broke off talks concerning sovereignty over the Falkland Islands (Islas Malvinas) and invaded the remote archipelago in April 1982. The conservative British government of Margaret Thatcher was taken by surprise but began at once to mobilize supplies, ships, and men to reconquer the islands some 8,000 miles from home. The United States was torn between loyalty to its NATO ally (and political friend of President Reagan) and the fear of antagonizing South Americans by siding with the "imperialists." But when U.S. diplomacy failed to resolve the dispute, the United States supplied Britain with intelligence data from American reconnaissance satellites. The Royal Navy and ground forces began operations in May, and the last Argentine defenders surrendered on June 14. In the wake of the defeat, the military junta in Buenos Aires gave way to democratization.

The Falklands War

Nicaragua and El Salvador. Problems in Central America, however, commanded the attention of the United States throughout the 1980s. In Nicaragua the broadly based Sandinista revolutionary movement challenged the oppressive regime of Anastasio Somoza Debayle, whose family had ruled the country since the 1930s. In accordance with its human rights policies, the Carter administration cut off aid to Somoza, permitting the Sandinistas to take power in 1979. They appeared to Americans as democratic patriots and received large sums of United

States aid. But a radical faction soon took control of the revolution, and moderates either departed or were forced out of the government in Managua. The Sandinistas then socialized the economy, suppressed freedom of the press and religion, and established close ties to Cuba and other Soviet-bloc countries. By the time Reagan took office, neighbouring El Salvador had also succumbed to violence among leftist insurgents, authoritarian landowners supporting right-wing "death squads," and a struggling reformist government. Reagan vigorously affirmed a last-minute decision by Carter to grant military aid to the Salvadoran government. Nicaragua and Cuba were identified as the sources of the insurgency. Americans became increasingly confused by evidence of atrocities on all sides and were again torn between their desire to promote human rights and their determination to halt the spread of Communism. Opponents of U.S. involvement warned of another Vietnam in Central America, while supporters warned of another Cuba.

Nicaragua, meanwhile, built up one of the largest armies in the world in proportion to population, expanded its port facilities, and received heavy shipments of arms from the U.S.S.R. The CIA used this military buildup to justify the secret mining of Nicaraguan harbours in February 1984, which was, when revealed, universally condemned. The CIA also secretly organized and supplied a force of up to 15,000 anti-Sandinista "freedom fighters," known as the Contras, across the border in Honduras and Costa Rica, while U.S. armed forces conducted joint maneuvers with those states along the Nicaraguan border. The ostensible purpose of such exercises was to interdict the suspected flow of arms from Nicaragua to the Salvadoran rebels. In fact, American policy aimed at provoking a popular revolt in hopes of overthrowing the Sandinistas altogether.

Cuban and Soviet influence with leftist governments on the Caribbean islands of Jamaica, Trinidad, and Grenada also appeared to be on the increase, a trend that the Reagan administration tried to counter with its 1982 Caribbean Basin Initiative, an Alliance for Progress confined to the islands. Grenada, a tiny island that won independence from Britain in 1974, had initially come under the control of Sir Eric Gairy, whose policies and conduct verged on the bizarre. In March 1979, Gairy was overthrown by the leftist New Jewel Movement led by the charismatic Maurice Bishop. Over the next several years the regime socialized the country, signed mutual-assistance agreements with Soviet-bloc states, and hastened construction of a large airstrip that the United States feared would be used by Soviet aircraft. But the evident incompetence of the New Jewel leadership prompted a split in 1982 between Bishop's supporters and hard-line Leninists. In October 1983 the revolution came apart when Bishop was arrested and, when protest demonstrations broke out, was shot. The Organization of East Caribbean States invited American intervention, and U.S. forces, together with small contingents from neighbouring islands, landed on Grenada to restore order and protect American medical students. Free elections returned a moderate government to Grenada in 1984, but the self-destruction and overthrow of the New Jewel Movement, while a setback for Castroism in the region, also lent credence to Nicaragua's often and loudly voiced fear of an American invasion.

The invasion of Grenada

The U.S. public emphatically supported the Grenadian intervention but was split almost evenly on the question of support for the Nicaraguan Contras. While the Reagan Doctrine of supporting indigenous rebels, such as Jonas Savimbi's UNITA in Angola or the *mujāhidīn* in Afghanistan, appeared to be a low-risk means of countering Soviet influence, Americans remained nervous about the possibility of deeper United States involvement. Congress reflected this public ambivalence by first approving funds for the Contras, then restricting the ability of federal agencies to raise or spend funds for the Contras, then reversing itself again. In 1986 investigations of the secret U.S. arms sales to Iran revealed that National Security Council officials had kept supplies flowing to the Contras while the congressional restrictions were in effect by soliciting funds from private contributors and friendly Arab states and by diverting the profits from the Iranian arms sales. In 1987 Congress launched elaborate investigations into this "Iran-Contra Affair" that virtually paralyzed U.S. foreign policy in two regions of the world for more than a year.

THE GORBACHEV ERA

In the early 1980s the Reagan administration's rebuilding of U.S. defenses, restoration of steady economic growth, and self-confident assertion of American principles seemed to have regained the initiative in superpower relations for the United States. The Soviet Union, by contrast, suffered a prolonged leadership crisis and chronically poor economic performance. In 1985 the momentum shifted when the comparatively young and reform-minded Mikhail Gorbachev was elevated to the post of Communist Party general secretary. Unlike previous Soviet leaders, Gorbachev frankly admitted the damage done to the Soviet economy by bureaucratic rigidity, corruption, secrecy, and the lack of market incentives. Though his power was far from consolidated, Gorbachev launched an aggressive campaign on behalf of *perestroika* (economic restructuring) and *glasnost* (publicity, or openness), and his frankness and skill in dealing with the Western press persuaded many Europeans and Americans that real change had become possible in the Soviet Union.

Gorbachev also made surprising forays in foreign policy, including a concerted push for new arms control accords and negotiations with the United States, Pakistan, and the *mujāhidīn* in Afghanistan. In October 1986 Gorbachev met Reagan at an impromptu summit at Reykjavík and proposed dramatic cuts in long-range missile forces. He withdrew the offer when Reagan refused to renounce his SDI program. The following year Gorbachev caught the United States off balance again by suddenly agreeing to an earlier U.S. zero option proposal on intermediate-range nuclear forces (INF) in Europe. This time, negotiations at Geneva eventuated in an INF treaty signed at a Washington summit meeting in 1987. The U.S. Senate ratified the accord on the strength of its having eliminated an entire category of nuclear weapons and having allowed, for the first time, extensive on-site inspection in the Soviet bloc. But critics feared that stripping Europe of nuclear weapons would only enhance the value of the Warsaw Pact's superiority in conventional forces and require the United States to respond to an attack on western Europe either by using its strategic nuclear missiles or by abandoning Europe altogether. NATO governments cheered the arms pact but were reminded of why they had come to rely on nuclear defenses in the first place. Reagan and Gorbachev met again in Moscow in June 1988 to discuss an even bolder proposal: the reduction of superpower strategic nuclear weapons by 50 percent.

Gorbachev also carried through the Afghan negotiations to a May 1988 agreement providing for the withdrawal of Soviet troops and a cut-off of foreign military aid to the warring Afghan parties, leaving open the question of whether the U.S.S.R. would permit its client government in Kabul to fall. Meanwhile, the Reagan administration, weakened by the Iran-Contra scandal, also trimmed its sails with regard to Third World competition. The corrupt Filipino dictator Ferdinand Marcos, a long-standing ally of the United States, lost his grip on power in 1986 when crowds backed by leading elements in the church, press, labour unions, and part of the army rose up to demand his resignation. Rather than try to prop up Marcos, Reagan sent a personal envoy to Manila to engineer his departure and a transfer of power to Corazon Aquino, the wife of a popular opposition leader who had been murdered. In Central America, however, Reagan appeared to lose control of events. The U.S. Congress refused to renew military aid to the Contra rebels in Nicaragua, while Costa Rican president Oscar Arias preempted Reagan by negotiating a comprehensive Central American peace plan with all neighbouring countries. The success of the plan depended in large part on whether the Nicaraguan government carried out its promises regarding freedom of the press and religion and free elections.

As the 1980s ended it appeared that the renewed Cold War was again giving way to détente. Western leaders were left to meditate on several fundamental questions:

Was Gorbachev sincere in his effort to liberalize the Soviet bloc and scale back its foreign involvement? If so, would he survive his opposition at home? If he were both sincere and successful, would a more dynamic, efficient Soviet Union be good or bad for the West in the long run?

FORCES THAT SHAPE WORLD POLITICS

World politics since 1900 is a chronicle of the impact on war and diplomacy of the explosion of technological and organizational power. But it is also the story of the transformation of a Eurocentric great-power system dominated by states with similar Christian cultures into a global congeries of sovereignties, great and small, whose cultures are not merely dissimilar but often seem mutually incomprehensible. What began to disappear in 1914, and was largely gone by 1970, was a common structure of values and norms of behaviour among governments that shape world events. The transition was accompanied by unprecedented spasms of violence and sudden shifts in the fortunes of states. The 30 years' civil war among the European powers (1914–45) turned the old kaleidoscopic balance of six or seven great powers into an ellipse focused on the two superpowers. The 40 years afterward (1945–85) brought a rediffusion of power to second- and third-rank states and revealed new limits in the ability of states wielding traditional military or economic power (or even nuclear power) to control events beyond their borders even as they learned to mobilize their resources ever more effectively at home.

Failure of shared values

Hence all myths of progress, or even the simple belief in a linear logic to the course of history, went into remission. The 19th-century liberal myth of history as the progress of freedom broke on the experience of World War I, the Fascist and Communist episodes, and the failure of democracy and human rights to find many havens in the decolonized world. Freedom, in the Euro-American constitutional sense, has not fared well in the 20th century. But the 20th-century Marxist-Leninist myth that history is the product of inexorable laws tending toward proletarian socialism has broken on the failure of the U.S.S.R. and other Communist states to outperform freer societies even by material standards, to say nothing of qualitative standards of human fulfillment. But to go to the other extreme—to assert that the present age is one of irreversible decay and the future one of apocalypse—would be just as presumptuous. It cannot be said that nuclear war, or the triumph of a given form of society, or world government, or the perfection of freedom through technology, or the extirpation of freedom through biologic and psychological conditioning, is somehow inevitable. Whether or not human beings enjoy free will, they cannot know, much less make, their own future.

What at least are the parameters of world events in the near future? What trends will future historians deem most important? Those following, because of the range of possibilities to which they give rise, are likely candidates.

Demography and food. On the most fundamental level the 20th century has been the age in which the very ability of the Earth's resources to continue to support the human race was called into question. World population nearly doubled from 1,300,000,000 in 1900 to 2,500,000,000 in 1950, and doubled again to 5,000,000,000 by 1987. Estimates were that the 6,000,000,000 mark would be reached by 2000, the 10,000,000,000 mark by 2050. By the 1960s these numbers revived the fears of global famine attended by widespread violence that were first raised by Thomas Malthus early in the 19th century. But economic history suggests that the cause of Europe's "vital revolution" of the 18th to 20th centuries—the Industrial Revolution—was also the cure. Nations appear to pass through a demographic cycle in which falling death rates due to advances in medicine, public health, urbanization, and industrial production all stimulate rapid expansion of population until a level of per capita affluence is reached at which societies cease to reproduce themselves in great numbers and come to prefer the personal freedom and quality of life inherent in smaller families. Indeed, by the 1980s the birth rate of Caucasian populations of North America and Europe had declined to the point of zero population growth or, in the case of West Germany, Sweden, and others, below it. Similar drops in birth rate have occurred in Japan and the prosperous states of the Asian rim (Korea, Hong Kong, Singapore). By the 1980s the population explosion was occurring exclusively in the underdeveloped regions of South Asia, Africa, and Latin America and among racial minorities in more developed countries. The result is that the percentage of mankind residing in the industrial world, and of Caucasians in particular, has dropped steadily. The potential political consequences of this trend range from the upheavals attending massive famine and violence in the underdeveloped world to subtler changes in "First World" political cultures caused by the increase of Third World immigrants, guest workers, and refugees. In places where the industrial and underdeveloped countries directly abut, as along the U.S.–Mexico border or in Soviet Central Asia, tensions have already become evident.

The Malthusian question

While each country presents a unique combination of climate, arable land, fertility, cultural traditions, and available technology, one general reason the global population explosion threatens famine is the dismal performance of collectivized or planned agriculture. The United States, Canada, Australia, and the European Community produce record surpluses almost at will, and populous India, although prone to climatic disasters, is self-sufficient in food, while agriculture in Soviet and Third World socialist countries shows regular annual shortfalls. Unfortunately, landowning or bureaucratic elites in those countries have a vested interest in inefficient forms of land tillage. In many African states tribal hatreds and local wars only magnify the effects of drought, collectivist policies, and primitive technology. The prospect of famine or disease causing catastrophic mortality is a growing possibility, especially in Africa, where the demographic cycle is still in its first stage. Elsewhere, the prospects for the poor and populous nations of the Third World depend on how soon they can reach the stage in the demographic cycle when birthrates decrease. It could occur benignly, as a concomitant of higher per capita incomes, or forcibly, through state-enforced birth control or infanticide. It is likely, however, that by natural or artificial check the rate of increase in global population will be arrested over the course of the next century.

Failure of collective agriculture

Technology and technocracy. A second explosion characteristic of the 20th century is technological. It, too, began early in the Industrial Revolution of the 18th century. But unlike population in the advanced countries, the pace of technological advance did not ultimately slow down but continues to produce successive industrial revolutions including, most recently, those based on atomic power, the jet engine, electronics, computers, spaceflight, lasers, and bioengineering. But the political consequences of each new increment in human power over the material world are unfathomable. Have nuclear weapons rendered obsolete the accumulated lessons of strategy, diplomacy, and war? Or does human conflict continue according to its ancient patterns despite the risk of nuclear holocaust? Do computers magnify the ability of governments to control their populations, perhaps in totalitarian fashion, or will they decentralize the control of information and liberate the individual? To the extent that knowledge is power, the quest for new knowledge and technique is ipso facto a political activity. And the trend appears to be one toward ever greater government stimulation and regulation of new technology for the purposes of military and economic security.

The phenomenon of planned research and development, the "invention of invention," began in the late 19th century and received a powerful impetus from the World Wars and Cold War. The Soviets were the first to institute large-scale state mobilization of research in peacetime, and the scientific and engineering achievements of World War II impressed many in the market-oriented West with the social utility of state-supported research and development. Cold War competition, the high cost of research facilities in such fields as spaceflight and high-energy physics, and the role of technological prowess as a primary indicator of national power (especially after Sputnik) led governments

Technological competition

to expand their mobilization of science and technology. By the 1960s, France, Britain, West Germany, Japan, China, and India, as well as the superpowers, faced the choice of finding ways of keeping up with the pace of technology, at least in targeted specialties, or else falling behind in the ever-quickening global competition. This process meant a sharp increase in state power, since such mobilization required the coordination of government, private corporations, and universities, as well as state planning, budgeting, management, and regulation.

By the late 1960s, when the United States was racing to the Moon and was still the apparent leader of the world economy, Europeans feared that economies of scale would prevent their countries from effectively competing in high technology. Jean-Jacques Servan-Schreiber wrote a best-seller entitled *The American Challenge,* and many European leaders feared a widening technology gap. In fact, events of the next few years proved that technocratic force-feeding of progress did not guarantee meaningful results. By the late 1970s other factors, including a commercial failure to exploit new technologies, declining labour productivity, loss of will, and determined foreign challenges to American primacy in specific technological sectors, combined to reduce overall U.S. leadership. American automobile, steel, and electronics industries were no longer able to compete against German, Japanese, and other imports even in their own market. By the 1980s mismanagement of the U.S. space program destroyed competitiveness in the pursuit that once symbolized American superiority. The United States was also hampered in world commerce by aggressive (some said unfair) foreign trade practices and in military competition by an aggressive Soviet campaign to acquire advanced Western technology by purchase or espionage.

The condition of permanent and planned technological revolution in the late 20th century has therefore created an anomalous situation in which all states are, or feel themselves to be, functionally "backward" all the time and racing against the projected innovations of others. Neither the American liberal model of high technology tying the world together with bonds of interdependence, nor the Soviet model of high technology tending to the triumph of socialism and the dissolution of the nation-state, has come to pass. Instead, what may with justice be called the Gaullist model of international affairs has triumphed in an age of technological revolution, in which national technocracies mobilize for competition against rival national technocracies.

Whether this pace of change can or should continue is a question of value judgments. Critics of the materialism that drives high technology, or the centralized power complexes it creates, may protest the continuing diversion of resources into research and development. By the late 1980s a series of spectacular failures in high technology—the explosion of the space shuttle *Challenger,* the toxic chemical leak in Bhopāl, India, and the accident in the Chernobyl nuclear reactor in the U.S.S.R.—seemed to hint that faith in the saving grace of ever more complex technology was a form of hubris. But so long as the well-being of the developed and developing societies seems to depend on technological progress, it is difficult to see how the engine of change can be shut down.

Failures of technology

The military balances. In one respect nuclear weapons and the rush to high technology changed the strategic picture forever, in that two states, if they chose to do so, could obliterate any other nation or group of nations in the world. Compared to the likely results of nuclear war, even the "total" wars of 1914 and 1939 were decidedly limited. But nuclear weapons, by their very destructiveness, also proved to be disproportionate to all real-world conflicts that arose after 1945. Their diplomatic use as the ultimate threat behind a state's policies also declined with the advent of bipolar parity. Perhaps this helps to explain why far fewer countries actually acquired nuclear weapons than was feared in the 1960s. Only the United States, the Soviet Union, Britain, France, China, and India openly, and perhaps Pakistan, South Africa, and Israel secretly, possess nuclear capabilities, and those nations presumably would contemplate using them only to ward off a threat to national survival (which the very possession of nuclear weapons is designed to deter). The possibility remains, however, of nuclear war occurring by accident, by unplanned escalation of a crisis situation, or by a fanatical government in Iran, Libya, or elsewhere acquiring and using an atomic bomb regardless of the consequences.

A significant trend in nuclear weapons since the early 1960s has been toward reliance on warheads of smaller and smaller yield, made possible by more accurate missile-guidance systems. While the number of warheads deployed has been multiplied by the MIRVing of long-range missiles and the proliferation of medium-range and cruise missiles in the European theatre, the United States and the U.S.S.R. came to rely chiefly on bombs in the 200- to 500-kiloton range, rather than the megaton weapons of the 1950s and 1960s. The total megatonnage in the U.S. stockpile in 1980 was less than one-fourth what it had been in 1960. But the fact that this trend could reduce collateral damage and civilian deaths in a limited exchange of weapons aimed at military targets could be viewed as damaging to world security, since it could conceivably make nuclear war less "unthinkable."

After 1945, perhaps because of nuclear deterrence, the great powers remained at peace with each other for a longer time than ever before in modern history. But at the same time the nuclear age has been unusually violent at lower technological levels of conflict. A reasonable estimate of the number of wars after 1945 would approach 150, with total casualties in the range of 20,000,000. Virtually all of these wars occurred in the less developed continents, where more than 100 new sovereignties came into existence after World War II. The great powers were involved directly in several of the wars as they struggled with each other through proxy states. But beyond those conflicts, the freedom of Third World countries to engage in warfare with each other or even (as in Vietnam) against major powers was a product of the international trade in modern weapons. In the decade after 1975 alone, 116 nations of the Third World spent almost $250,000,000,000 on imported arms, roughly a third from the United States, a third from the U.S.S.R., and the rest mostly from Britain and France. This proclivity of Third World governments to spend their precious revenue or loans on armaments (albeit with the encouragement of the Great Powers) not only slowed their economic progress but also created little balances of power, with dynamics of their own, in every corner of the world. Finally, small nations and political movements chose increasingly to assert their wills by the irregular exercise of force through guerrilla war and terrorism. Global strategy and tactics, therefore, became probably more complicated and less predictable than at any time in modern history, raising fears that the Middle East or Southeast Asia might someday serve, as the Balkans did in 1914, to spark a superpower conflict. The analogy with 1914 is also tempting insofar as state-supported terrorism, which triggered World War I, again became prevalent in the world. Incidents of terrorism increased from about 180 in 1968 to more than 750 in 1980, and deaths due to terrorist action from less than 100 to more than 600. The targets were almost exclusively Western, not Soviet-bloc, lives and property.

Proliferation of war

The universal reliance on deterring, rather than repelling, enemy attack in the nuclear age and the continuing technological revolution in nuclear and conventional weapons also placed an unprecedented premium on the size and quality of forces in being, rather than on a state's ability to gear up for war once it had begun. Hence peacetime military spending throughout the world was at an all-time high. U.S. and Soviet military budgets in the mid-1980s were both on the order of $150,000,000,000 per year, or over 6 percent of gross national product in the case of the United States, over 12 percent for the U.S.S.R. The total for the world as a whole was around $550,000,000,000, one-fifth of which was accounted for by the Third World. The trends in the superpower balance, however, clearly favoured the U.S.S.R. Not only had the Soviets' buildup since the 1960s provided them with numerical superiority in many categories of nuclear weapons, but they also remained predominant in conventional military

Armaments spending

strength while the Soviet fleet grew into the largest in the world. A Soviet Mediterranean squadron made its appearance in 1964, an Indian Ocean squadron in 1968, and a large Pacific fleet in the 1980s. The U.S.S.R. defended its enormous military budget by reference to the fact that all the other nuclear powers and industrial giants in the world (North America, western Europe, China, and Japan) were hostile to it. Whatever the cause and effect, Western observers viewed the combination of endemic Soviet economic weakness and cresting Soviet military might as potentially volatile.

Economic growth. Military and political power ultimately rests on an economic foundation. That of the Western alliance was still clearly superior. While the Soviet bloc's share of world manufacturing production rose from 16 percent in 1953 to 21.1 percent in 1980, it still lagged far behind that of the market economies (77.5 percent in 1953; 66.9 percent in 1980). The purely American share continued to decline, however, from its artificial post-World War II high (44.7 percent in 1953) to 31.5 percent in 1980, signaling a shift of relative power toward Japan and the industrializing Third World states. In 1959, 71 percent of the world's 150 largest firms were American; in 1976, only 44 percent were United States-based. In 1970 the world's 10 biggest banks were American; only two still were in 1980. Real growth rates in the developed world also leveled off after the booming 1960s. The average annual United States rate of 3.9 percent fell to 2.9 percent during the inflationary 1970s; Japanese annual real growth fell by more than half to 4.5 percent; French growth from 5.6 to 3.3 percent; West German from 4.7 to 2.6 percent; and British from 2.9 to 1.7 percent. American growth reached moderately healthy levels again in the 1980s, but at the cost of stupendous budget deficits in excess of $200,000,000,000 per year. Britain in that period also showed signs of escaping the malaise dating to World War I. Still, the shifts in world economic power, a function of the inevitable diffusion of technology combined with the ability of mobile capital to seek lower labour costs in less developed countries, were shrinking the command of world resources once exercised by the United States and western Europe. The U.S.S.R. similarly suffered a declining growth rate in the 1970s (3.3 percent) and 1980s and fell so far behind in computer technology and other critical new fields that the Japanese GNP had almost surpassed that of the U.S.S.R. by the late 1980s.

Decline of growth rates

What these numbers signified was that the Soviet promise of overtaking the West would not be fufilled. In fact, if the trends of the 1980s were to continue, the Soviet share of world economic power would soon begin to decline. But the U.S. economy had been losing ground relative to the world as a whole since the 1950s and could not sustain its mid-1980s level of military and domestic spending. This trend accounted for growing protectionist sentiment in the United States as well as for demands that the allies (especially Japan) who benefited from American protection share more of the burden of the common defense. To the extent this occurred, however, Washington would lose still more leverage over the policies of its allies.

The decline in economic performance among the developed countries, combined with the relative failure of many Third World development programs and the punitive effects of the 1970s price revolution in oil, also created a severe world financial crisis. The United States and Britain relied on Arab, Japanese, and other foreign investors to buy their government bonds and banknotes, while developing countries amassed an outstanding debt of $600,000,000,000 by 1983 (despite having received another $200,000,000,000 in foreign aid over the previous 20 years). Argentina, Brazil, Chile, Mexico, and the Philippines alone accounted for almost half the Third World debt. This sea of liabilities was fed by the largesse of international charities such as the International Monetary Fund and World Bank, poor judgment on the part of private bankers, political underwriting of loans by the U.S. government, and the need to recycle Arab petrodollars to Third World countries injured by oil price hikes.

Third World debt

Another economic worry that first attracted attention in the 1970s was the possible depletion of the world's raw materials, a fear aptly dramatized by the 1973 Arab oil boycott and given formal expression in the Club of Rome's apocalyptic warnings against unbridled industrial growth. In 1974 the U.S. Bureau of Mines estimated that at then-current rates of increase in usage, the world would run out of oil in 74 years—even if known reserves increased five times over that period. Similar calculations left the world less than 100 years of aluminum, copper, lead, natural gas, nickel, and tin.

The spread of modern industry around the world also caused a degree of pollution of the air, sea, and land environments that by 1970 some scientists considered critical. In the Western world, unilateral and cooperative efforts achieved a great deal in reducing pollution levels, but the environmental movement had far less influence in the Soviet bloc and Third World, where demand for productivity far outweighed concern over pollution. The long-range consequences of testing the global ecosystem's resiliency were only beginning to be understood.

International organization. The kind of multilateral cooperation required for global planning in such matters as resource allocation and pollution control seemed unlikely to develop short of a planetwide crisis. Organizations with purely administrative or humanitarian duties, like the International Telecommunication Union or the World Health Organization, were able to function well in spite of political rivalries, but others involved in more sensitive pursuits were paralyzed by continuing East–West conflicts between the Soviet bloc and the Western alliance, combined with growing North–South conflicts between the rich and poor nations. When the UN was founded, democracies comprised a majority of the 51 member states, and their hope was that the organization would foster free trade, human rights, pluralistic regimes, and compromise solutions to global problems. By 1980 the UN had expanded to more than 150 members, three-quarters of which were one-party states. Such regimes enjoyed centralized political control of their domestic societies and economies and demanded similar control of international regimes.

East–West, North–South conflicts

Neo-Marxist economists and "dependency" theorists defended the impatience of the developing countries, arguing that the gap between rich and poor nations could only be narrowed by a transfer of power over decision making rather than by the action of trade and investment. The hypocrisy of oppressive Third World regimes that criticized the Western democracies while patronizing the U.S.S.R. was nonetheless palpable.

Arguments over how to manage the global problems in every sphere touched on the great debate over the comparative wisdom of reliance on natural processes (the market, the balance of power, the patterns of demography) versus reliance on institutional controls. It was, at bottom, a debate over the primacy of means or ends and over the definition of freedom.

TOUR D'HORIZON

Within the context shaped by these forces of world politics, the governments of the world struggled to identify their choices. The U.S.S.R. remained a military giant dominating the Eurasian landmass and able to project force into all regions of the world. But short of an extremely risky war of expansion, the Soviets could make use of their military might only to intimidate the weaker links in their chain of potential enemies. Moreover, the Soviets were reaching the limit of what they could afford to spend on the military, in light of the poor performance of their planned economy. The Soviet regime also faced discontent among its subject nationalities, especially the rapidly expanding Muslim peoples of Central Asia. The accession of the relatively youthful Gorbachev, and his apparent commitment to *perestroika* and *glasnost* held out hope of genuine change in the U.S.S.R. But to repudiate Leninism too obviously would also undermine the ideological legitimacy of the Communist Party and the Soviet ruling class.

The Soviet Union

The Chinese seemed to have chosen pragmatism over ideology in their Four Modernizations program and commitment to détente with the United States. But the slogan itself suggested the limits the Chinese Communist Party hoped to place on modernization: Reform was to be con-

fined to the agricultural, industrial, technical, and military spheres and not be allowed to spill over into the political. Hence the course laid out by Deng Xiaoping bore a certain comparison with that of Stalin. China's challenges in the area of population control and development were so formidable that talk of China becoming a third superpower was premature.

An added worry for both Communist giants was the rising economic power of Japan. Japan had become a powerhouse of such proportions that Japanese capital alone would have a serious impact on the development of East Asia and international finance. Japan was heavily dependent on open international markets to sustain its prosperity and, despite its tradition of single-minded competitiveness, had to take care not to provoke its customers into restrictive trade measures. The United States could at any time increase pressure on Japan to assume a larger role in the defense of Pacific sea lanes. The Japanese would then face the decision of whether to retire their memories of World War II and take their place as a major power. To do so without provoking threatening responses from China and the U.S.S.R. would be a major challenge for Tokyo and Washington alike.

The future of South Asia depended heavily on the outcome of the Soviet–Afghan war and the course taken by Iran upon the death of Khomeini. A continuing fundamentalist upsurge in the Muslim world could pose a serious threat to European and Japanese sources of oil. The Israelis, meanwhile, could not rely forever on the division of their enemies, their own military prowess, the quiescence of the million hostile Palestinians in the occupied West Bank and Gaza, and financial aid from the United States. Indeed, violent Palestinian protests on the West Bank in 1988 seriously damaged the Israeli cause in the United States and in world opinion. But neither could Israel afford to relinquish its territorial buffers without firm guarantees of security.

The western European governments remained in a quandary, caught between their dependence on the United States for defense and their desire for relaxation of tensions with the Soviet bloc, between their desire for denuclearization of Europe and their unwillingness to expand their own conventional forces. As long as this situation obtained, the Europeans would be perceived by Americans as free riders. The Common Market countries together had a GNP larger than that of the U.S.S.R. and two-thirds the size of the American, and had since 1950 enjoyed an age of peace and prosperity unprecedented in their history. An American pullout from Europe carried the risk that the Europeans would gradually submit to Soviet pressure rather than unite and defend themselves. But NATO had proved itself an unusually, perhaps uniquely, resilient alliance system, and a crisis in Western solidarity seemed likely only in the case of a decisive leftist-pacifist victory in the domestic politics of Britain or West Germany.

Latin America seemed to be a new storm centre in world politics. The population explosion and awakening of the peasant masses in this heretofore stagnant and quiescent continent presented opportunities for the Soviets, challenges to the United States, and grave dangers for the Latin-American peoples themselves. They faced terrific problems of development that, however they were met or not met, would tend to destabilize domestic politics. The United States was especially concerned with the region, not only because of its proximity but also because of the complex created in the American mind by Communist Cuba, which sat like a West Berlin inside the American sphere of influence. Thus, Central American and Caribbean states had only restricted room to maneuver in their domestic politics and little chance of working out their destinies wholly on their own. What Latin America appeared to need was the kind of regional leadership and inspiration it had not seen since the days of Simón Bolívar. But the largest states, like Brazil, Argentina, and Mexico, had failed to produce such leaders, while regional organizations were either dominated by the United States or lacked sufficient prestige and force to act effectively.

The United States faced problems of adjustment greater than that of any power except perhaps the U.S.S.R. American economic power would continue to shrink, partly because of the diffusion of manufacturing industries and technology around the world and partly because of the high percentage of U.S. brainpower and resources devoted to military-related research. The U.S. position in world affairs was not unlike that of the British Empire around the turn of the century, when Britain was still the overall naval, financial, commercial, and technological leader but was challenged by numerous industrializing powers anxious to build navies, markets, and colonial empires of their own. No longer able to defend the world order alone, Britain concluded a series of alliances to contain its rivals, especially Germany, and in the process lost some of its control over its own diplomacy. The United States had now to choose between maintaining its level of involvement in world affairs and deferring increasingly to allied opinion, or reverting to its other and older tradition of isolationism. But to the extent that isolationism was no longer an option, future American presidents would face the challenge of building a domestic base of support for foreign policies that neither fled from the realities of the world—including Soviet espionage and military power, global terrorism, and Third World poverty—nor undermined the values—including personal liberty, free enterprise, and limited government—that make the United States and its Western allies worth defending in the first place.

apan

Western Europe

Latin America

The United States

BIBLIOGRAPHY

General works: Important international histories of the 20th century are WILLIAM R. KEYLOR, *The Twentieth-Century World: An International History* (1984); and FELIX GILBERT, *The End of the European Era, 1890 to the Present,* 3rd ed. (1984). See also PAUL JOHNSON, *Modern Times: The World from the Twenties to the Eighties* (1983), episodic but insightful; and RENÉ ALBRECHT-CARRIÉ, *A Diplomatic History of Europe Since the Congress of Vienna,* rev. ed. (1973), a standard survey. The interpretive essays by LUDWIG DEHIO, *The Precarious Balance: Four Centuries of the European Power Struggle* (1962; originally published in German, 1948); and HAJO HOLBORN, *The Political Collapse of Europe* (1951, reprinted 1982), put the 20th century in a longer perspective. On war and intelligence, see MICHAEL HOWARD, *War in European History* (1976), and *War and the Liberal Conscience* (1978); THEODORE ROPP, *War in the Modern World* (1959, reprinted 1981); and ERNEST R. MAY (ed.), *Knowing One's Enemies: Intelligence Assessment Before the Two World Wars* (1984). Theoretical works on the nature of international relations include HANS J. MORGENTHAU and KENNETH W. THOMPSON, *Politics Among Nations: The Struggle for Power and Peace,* 6th ed. (1985); KENNETH N. WALTZ, *Man, the State, and War: A Theoretical Analysis* (1959, reprinted 1965); and F.H. HINSLEY, *Power and the Pursuit of Peace: Theory and Practice in the History of Relations Between States* (1963). JULIUS STONE, *Visions of World Order: Between State Power and Human Justice* (1984), explores the laws governing international relations in the modern world. Key terms and concepts of international politics are analyzed in DAVID WEIGALL, *Britain & the World, 1815–1986: A Dictionary of International Relations* (1987); and, in a larger work, EDMUND JAN OSMÁNCZYK, *The Encyclopedia of the United Nations and International Agreements* (1985).

World War I: Works on the origins of World War I include LUIGI ALBERTINI, *The Origins of the War of 1914,* 3 vol. (1952–57, reprinted 1980; originally published in Italian, 1942–43); LAURENCE LAFORE, *The Long Fuse: An Interpretation of the Origins of World War I* (1965, reprinted 1981); DWIGHT E. LEE, *The Outbreak of the First World War: Causes and Responsibilities,* 4th ed. (1975); V.R. BERGHAHN, *Germany and the Approach of War in 1914* (1973); ZARA S. STEINER, *Britain and the Origins of the First World War* (1977); and JAMES JOLL, *The Origins of the First World War* (1984). Diplomacy of the war years is explored in GERD HARDACH, *The First World War, 1914–1918* (1977; originally published in German, 1973); BERNADOTTE E. SCHMITT and HAROLD C. VEDELER, *The World in the Crucible, 1914–1919* (1984); Z.A.B. ZEMAN, *The Gentlemen Negotiators* (1971; U.K. title, *A Diplomatic History of the First World War*); and ARNO J. MAYER, *Political Origins of the New Diplomacy, 1917–1918* (1959, reissued 1967 as *Wilson vs. Lenin*). (*Peacemaking 1919*): The history of the Paris Peace Conference is found in the reminiscences of the principal participants, which are regrettably dated and tendentious, except for HAROLD NICOLSON, *Peacemaking, 1919* (1933, reprinted 1984), a memoir of lasting value. On Wilsonianism, see N. GORDON LEVIN, JR., *Woodrow Wilson and World Politics: America's Response to War and Revolution* (1968). On the Peace Conference and the Russian problem, see ARNO J. MAYER, *Politics and Diplo-*

macy of Peacemaking: Containment and Counterrevolution at Versailles, 1918–1919 (1967); GEORGE F. KENNAN, *Russia and the West Under Lenin and Stalin* (1961); and STEPHEN WHITE, *The Origins of Detente: The Genoa Conference and Soviet-Western Relations, 1921–1922* (1985). On French security during and after 1919, see WALTER A. MCDOUGALL, *France's Rhineland Diplomacy, 1914–1924: The Last Bid for a Balance of Power in Europe* (1978); and MELVYN P. LEFFLER, *The Elusive Quest: America's Pursuit of European Stability and French Security, 1919–1933* (1979). On reparations at the Peace Conference, see MARC TRACHTENBERG, *Reparation in World Politics: France and European Economic Diplomacy, 1916–1923* (1980). On war debts, see DENISE ARTAUD, *La Question des dettes interalliées et la reconstruction de l'Europe (1917–1929),* 2 vol. (1978, thesis). See also JOHN MAYNARD KEYNES, *The Economic Consequences of the Peace* (1919, reissued 1971); and ÉTIENNE MANTOUX, *The Carthaginian Peace: or, The Economic Consequences of Mr. Keynes* (1946, reprinted 1979).

(*The fragile 1920s*): PIERRE RENOUVIN, *War and Aftermath, 1914–1929* (1968; originally published in French, 1957); and RAYMOND J. SONTAG, *A Broken World, 1919–1939* (1971), provide excellent historical summaries. On economic history, see DEREK H. ALDCROFT, *From Versailles to Wall Street, 1919–1929* (1977). A keen portrayal of the statesmen of the period is offered in GORDON A. CRAIG and FELIX GILBERT (eds.), *The Diplomats: 1919–1939* (1953; reissued in 2 vol., 1971–72). On the settlement in East Asia and U.S.–Japanese–Chinese relations, see AKIRA IRIYE, *After Imperialism: The Search for a New Order in the Far East, 1921–1931* (1965), and *Across the Pacific: An Inner History of American-East Asian Relations* (1967). On U.S. policy in Latin America, see GORDON CONNELL-SMITH, *The United States and Latin America: An Historical Analysis of Inter-American Relations* (1974). For the broadest overview of European diplomacy in the 1920s, reinterpreted in light of new documentation, see CHARLES S. MAIER, *Recasting Bourgeois Europe: Stabilization in France, Germany, and Italy in the Decade After World War I* (1975). On Franco-German relations, see JACQUES BARIÉTY, *Les Relations franco-allemandes après la première guerre mondiale* (1977). On the settlements of the mid-1920s, see STEPHEN A. SCHUKER, *The End of French Predominance in Europe: The Financial Crisis of 1924 and the Adoption of the Dawes Plan* (1976); DAN P. SILVERMAN, *Reconstructing Europe After the Great War* (1982); and JON JACOBSON, *Locarno Diplomacy: Germany and the West, 1925–1929* (1972). The U.S.S.R. is covered exhaustively and insightfully in ADAM B. ULAM, *Expansion and Coexistence: Soviet Foreign Policy, 1917–73,* 2nd ed. (1974). The U.S.-Soviet contacts of the 1920s are explored in JOAN HOFF-WILSON, *Ideology and Economics: U.S. Relations with the Soviet Union, 1918–1933* (1974). On Geneva, see F.P. WALTERS, *A History of the League of Nations,* 2 vol. (1952, reprinted 1986); and GEORGE SCOTT, *The Rise and Fall of the League of Nations* (1973). Eastern European diplomacy is expertly covered by PIOTR S. WANDYCZ, *France and Her Eastern Allies, 1919–1925: French-Czechoslovak-Polish Relations from the Paris Peace Conference to Locarno* (1962, reprinted 1974); and F. GREGORY CAMPBELL, *Confrontation in Central Europe: Weimar Germany and Czechoslovakia* (1975).

(*Origins of World War II*): E.M. ROBERTSON (ed.), *The Origins of the Second World War: Historical Interpretations* (1971), compiles articles on the debate over A.J.P. TAYLOR, *The Origins of the Second World War* (1961; reissued with a new introduction, 1983), which is still good on British and French policy, but idiosyncratic and controversial on Hitler. ANTHONY P. ADAMTHWAITE, *The Making of the Second World War,* 2nd ed. (1979), offers an informative historical summary. See also the standard PIERRE RENOUVIN, *World War II and Its Origins: International Relations, 1929–1945* (1968; originally published in French, 1958); and the classic memoir by WINSTON CHURCHILL, *The Gathering Storm* (1948, reissued 1985). Nazi diplomacy is covered in detail in GERHARD L. WEINBERG, *The Foreign Policy of Hitler's Germany: Diplomatic Revolution in Europe, 1933–36* (1970), and *The Foreign Policy of Hitler's Germany: Starting World War II, 1937–1939* (1980). Other good interpretations include ALAN BULLOCK, *Hitler, a Study in Tyranny,* rev. ed. (1962); KLAUS HILDEBRAND, *The Foreign Policy of the Third Reich* (1973; originally published in German, 1971); and EBERHARD JÄCKEL, *Hitler's Weltanschauung: A Blueprint for Power* (1972; reissued as *Hitler's World View,* 1981; originally published in German, 1969). On Fascist Italy, see ELIZABETH WISKEMANN, *The Rome-Berlin Axis: A Study of the Relations Between Hitler and Mussolini,* new rev. ed. (1966); and DENIS MACK SMITH, *Mussolini's Roman Empire* (1976). On France, see ANTHONY P. ADAMTHWAITE, *France and the Coming of the Second World War, 1936–1939* (1977). On British appeasement, see MARTIN GILBERT, *The Roots of Appeasement* (1966); A.L. ROWSE, *Appeasement: A Study in Political Decline, 1933–1939* (1961); and TELFORD TAYLOR, *Munich: The Price of Peace* (1979). On the United States, see MANFRED JONAS, *Isolationism in America, 1935–1941* (1966); ROBERT A. DIVINE, *The Illusion of Neutrality* (1962); and ARNOLD A. OFFNER, *American Appeasement: United States Foreign Policy and Germany, 1933–1938* (1969, reissued 1976). The economic collapse of the 1930s is covered in CHARLES P. KINDLEBERGER, *The World in Depression, 1929–1939,* rev. ed. (1986); and its diplomatic effects in DAVID E. KAISER, *Economic Diplomacy and the Origins of the Second World War: Germany, Britain, France, and Eastern Europe, 1930–1939* (1980). On military preparations, see DONALD CAMERON WATT, *Too Serious a Business: European Armed Forces and the Approach to the Second World War* (1975); and ROBERT J. YOUNG, *In Command of France: French Foreign Policy and Military Planning, 1933–1940* (1978). On the origins of the Pacific war, see ARNOLD A. OFFNER, *The Origins of the Second World War: American Foreign Policy and World Politics, 1917–1941* (1975, reprinted 1986).

World War II and after: A monumental survey of European politics during World War II is presented in SIR LLEWELLYN WOODWARD, *British Foreign Policy in the Second World War,* 5 vol. (1962–70). Other works on diplomatic developments include ROBERT A. DIVINE, *The Reluctant Belligerent: American Entry into World War II,* 2nd ed. (1979), and *Roosevelt and World War II* (1969); HERBERT FEIS, *Churchill, Roosevelt, and Stalin: The War They Waged and the Peace They Sought,* 2nd ed. (1967); ANDREAS HILLGRUBER, *Hitlers Strategie: Politik und Kriegführung, 1940–1941,* 2nd ed. (1982); MACGREGOR KNOX, *Mussolini Unleashed, 1939–1941: Politics and Strategy in Fascist Italy's Last War* (1982); WILLIAM H. MCNEILL, *America, Britain, and Russia: Their Co-operation and Conflict, 1941–1946* (1953, reprinted 1970); ALAN S. MILWARD, *War, Economy, and Society, 1939–1945* (1977); and GORDON WRIGHT, *The Ordeal of Total War, 1939–1945* (1968). (*General histories*): For summaries of global relations since 1945, see PETER CALVOCORESSI, *World Politics Since 1945,* 5th ed. (1987); PETER LANE, *Europe Since 1945* (1985); ROBERT A. DIVINE, *Since 1945: Politics and Diplomacy in Recent American History,* 3rd ed. (1985); RAYMOND ARON, *The Imperial Republic: The United States and the World, 1945–1973* (1974, reprinted 1982; originally published in French, 1973); PAUL Y. HAMMOND, *Cold War and Détente: The American Foreign Policy Process Since 1945* (1975); JOHN LEWIS GADDIS, *Strategies of Containment: A Critical Appraisal of Postwar American National Security Policy* (1982); LAWRENCE FREEDMAN, *The Evolution of Nuclear Strategy* (1981); and a revisionist work by WALTER LAFEBER, *America, Russia, and the Cold War, 1945–1984,* 5th ed. (1985). On the Middle East, see TREVOR N. DUPUY, *Elusive Victory: The Arab-Israeli Wars, 1947–1974* (1978, reissued 1984); RITCHIE OVENDALE, *The Origins of the Arab-Israeli Wars* (1984); and GIDEON RAFAEL, *Destination Peace: Three Decades of Israeli Foreign Policy* (1981). On European recovery, see WALTER LAQUEUR, *The Rebirth of Europe* (1970); and RICHARD MAYNE, *The Recovery of Europe: From Devastation to Unity* (1970).

(*Origins of the Cold War*): On the Stalin-Truman years, see HARRY S. TRUMAN, *Memoirs,* 2 vol. (1955–56, reprinted 1986–87); DEAN ACHESON, *Present at the Creation: My Years in the State Department* (1969, reprinted 1987); GEORGE F. KENNAN, *Memoirs,* 2 vol. (1967–72); and DWIGHT D. EISENHOWER, *The White House Years,* 2 vol. (1963–65). Insightful histories include JOHN LEWIS GADDIS, *The United States and the Origins of the Cold War, 1941–1947* (1972), and *The Long Peace: Inquiries into the History of the Cold War* (1987); PAUL SEABURY, *The Rise and Decline of the Cold War* (1967); LOUIS J. HALLE, *The Cold War as History* (1967, reprinted 1971); DANIEL YERGIN, *Shattered Peace: The Origins of the Cold War and the National Security State* (1977); and HUGH THOMAS, *Armed Truce: The Beginnings of the Cold War, 1945–46* (1986).

The following are works of scholarship on the Cold War by authors who clearly regarded themselves as left-revisionist: WILLIAM APPLEMAN WILLIAMS, *The Tragedy of American Diplomacy,* 2nd rev. ed. (1972); GABRIEL KOLKO, *The Roots of American Foreign Policy: An Analysis of Power and Purpose* (1969); GAR ALPEROVITZ, *Atomic Diplomacy: Hiroshima and Potsdam: The Use of the Atomic Bomb and the American Confrontation with Soviet Power,* rev. ed. (1985); DENNA F. FLEMING, *The Cold War and Its Origins, 1917–1960,* 2 vol. (1961); and DAVID HOROWITZ, *The Free World Colossus: A Critique of American Foreign Policy in the Cold War,* rev. ed. (1971). ROBERT J. MADDOX, *The New Left and the Origins of the Cold War* (1973), provides a critique of their logic and use of evidence.

The Soviet side is discussed in VOJTECH MASTNY, *Russia's Road to the Cold War: Diplomacy, Warfare, and the Politics of Communism, 1941–1945* (1979); ADAM B. ULAM, *The Rivals: America and Russia Since World War II* (1971, reprinted 1983); DAVID HOLLOWAY, *The Soviet Union and the Arms Race* (1983); and THOMAS W. WOLFE, *Soviet Power and Europe, 1945–1970* (1970). MARSHALL D. SHULMAN, *Stalin's Foreign Policy Reappraised* (1963, reissued 1985); and WILLIAM TAUBMAN, *Stalin's American Policy: From Entente to Detente to Cold War* (1982),

are sympathetic accounts. On the "wise men" surrounding Truman during the late 1940s, see the critique by LLOYD C. GARDNER, *Architects of Illusion: Men and Ideas in American Foreign Policy, 1941–1949* (1970); and a later, more sympathetic work WALTER ISAACSON and EVAN THOMAS, *The Wise Men: Six Friends and the World They Made: Acheson, Bohlen, Harriman, Kennan, Lovett, McCloy* (1986). The standard earlier work on atomic policy is *A History of the United States Atomic Energy Commission,* vol. 1 by RICHARD G. HEWLETT and OSCAR E. ANDERSON, *The New World, 1939/46* (1962), and vol. 2 by RICHARD G. HEWLETT and FRANCIS DUNCAN, *Atomic Shield, 1947/1952* (1969). Making use of declassified material is a later work by GREGG HERKEN, *The Winning Weapon: The Atomic Bomb in the Cold War, 1945–1950* (1980). See also HERBERT F. YORK, *The Advisors: Oppenheimer, Teller, and the Superbomb* (1976); and ROBERT A. DIVINE, *Blowing on the Wind: The Nuclear Test Ban Debate, 1954–1960* (1978). On the origins of the Korean war, see BRUCE CUMINGS (ed.), *Child of Conflict: The Korean-American Relationship, 1943–1953* (1983); and on the war itself, DAVID REES, *Korea: The Limited War* (1964).

(*"Total" Cold War, 1957–72*): The concept of "total Cold War" is described in WALTER A. MCDOUGALL, *The Heavens and the Earth: A Political History of the Space Age* (1985). World trends after Sputnik are also the subject of W.W. ROSTOW, *The Diffusion of Power: An Essay in Recent History* (1972). See also the extremely interesting memoirs by NIKITA KHRUSHCHEV, *Khrushchev Remembers,* trans. from Russian (1970), and *Khrushchev Remembers: The Last Testament,* trans. from Russian (1974); RICHARD M. NIXON, *RN: The Memoirs of Richard Nixon* (1978); and HENRY KISSINGER, *White House Years* (1979). Also see ARTHUR M. SCHLESINGER, JR., *A Thousand Days: John F. Kennedy in the White House* (1965, reprinted 1983); ROGER HILSMAN, *To Move a Nation: The Politics of Foreign Policy in the Administration of John F. Kennedy* (1967); GRAHAM T. ALLISON, *Essence of Decision: Explaining the Cuban Missile Crisis* (1971); GLENN T. SEABORG, *Kennedy, Khrushchev, and the Test Ban* (1981); and DESMOND BALL, *Politics and Force Levels: The Strategic Missile Program of the Kennedy Administration* (1980). On the Sino-Soviet split, see ALFRED D. LOW, *The Sino-Soviet Dispute: An Analysis of the Polemics* (1976), continued in his *Sino-Soviet Confrontation Since Mao Zedong: Dispute, Detente, or Conflict?* (1987); DONALD S. ZAGORIA, *The Sino-Soviet Conflict, 1956–1961* (1962, reissued 1969); and WILLIAM E. GRIFFITH, *The Sino-Soviet Rift, Analyzed and Documented* (1964; reissued as *Sino-Soviet Relations, 1964–1965,* 1967). On the phenomenon of Gaullism, see CHARLES DE GAULLE, *Memoirs of Hope: Renewal and Endeavor* (1971; originally published in French, 1970); W.W. KULSKI, *De Gaulle and the World: The Foreign Policy of the Fifth French Republic* (1966); and WILFRID L. KOHL, *French Nuclear Diplomacy* (1971). On postwar German policies, see WILLIAM E. GRIFFITH, *The Ostpolitik of the Federal Republic of Germany* (1978); GERHARD WETTIG, *Community and Conflict in the Socialist Camp: The Soviet Union, East Germany, and the German Problem, 1965–1972* (1975; originally published in German, 1972–73); and PETER H. MERKL, *German Foreign Policies, West & East: On the Threshold of a New European Era* (1974).

(*Third World countries*): General works on European decolonization include JOHN D. HARGREAVES, *The End of Colonial Rule in West Africa: Essays in Contemporary History* (1979); PROSSER GIFFORD and W. ROGER LEWIS (eds.), *The Transfer of Power in Africa: Decolonization, 1940–1960* (1982); and ANN WILLIAMS, *Britain and France in the Middle East and North Africa, 1914–1967* (1968); HUGH THOMAS, *Suez* (1967; U.K. title, *Suez Affair*). On Soviet penetration of the Third World, see ROBERT C. HORN, *Soviet-Indian Relations: Issues and Influence* (1982); CHRISTOPHER STEVENS, *The Soviet Union and Black Africa* (1976); and ROBERT H. DONALDSON (ed.), *The Soviet Union in the Third World: Successes and Failures* (1981). The Vietnam War is treated in WILLIAM S. TURLEY, *The Second Indochina War: A Short Political and Military History, 1954–1975* (1986); STANLEY KARNOW, *Vietnam: A History* (1983); and GEORGE C. HERRING, *America's Longest War: The United States and Vietnam, 1950–1975,* 2nd ed. (1986). On U.S. involvement, see DAVID HALBERSTAM, *The Best and the Brightest* (1972, reprinted 1983); on the Tet Offensive, PETER BRAESTRUP, *Big Story: How the American Press and Television Reported and Interpreted the Crisis of Tet 1968 in Vietnam and Washington,* 2 vol. (1977); on U.S. bombing strategy, WALLACE J. THIES, *When Governments Collide: Coercion and Diplomacy in the Vietnam Conflict, 1964–1968* (1980); and on American military mistakes, HARRY G. SUMMERS, JR., *On Strategy: A Critical Analysis of the Vietnam War* (1982).

(*The Global Village since 1972*): For the contemporary period, memoirs become increasingly important. All the principals in the Carter administration produced lengthy accounts: JIMMY CARTER, *Keeping Faith: Memoirs of a President* (1982); CYRUS VANCE, *Hard Choices: Critical Years in America's Foreign Policy* (1983); and ZBIGNIEW BRZEZINSKI, *Power and Principle: Memoirs of the National Security Adviser, 1977–1981* (1983). A fine summary of the administration is GADDIS SMITH, *Morality, Reason, and Power: American Diplomacy in the Carter Years* (1986). China since 1970 is the subject of ROY MEDVEDEV, *China and the Superpowers,* trans. from Russian (1986); and C.G. JACOBSEN, *Sino-Soviet Relations Since Mao: The Chairman's Legacy* (1981). Middle Eastern diplomacy is expertly analyzed in BAHGAT KORANY and ALI E. HILLAL DESSOUKI, *The Foreign Policies of Arab States* (1984); and general Third World problems in STEPHEN D. KRASNER, *Structural Conflict: The Third World Against Global Liberalism* (1985). Soviet policy is the subject of ADAM B. ULAM, *Dangerous Relations: The Soviet Union in World Politics, 1970–1982* (1983); RICHARD F. STAAR, *USSR Foreign Policies After Detente,* rev. ed. (1987); and ROBERTA GOREN, *The Soviet Union and Terrorism* (1984). A thorough account of the decline of detente between the United States and U.S.S.R. is given in RAYMOND L. GARTHOFF, *Détente and Confrontation: American–Soviet Relations from Nixon to Reagan* (1985).

(*Arms race and disarmament*): Specific issues of armament and disarmament are discussed in NATIONAL ACADEMY OF SCIENCES (U.S.), *Nuclear Arms Control: Background and Issues* (1985); CURT GASTEYGER, *Searching for World Security: Understanding Global Armament and Disarmament* (1985); and WILLIAM T. LEE and RICHARD F. STAAR, *Soviet Military Policy Since World War II* (1986). For divergent views on the future of nuclear weapons, see KEITH B. PAYNE, *Strategic Defense: "Star Wars" in Perspective* (1986); CRAIG SNYDER (ed.), *The Strategic Defense Debate: Can "Star Wars" Make Us Safe?* (1986); JAMES H. WYLLIE, *European Security in the Nuclear Age* (1986); DONALD M. SNOW, *The Necessary Peace: Nuclear Weapons and Superpower Relations* (1987); ANGELO CODERILLA, *While Others Build: A Commonsense Approach to the Strategic Defense Initiative* (1988); and especially FREEMAN DYSON, *Weapons and Hope* (1984); and ROBERT M. LAWRENCE, *Strategic Defense Initiative: Bibliography and Research Guide* (1987).

(W.A.McD.)

International Trade

International trade includes all economic transactions that are made between countries. Among the items commonly traded are consumer goods, such as television sets and clothing; capital goods, such as machinery; and raw materials and food. Other transactions involve services, such as travel services and payments for foreign patents. International trade transactions are facilitated by international financial payments, in which the private banking system and the central banks of the trading nations play important roles.

International trade and the accompanying financial transactions are conducted generally toward the purpose of providing a nation with commodities it lacks in exchange for those that it produces in abundance; such transactions, functioning with other economic policies, generally improve the standard of living of a nation. This article provides a historical and contemporary overview of the structure of international trade, of the classic controversy over free versus controlled trade, and of the problems that arise in transactions between nations. (Ed.)

For statistical information on international trade, see the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Five, sections 531 through 536.

This article is divided into the following sections:

The theory of international trade 825
History 825
Comparative-advantage analysis 825
Sources of comparative advantage 827
State interference in international trade 828
Methods of interference 828
Arguments for and against interference 829
International trade arrangements 830
History of modern trade policies 830
Trade agreements 831
Economic integration 833
Patterns of trade 839
Degrees of national participation 839
Trade among developed countries 839
Trade between developed and developing countries 839
Soviet foreign trade 839
International commodity trade 840
Primary commodity markets 840
Price movements 840
Interests of the less-developed countries 842
OPEC and oil 842
International payments and exchange 843
Balance-of-payments accounting 843
Adjusting for fundamental disequilibrium 845
Foreign exchange markets 846
The gold standard 846
The International Monetary Fund 847
The IMF system of parity (pegged) exchange rates 847
Floating exchange rates 852
The international debt crisis 853
Bibliography 853

The theory of international trade

HISTORY

Accounts of barter of goods or of services among different peoples can be traced back almost as far as the record of human history. International trade, however, is specifically an exchange between members of different nations, and accounts and explanations of such trade begin (despite fragmentary earlier discussion) only with the rise of the modern nation-state at the close of the European Middle Ages. As political thinkers and philosophers began to examine the nature and function of the nation, trade with other nations became a particular topic of their inquiry. It is, accordingly, no surprise to find one of the earliest attempts to describe the function of international trade within that highly nationalistic body of thought now known as "mercantilism." Mercantilist analysis, which reached the peak of its influence upon European thought in the 16th and 17th centuries, focused directly upon the welfare of the nation. It insisted that the acquisition of wealth, particularly wealth in the form of gold, was of paramount importance for national policy. Mercantilists took the virtues of gold almost as an article of faith; consequently, they never undertook to explain adequately why gold deserved such a high priority in their economic plans.

Mercantilism

The trade policy dictated by mercantilist philosophy was accordingly simple: encourage exports, discourage imports, and take the proceeds of the resulting export surplus in gold. Because of their nationalistic bent, mercantilist theorists either brushed aside or else did not realize that, from an international viewpoint, this policy would necessarily prove self-defeating. The nation that successfully gains an export surplus must ordinarily do so at the expense of one or more other nations that record a matching import surplus. Mercantilists' ideas often were intellectually shallow, and indeed their trade policy may have been little more than a rationalization of the interests of a rising merchant class that wanted wider markets—hence the emphasis on expanding exports—coupled with protection against competition in the form of imported goods. Yet mercantilist policies, as will be noted later, are by no means completely dead today.

COMPARATIVE-ADVANTAGE ANALYSIS

Adam Smith and David Ricardo

The British school of "classical economics" began in no small measure as a reaction against the inconsistencies of mercantilist thought. Adam Smith was the 18th-century founder of this school; his famous work, *The Wealth of Nations,* is in part an antimercantilist tract. In *The Wealth of Nations,* Smith emphasized the importance of specialization as a source of increased output, and he treated international trade as a particular instance of specialization: in a world where productive resources are scarce and human wants cannot be completely satisfied, each nation should specialize in the production of goods it is particularly well equipped to produce; it should export part of this production, taking in exchange other goods that it cannot so readily turn out. Smith did not expand these ideas at much length; but David Ricardo, the second great classical economist, developed them into the "principle of comparative advantage," a principle still to be found, much as Ricardo spelled it out, in every international trade text.

Simplified theory of comparative advantage. For clarity of exposition, the theory of comparative advantage is usually first outlined as though only two countries and only two commodities were involved, although it is by no means limited to such cases. Again for clarity, the cost of production is usually measured only in terms of labour time and effort; the cost of a unit of cloth, for example, might be given as two hours of work. The two countries will here be A and B, and the two commodities wine and cloth. The labour time required to produce a unit of either commodity in either country is as follows:

Cost of Production (labour time)

	country A	country B
wine (1 unit)	1 hour	2 hours
cloth (1 unit)	2 hours	6 hours

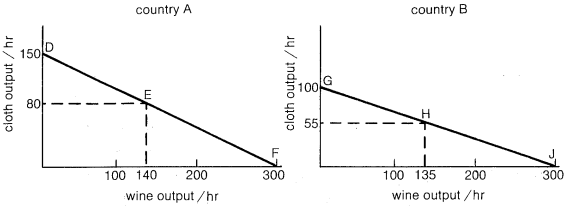
As compared with country A, country B is productively inefficient. Its workers need more time to turn out a unit of wine or a unit of cloth. This relative inefficiency may result from differences in climate, in worker training or skill, or in the amount of available tools and equipment, or from numerous other possible reasons. Ricardo took it for granted that such differences do exist, and he was not concerned with their origins.

Country A is said to have an absolute advantage in the production of both wine and cloth because it is more efficient in the production of both goods. Accordingly, A's absolute advantage seemingly invites the conclusion that country B could not possibly compete with country A, and indeed that if trade were to be opened up between them, country B would be competitively overwhelmed. Ricardo insisted that this conclusion is false. The critical factor is that country B's disadvantage is less pronounced in wine production, in which its workers require only twice as much time for a single unit as do the workers in A, than it is in cloth production, in which the required time is three times as great. This means, Ricardo pointed out, that country B will have a comparative advantage in wine production. Both countries will profit, in terms of the real income they enjoy, if country B specializes in wine production, exporting part of its output to country A, and if country A specializes in cloth production, exporting part of its output to country B. Paradoxical though it may seem, it is preferable for country A to leave wine production to country B, despite the fact that A's workers can produce wine of equal quality in half the time that B's workers can do so.

Ricardo's conclusion

To illustrate this conclusion, it can be considered that country A's total labour force consists of 300 workers. Disregarding the possibility of trade with B, A then has a choice of various outputs of cloth and of wine, depending on the number of workers engaged in each of the two occupations. This range of choices is illustrated by the line DEF in the accompanying diagram. If all 300 labourers work on cloth production, total hourly cloth output will be 150 units (point D in the diagram), since each such unit requires two hours' labour. At the other extreme, if

all labour works on wine production, wine output will be 300 units per hour (point F). Any intermediate point on the line DEF is possible. Point E, for example, indicates 80 units of cloth produced each hour (160 workers so employed) and 140 units of wine (employing the other 140 workers). DEF is country A's "production possibility" line. If it does not trade with country B and so can consume only what it produces itself, DEF will also be country A's "consumption possibility" line; it will choose some point thereon, depending on the preferences of its citizens for wine and cloth. DEF represents the limit of production and consumption possibilities; points above and to the right of DEF are unattainable.



Production possibilities for cloth and wine (see text).

In the right-hand diagram, the line GHJ has exactly the same production and consumption significance for country B—assuming its total force to be 600 workers (so as to make it roughly equal to A in total output capacity). The position of the line GHJ reflects the fact that labour in country B requires two hours to produce a unit of wine, and six hours for a unit of cloth.

One may consider that A and B are initially isolated from one another. Country A has chosen point E (80 cloth, 140 wine) as its production–consumption point. Country B has chosen point H (55 cloth, 135 wine). The opportunity of free trade between the two countries is now opened up. If both countries want to attain the higher levels of production and consumption available to them through specializing on and trading of the product for which they have a comparative advantage, Country A will shift its entire labour force to cloth production, and Country B will shift its entire labour force to wine production. A possible barter rate (setting aside the detail of how this would be worked out) might be one cloth for two and one-half wine. Country A might then choose to export 60 units of its hourly cloth output of 150, keeping the other 90 for domestic consumption. In exchange for this 60 cloth (at the 1-for-2½ exchange rate) it would receive 150 wine. A's real income position is thus improved in comparison with pretrade point E: cloth for domestic consumption has risen from 80 to 90, and wine consumption has risen from 140 to 150. Country B enjoys a similar gain. In comparison with pre-trade point H, its cloth consumption has risen from 55 to 60, and wine consumption has risen from 135 to 150.

rice
ıcentives
› trade

The incentive to export and to import can be explained in price terms. In country A (before international trade), the price of cloth ought to be twice that of wine, since a unit of cloth requires twice as much labour effort. If this price ratio is not satisfied, one of the two commodities will be overpriced and the other underpriced. Labour will then move out of the underpriced occupation and into the other, until the resulting shortage of the underpriced commodity drives up its price. In country B (again, before trade), a cloth unit should cost three times as much as a wine unit, since a unit of cloth requires three times as much labour effort. Hence, a typical before-trade price relationship, matching the underlying real cost ratio in each country, might be as follows:

	country A	country B
Price of wine per unit	$5	£1
Price of cloth per unit	$10	£3

The absolute levels of price do not matter. All that is necessary is that in each country the ratio of the two prices should match the labour–cost ratio.

As soon as the opportunity of exchange between the two countries is opened up, the difference between the wine–cloth price ratio in country A (namely, 5:10, or 1:2) and that in country B (which is 1:3) provides the opportunity of a trading profit. Cloth will begin to move from A to B, and wine from B to A. A trader in A, starting with a capital of $10 for example, would buy a unit of cloth, sell it in B for £3, buy 3 units of B's wine with the proceeds, and sell this in A for $15. (This example assumes, for simplicity, that costs of transporting goods are negligible or zero. The introduction of transport costs complicates the analysis somewhat, but it does not change the conclusions, unless these costs are so high as to make trade impossible.)

So long as the ratio of prices in country A differs from that in country B, the flow of goods between the two countries will steadily increase as traders become increasingly aware of the profit to be obtained by moving goods between the two countries. Prices, however, will be affected by these changing flows of goods. The wine price in country A, for example, can be expected to fall as larger and larger supplies of imported wine become available. Thus A's wine–cloth price ratio of 1:2 will fall. For comparable reasons, B's price ratio of 1:3 will rise. When the two ratios meet, at some intermediate level (in the example earlier, at 1:2½), the flow of goods will stabilize.

John Stuart Mill's view

Amplification of the theory. At a later stage in the history of comparative-advantage theory, the English philosopher and political economist John Stuart Mill showed that the determination of the exact after-trade price ratio was a supply-and-demand problem. At each possible intermediate ratio (within the range of 1:2 and 1:3), country A would want to import a particular quantity of wine and export a particular quantity of cloth. At that same possible ratio, country B would wish to import and export particular amounts of cloth and of wine. For any intermediate ratio taken at random, however, A's export-import quantities probably will not match those of B. Ordinarily, there will be just one intermediate ratio at which the quantities correspond; that is the final trading ratio at which quantities exchanged will stabilize. (Once they have stabilized, there is no longer any profit in exchanging goods. Even with such profits eliminated, however, there is no reason why A producers should want to stop selling part of their cloth in B, since the return there is as good as that obtained from domestic sales. Any falloff in the amounts exported and imported would reintroduce profit opportunities.)

Variations of the labour-cost model

In the elementary labour-cost example used above, there will be complete specialization: country A's entire labour force will move to cloth production and country B's to wine production. More elaborate comparative-advantage models recognize production costs other than labour (that is, the costs of land and of capital). In such models, part of country A's wine industry may survive and compete effectively against imports, as may also part of B's cloth industry. The models can be expanded in other ways: to take account of more than two countries, or more than two commodities, and of transport costs. The essential conclusions, however, come from the elementary model used above, so that this model, despite its simplicity, still provides a workable outline of the theory. (It should be noted that even the most elaborate comparative-advantage models continue to rely on certain simplifying assumptions without which the basic conclusions do not necessarily hold. These assumptions are discussed below.)

As noted earlier, the effect of this analysis is to correct any false first impression that low-productivity countries are at a hopeless disadvantage in trading with high-productivity ones. The impression is false, that is, if one assumes, as comparative advantage theory does, that international trade is an exchange of goods between countries. It is pointless for country A to sell goods to country B, whatever its labour-cost advantages, if there is nothing that it can profitably take back in exchange for its sales. With one exception, there will always be at least one commodity that a low-productivity country such as B can successfully export. Country B must of course pay a price for its low productivity, as compared with A; but that price is a lower per capita domestic income and not a disadvantage in international trading. For trading purposes, absolute productivity levels are unimportant; country B will always

find one or more commodities in which it enjoys a comparative advantage; that is, a commodity in the production of which its absolute disadvantage is least. The one exception is that case in which productivity ratios, and consequently pre-trade price ratios, happen to match one another in two countries. Such would have been the case had country B required four labour hours (instead of six) to produce a unit of cloth. In this particular circumstance, there would be no incentive for either country to engage in trade, and no gain from trading. In a two-commodity example such as that employed, it might not be unusual to find matching productivity and price ratios. But as soon as one moves on to cases of three and more commodities, the statistical probability of encountering precisely equal ratios becomes very small indeed.

Value of the theory of comparative advantage

The major purpose of the theory of comparative advantage is to illustrate the gains from international trade. Each country can gain by specializing in those occupations in which it is relatively efficient; it should export part of that production and take in exchange those goods in whose production it is, for whatever reason, at a comparative disadvantage. The theory of comparative advantage thus provides a strong argument for free trade—and indeed for a laissez-faire attitude with respect to trade. The supporting argument is simple: specialization and free exchange among nations yield higher real income for the participants.

The fact that a country will enjoy higher real income as a consequence of the opening up of trade does not mean, of course, that every family or individual within the country must share in that benefit. Producer groups affected by import competition obviously will suffer, to at least some degree. Comparative-advantage theorists concede that free trade would affect the relative income position of such groups, and perhaps even their absolute income level. But they insist that the special interest of these groups clashes with the total national interest, and the most that they are usually willing to concede is the possible need for temporary protection against import competition, in order that the persons affected may have sufficient time to move to another occupation.

Nations do, of course, maintain tariffs and other barriers to imports. For discussion of the reasons for this seeming clash between actual policies and the lessons of the theory of comparative advantage, see below *State interference in international trade.*

SOURCES OF COMPARATIVE ADVANTAGE

Theories of productivity variation

As already noted, British classical economists simply accepted the fact that productivity differences exist between countries; they made no concerted attempt to explain which commodities a country would export or import. During the 20th century, international economists have offered a number of theories in an effort to explain why countries have differences in productivity, the factor that determines comparative advantage and the pattern of international trade.

Natural resources. First, countries can have an advantage because they are richly endowed with a particular natural resource. For example, countries with plentiful oil resources can generally produce oil inexpensively. Thus, Saudi Arabia produces oil very cheaply, giving it a comparative advantage in oil, and it exports oil in order to finance its purchases of imports. Similarly, countries with large forests generally are the major exporters of wood, paper, and paper products. The supply available for export also depends on domestic demand. Canada, for example, has large quantities of lumber available for export to the United States, not only because of its large areas of forest but also because its small population consumes little of the supply, leaving much of the lumber available for export. Climate is another natural resource that provides an export advantage. Thus, for example, bananas are exported by Central American countries, not Iceland or Finland.

Factor endowments: the Heckscher-Ohlin theory. The proposition that countries with plentiful natural resources generally have a comparative advantage in products using those resources is obvious and straightforward. A related, but much more subtle, explanation was put forward by two Swedish economists, Eli Heckscher and Bertil Ohlin. Ohlin's work was built upon that of Heckscher. In recognition of his ideas as described in his path-breaking book, *Interregional and International Trade* (1933), Ohlin shared the Nobel Prize for Economics in 1977.

The work of Heckscher and Ohlin

The Heckscher-Ohlin theory focuses on the two most important factors of production, labour and capital. Some countries are relatively well-endowed with capital; the typical worker has plenty of machinery and equipment at his disposal. In such countries, wage rates generally are high. Products requiring much labour—such as textiles, sporting goods, and simple consumer electronics—tend as a result to be more expensive than in countries with plentiful labour and low wage rates. On the other hand, goods requiring much capital and only a little labour (automobiles and chemicals, for example) tend to be relatively inexpensive in countries with plentiful and cheap capital. Thus, countries with abundant capital should generally be able to produce capital-intensive goods relatively inexpensively, exporting them in order to pay for imports of labour-intensive goods.

In the Heckscher-Ohlin theory it is not the absolute amount of capital that is important; rather, it is the amount of capital per worker. A small country like Luxembourg has much less capital in total than India, but Luxembourg has more capital per worker. Accordingly, the Heckscher-Ohlin theory predicts that Luxembourg will export capital-intensive products to India and import labour-intensive products in return.

Despite its plausibility the Heckscher-Ohlin theory is frequently at variance with the actual patterns of international trade. As an explanation of what countries actually export and import, it is much less accurate than the more obvious and straightforward natural resource theory.

One early study of the Heckscher-Ohlin theory was carried out by Wassily Leontief, a Russian-born U.S. economist. Leontief observed that the United States was relatively well-endowed with capital. According to the theory, therefore, the United States should export capital-intensive goods and import labour-intensive ones. He found that the opposite was in fact the case: U.S. exports are generally more labour intensive than the type of products that the United States imports. Because his findings were the opposite of those predicted by the theory, they are known as the Leontief Paradox.

The Leontief Paradox

Economies of large-scale production. Even if countries have quite similar climates and factor endowments, they may still find it advantageous to trade. Indeed, economically similar countries often carry on a large and thriving trade. The prosperous industrialized countries have become one another's best customers. A main reason for this situation lies in what is called the economies of large-scale production.

For many products, there are advantages in producing on a large scale; costs become lower as more is produced. Thus, for example, automobiles can be made more cheaply in a factory producing 100,000 units than in a small factory producing only 1,000 units. This means that countries have an incentive to specialize in order to reduce costs. To sell a large volume of output, they may have to look to export markets.

The smaller the country, and the more limited its domestic market, the more incentive it has to look to international trade as a way of gaining the advantages of large-scale production. Thus, Luxembourg or Belgium has much more to gain relatively than does the United States. Indeed, the advantages of large-scale production have been one of the major sources of gain from the establishment of the European Economic Community (EEC), which was formed for the purpose of providing free trade between most western European countries.

Even a large country such as the United States, however, can gain in some cases by exporting in order to lengthen production lines. For example, the Boeing Company has been able to produce airplanes more cheaply because it is able to sell large numbers of aircraft to other countries. The importing countries also gain because they can buy aircraft abroad more cheaply than they could produce them at home.

Technology. Technological development can also provide a distinctive trade advantage. The relatively advanced countries—particularly the United States, Japan, and those of western Europe—are the principal exporters of high-technology products such as computers and precision machinery.

One important aspect of technology is that it can change rapidly. This is perhaps most obvious in the computer field, where productivity has increased, and costs have fallen sharply since the early 1960s. Such rapid changes present several challenges. For the countries that are not in the front rank, it raises the question of whether they should import high-technology products or attempt to enter the circle of the most advanced nations. For the countries that have held the technological lead in the past, there is always the possibility that they will be successfully overtaken by newcomers. This occurred in the second half of the 20th century when Japan advanced technologically in its automobile production to the point where it could challenge the automobile leadership of North America and Europe. Japan quickly became the world's foremost producer of automobiles.

Technological lead changes

The product cycle. The spread of technology across national boundaries means that comparative advantage can change. The most technologically advanced countries generally have the advantage in making new products, but as time passes other countries may gain the advantage. For example, many television sets were produced in the United States during the 1950s. As time passed, however, and technological change in the television industry became less rapid, there was less advantage in producing sets in the United States. Producers of television sets had an incentive to look to other locations, with lower wage rates. In time, the manufacturers established overseas operations in Taiwan, Hong Kong, and elsewhere. Concurrently, the United States turned to new activities, such as the manufacture of large mainframe computers and the development of computer software. (R.R./P.Wo.)

State interference in international trade

METHODS OF INTERFERENCE

Regardless of what comparative-advantage theory may say about the virtues of unrestricted trade, all nations interfere with international transactions to at least some degree. Tariffs may be imposed on imports—in some instances making them so costly as to bar completely the entry of the good involved. Quotas may limit the permissible volume of imports. State subsidies may be offered to encourage exports. Money-capital exports may be restricted or prohibited. Investment by foreigners in domestic plant and equipment may be similarly restrained.

These interferences may be simply the result of special-interest pleading, because particular groups suffer as a consequence of import competition. Or a government may impose restrictions because it feels impelled to take account of factors that comparative advantage sets aside. It is of interest to note that insofar as goods and services are concerned, the general pattern of interference follows the old mercantilist dictum of discouraging imports and encouraging exports.

A company that finds itself barred from an attractive foreign market by tariffs or quotas may be able to leapfrog the barrier simply by establishing a manufacturing plant within that foreign country. This policy of foreign plant investment has expanded enormously since the end of World War II. U.S. companies have taken the lead, investing particularly in western Europe, Canada, and South America. Industry in other developed countries has followed a similar pattern—some foreign companies establishing plants within the United States as well as in other areas of the world.

Foreign ownership of domestic industry

The governments of countries subject to this new investment find themselves in an ambivalent position. The establishment of new foreign-owned plants may mean more than simply the creation of new employment opportunities and new productive capacity; it may also mean the introduction of new technologies and superior business-control methods. But the government that welcomes such benefits must also expect complaints of "foreign control," an argument that will inevitably be pressed by domestic owners of older plants who fear a new competition that cannot be blocked by tariffs. This has been a pressing problem for many governments, particularly insofar as investment by U.S. firms is involved. Countries such as the United Kingdom and Canada have been liberal in their admissions policy; others, notably Japan, impose tight restrictions on foreign-owned plants. (R.R.)

Tariffs. A tariff, or duty, is a tax levied on a commodity when it crosses the boundary of a customs area. The boundary may be that of a nation or a group of nations that have agreed to impose a common tax on goods entering their territory. Tariffs are often classified as either protective or revenue. Protective tariffs are designed to shield domestic production from foreign competition by raising the price of the imported commodity. Revenue tariffs are designed to obtain revenue rather than to restrict imports. The two sets of objectives are, of course, not mutually exclusive. Protective tariffs, unless they are so high as to keep out imports, yield revenue. Revenue tariffs give some protection to any domestic producer of the duty-bearing goods. A transit duty, or transit tax, is a tax levied on commodities passing through a customs area en route to another country. Similarly, an export duty, or export tax, is a tax imposed on commodities leaving a customs area. Finally, some countries provide export subsidies; import subsidies are rarely used.

Protective and revenue tariffs

How tariffs work. Tariffs on imports may be applied in several ways. If they are imposed according to the physical quantity of an import (so much per ton, per yard, per item, etc.), they are called specific tariffs. If they are levied according to the value of the import, they are known as ad valorem tariffs.

Tariffs may differentiate among the countries from which the imports are obtained. They may, for instance, be lower between countries that have previously entered into special arrangements, such as the trade preferences accorded to each other by members of the Commonwealth.

Tariffs may affect the economy of the country imposing them in a number of ways. By raising the prices of imported goods, tariffs may encourage domestic production. As expenditures on domestic products rise, domestic employment tends to do likewise. This is why tariffs are favoured by industries that find themselves pressed by foreign competitors. The tariff may also encourage tendencies toward a monopolistic market structure to the extent that it lessens foreign competition, with a resulting decrease in the incentive to modernize or innovate. By increasing the price of an imported commodity a tariff may also reduce its consumption, depending on how consumers react to the higher price. The decrease in demand could be large enough in relation to the world market to force the price of the import down.

Economic effects of tariffs

Measuring the effects of tariffs. It is difficult to gauge the effect of tariff barriers among countries. First of all, how import demand responds to changes in tariffs depends on a variety of factors—the reaction of producers and consumers to price changes, the share of imports in domestic production and consumption, the substitutability of imports for domestic products, and so on. And the responsiveness to tariff levels will differ from country to country as well as from commodity to commodity. Thus, the amount of a tariff does not necessarily determine its restrictive effect. For another thing, such comparisons also must be restricted to commodities for which tariffs are the major protective device. This is generally true for nonagricultural commodities in developed countries, but other devices are often employed to protect agricultural production. In the third place, a tariff levied upon imports of raw materials will protect domestic raw material producers but will increase the costs of manufacturers using those raw materials. It is necessary, therefore, to distinguish between nominal and effective rates of protection.

The nominal rate of protection is the percentage tariff imposed on a product as it enters the country. For example, if a tariff of 20 percent of value is collected on clothing as it enters the country, then the nominal rate of protection is that same 20 percent.

Figuring the effective rate of protection

The effective rate of protection is a more complex concept: consider that the same product—clothing—costs $100 on international markets. The material that is imported to make the clothing (material inputs) sells for $60. In a free trade situation, a firm can charge no more than $100 for a similar piece of clothing (ignoring transportation costs). Importing the fabric for $60, the clothing manufacturer can add a maximum of $40 for labour, profit markup, rents, and the like. This $40 difference between the $60 cost of material inputs and the price of the product is called the value added.

The same situation may be considered with tariffs—say, 20 percent on clothing and 10 percent on fabric. The 20 percent tariff on clothing would raise the domestic price by $20 to $120, while a 10 percent tariff on fabrics would increase material costs to the domestic producer by $6 to $66. Protection would thus enable the firm to operate with a value-added margin of $54—the difference between the domestic price of $120 and the material cost of $66. The difference between the value added of $40 without tariff protection and that of $54 with it provides a margin of $14. This means that the effective rate of protection of the domestic processing activity—the ratio of $14 to $40—would be 35 percent. The effective rate of protection derived—35 percent—is greater than the nominal rate of only 20 percent. This will be the case whenever the tariff rate on the final product is greater than the tariff on inputs. Because countries generally do levy higher tariffs on final products than on inputs, effective rates of protection are generally higher than nominal rates—often much higher.

The effective rate of protection also depends on the share of value added in the product price. Effective rates can be very high if value added to the imported commodity is a small percentage or very low if value added is a large percentage of the total price. Thus, effective protection in one country may be much higher than that in another even though its nominal tariffs are lower, if it tends to import commodities of a high level of fabrication with correspondingly low ratios of value added to product price.

Nontariff barriers. Other government regulations and practices may also act as barriers to trade. Quotas or quantitative restrictions may prohibit the importation of certain commodities or limit the amounts imported. Such quotas are usually administered by requiring importers to have licenses to bring in particular commodities. Quotas raise prices just as tariffs do, but, being set in physical terms, their impact on imports is direct, with an absolute ceiling set on supply. Increased prices will not bring more goods in. There is also a difference between tariffs and quotas in their effect on revenues. With tariffs, the government receives the revenue: under quotas, the import license holders obtain a windfall in the form of the difference between the high domestic price and the low international price of the import.

"Voluntary" export restraints

Another barrier, which has become increasingly common during the past several decades, is the "voluntary" export restraint (VER). For example, in the 1980s the Japanese (under pressure from the United States) "voluntarily" limited their exports of automobiles to the U.S. market. Like quotas, VERs limit the quantity of trade and therefore tend to raise the prices of imported goods. In this case the VER made Japanese automobiles less available in the United States and raised the prices that U.S. consumers had to pay for such automobiles. VERs are usually not voluntary in any meaningful sense. In this example, the Japanese agreement to a VER on automobiles was an attempt to avoid a U.S. import quota. For the Japanese, a VER was preferable to a U.S. import quota because Japanese exporters could charge higher prices. The Japanese exporters, rather than U.S. importers, reaped much of the windfall from the VER. A VER is also easier to get rid of. In addition, a VER has a less damaging effect on the political relations between countries.

Still other barriers include state trading organizations and government procurement practices that may be used preferentially. In the United States, "buy American" legislation requires government procurement agencies to favour domestic goods. Customs classification and valuation procedures, health regulations, and marking requirements may also have a restrictive effect on trade. Finally, excise taxes may act as a barrier to trade if they are levied at higher rates on imports than on domestic goods.

Protectionism in the less-developed countries. The industrialization that has taken place in the late decades of the 20th century in some developing countries has been characterized by the expansion of import-competing industries behind high tariff walls. In many of those countries tariffs and various quantitative restrictions on manufactured goods are great. The effective rates of protection may be even greater, because the goods tend to be highly fabricated and the proportion of value added in production after importation is low. While some places such as Hong Kong, Taiwan, and Korea have oriented their manufacturing industries mainly toward export trade, those are exceptional cases in the developing world. More commonly the new industries seek to compete with foreign-made goods for the domestic market. High protection in these countries has often contributed to a slowdown in the production and export of primary commodities and has discouraged expansion of exports of manufactured goods. Furthermore, while domestic production of nondurable consumer goods has permitted rapid economic growth at an early stage, developing countries have encountered considerable difficulties in producing more sophisticated commodities. They suffer all of the disadvantages of small domestic markets, in addition to a lack of incentives for technological improvement.

(B.B./T.J.B./P.Wo.)

ARGUMENTS FOR AND AGAINST INTERFERENCE

Revenue. Developing nations in particular often lack the institutional machinery needed for effective imposition of income or corporation taxes. The governments of such nations may then finance their activity by resort to tariffs on imported goods, since such levies are relatively easy to administer. The amount of tax revenue obtainable through tariffs, however, is always limited. If the government tries to increase its tariff income by imposing higher duty rates, this may choke off the flow of imports and so reduce tariff revenue instead of increasing it.

Arguments on the protection of industry

Economic development. *Protection of domestic industry.* Probably the most common argument for tariff imposition is that particular domestic industries need tariff protection for survival. Comparative-advantage theorists will of course argue that the industry in need of such protection ought not to survive and that the resources so employed ought to be transferred to occupations having greater comparative efficiency. The welfare gain of citizens taken as a whole would more than offset the welfare loss of those groups affected by import competition; that is, total real national income would increase. An opposing argument would be, however, that this welfare gain would be widely diffused, so that the individual beneficiaries might not be conscious of any great improvement. The welfare loss, in contrast, would be narrowly and acutely felt. Although resources can be transferred to other occupations, just as comparative-advantage theory says, the transfer process is sometimes slow and painful for those being transferred. For such reasons, comparative-advantage theorists rarely advocate the immediate removal of all existing tariffs. They argue instead against further tariff increases—since increases, if effective, attract still more resources into the wrong occupation—and they press for gradual reduction of import barriers. (R.R.)

The infant-industry argument. Advocates of protection often argue that new and growing industries, particularly in less-developed countries, need to be shielded from foreign competition. They contend that costs decline with growth and that some industries must reach a minimum size before they are able to compete with well-established industries abroad. Tariffs can protect the domestic market until the industry becomes internationally competitive and, it is often argued, the costs of protection can be recouped after the industry has reached maturity. In short, the infant-industry argument is based principally on the idea that there are economies of large-scale production in many industries and that developing countries have difficulty in establishing such industries.

Weaknesses of the infant-industry theory

Advocates of such protection, however, can have their arguments turned against them. While an individual country can, in some circumstances, gain from protecting its infant industries, this protection is particularly costly for the international community as a whole. Where there are major advantages in large-scale production, there are also large advantages in relatively free international trade. By closing off markets, protection reduces the ability of firms to gain economies of large-scale by exporting. If a group of countries imposes infant-industry protection, it will split up the market; each country may end up with small-scale, localized, inefficient production, thus reducing the prosperity of all of the countries. One way in which developing nations have tried to deal with this problem has been through the establishment of customs unions or other regional groupings (see below *International trade arrangements*).

Infant-industry tariffs have been disappointing in other ways; the infant-industry argument is often abused in practice. In many developing countries, industries have failed to attain international competitiveness even after 15 or 20 years of operation and might not survive if protective tariffs were removed. The infant industry is probably better aided by production subsidies than by tariffs. Production subsidies do not raise prices and therefore do not curtail domestic demand; and the cost of the protection is not concealed in higher prices to consumers. Production subsidies, however, have the disadvantage of drawing upon government revenue rather than adding to it, which may be a serious consideration in countries at lower levels of development. (B.B./T.J.B./P.Wo.)

Unemployment. Tariffs or quotas are also sometimes proposed as a way to maintain domestic employment—particularly in times of recession. There is, however, near-unanimity among modern-day economists that proposals to remedy unemployment by means of tariff increases are misguided. Insofar as a higher tariff is effective for this purpose, it simply "exports unemployment"; the rise in domestic employment is matched by a drop in production in some foreign country. That other country, moreover, is likely to impose a retaliatory tariff increase. Finally, the tariff remedy for unemployment is a poor one because it is usually ineffective and because more suitable remedies are now available. It has come to be generally recognized that unemployment is far more efficiently dealt with by the implementation of proper fiscal and monetary policies.

National defense. A common appeal made by an industry seeking tariff or quota protection is that its survival is essential in the national interest: its product would be needed in wartime, when the supply of imports might well be cut off. The verdict of economists on this argument is fairly clear: the national-defense argument is frequently a red herring, an attempt to "wrap oneself in the flag," and insofar as an industry is essential, the tariff is a dubious means of ensuring its survival. Essential industries ought instead to be given a direct subsidy to enable them to meet foreign competition, with explicit recognition of the fact that the subsidy is a price paid by the nation in order to maintain the industry for defense purposes.

Autarky, or self-sufficiency. Many demands for protection, whatever their surface argument may be, are really appeals to the autarkic feelings that prompted mercantilist reasoning. (Autarky is defined as the state of being self-sufficient at the level of the nation.) A proposal for the restriction of free international trade can be described as autarkic if it appeals to those half-submerged feelings that the citizens of the nation share a common welfare and common interests, whereas foreigners have no regard for such welfare and interests and might even be actively opposed to them. And it is quite true that a country that has become heavily involved in international trade has given hostages to fortune: a part of its industry has become dependent upon export markets for income and for employment. Any cutoff of these foreign markets (brought about by recession abroad, by the imposition of new tariffs by some foreign country, or by numerous other possible changes) would be acutely serious; and yet it would be a situation largely beyond the power of the domestic government involved to alter. Similarly, another part of domestic industry may rely on an inflow of imported raw materials, such as oil for fuel and power. Any restriction of these imports could have the most serious consequences. The vague threat implicit in such possibilities often results in a yearning for autarky, for national self-sufficiency, for a life free of dependence on the hazards of the outside world.

Self-sufficiency as a basis for national welfare

There is general agreement that no modern nation, regardless of how rich and varied its resources, could really practice self-sufficiency, and attempts in that direction could produce sharp drops in real income. Nevertheless, protectionist arguments—particularly those made "in the interests of national defense"—often draw heavily on the strength of such autarkic sentiments.

The terms-of-trade argument. When a country imposes a tariff, foreign exporters have greater difficulty in selling their products. As their exports decline, they may cut prices in order to keep their sales from falling drastically. Thus, for example, when a tariff of $10.00 is imposed, foreign exporters may cut their price by, say, $6.00. The foreign exporter is being "taxed" when the tariff is imposed; the other $4.00 is reflected in a higher price to the consumer. The use of tariffs to tax foreign exporters in this way is known as the terms-of-trade argument for protection. The terms of trade represent the relative price of what a nation is exporting, compared with the price paid to foreigners for imported goods. When the price of what is being exported rises, or when the price paid to foreigners for imported goods falls (as it may when a nation imposes a tariff), terms of trade improve. (Terms of trade is discussed below under *International commodity trade.*)

Balance-of-payments difficulties. Governments may interfere with the processes of foreign trade for a reason quite different from those thus far discussed: shortage of foreign exchange. Under the international monetary system established after World War II and in effect until the 1970s, most governments tried to maintain fixed exchange rates between their own currencies and those of other countries. Even if not absolutely fixed, the exchange rate was ordinarily allowed to fluctuate only within a narrow range of values.

If balance-of-payments difficulties arise and persist, a nation's foreign exchange reserve runs low. In a crisis, the government may be forced to devalue the nation's currency. But before being driven to this, it may try to redress the balance by restricting imports or encouraging exports, in much the old mercantilist fashion.

The problem of reserve shortages became acute for many countries during the 1960s. Although the total volume of international transactions had risen steadily, there was not a corresponding increase in the supply of international reserves. By 1973 payment imbalances led to an end of the system of fixed, or pegged, exchange rates and to a "floating" of most currencies. (Balance of payments is discussed in detail below under *International payments and exchange.*) (R.R./P.Wo.)

Floating currencies

International trade arrangements

Many efforts have been made in modern times to promote trade among nations. The ways in which this may be attempted range from agreements among governments to reduce or eliminate trade barriers to more ambitious attempts to harmonize economic policies, as in the European Economic Community (EEC) established by the nations of western Europe.

HISTORY OF MODERN TRADE POLICIES

Mercantilism. Much of the modern history of international relations concerns efforts to promote freer trade among nations. The 17th century saw the growth of restrictive policies that later came to be known as mercantilism. The mercantilists held that economic policy should be nationalistic and aim to secure the wealth and power of the state. The concept was based on the conviction that national interests are inevitably in conflict—that one nation can increase its trade only at the expense of other nations. Thus, governments were led to impose price and wage controls, foster national industries, promote exports of finished goods and imports of raw materials, and pro-

hibit the exports of raw materials and the import of finished goods. The state endeavoured to provide its citizens with a monopoly of the resources and trade outlets of its colonies.

A typical illustration of the mercantilist spirit is the famous English Navigation Act of 1651, which reserved for the home country the right to trade with the colonies and prohibited the import of goods of non-European origin unless transported in ships flying the English flag. This law lingered on until 1849. A similar policy was followed in France.

The growth of free trade

Liberalism. A strong reaction against mercantilist attitudes began to take shape toward the middle of the 18th century. In France, the economists known as Physiocrats demanded liberty of production and trade. In England, as discussed above, Adam Smith demonstrated in his *The Wealth of Nations* (1776) the advantages of removing trade restrictions. Economists and businessmen voiced their opposition to excessively high and often prohibitive customs duties and urged the negotiation of trade agreements with foreign powers. This change in attitudes led to the signing of a number of agreements embodying the new liberal ideas, among them the Anglo-French Treaty of 1786, which ended what had been an economic war between the two countries.

After Adam Smith, the basic tenets of mercantilism were no longer considered defensible. This did not, however, mean that nations abandoned all mercantilist policies. Restrictive economic policies were now justified by the claim that, up to a certain point, the government should keep foreign merchandise off the domestic market in order to shelter national production from outside competition. To this end, customs levies were introduced in increasing number, replacing outright bans on imports, which became less and less frequent.

In the middle of the 19th century, customs walls effectively sheltered many national economies from outside competition. The French tariff of 1860, for example, charged extremely high rates on British products: 60 percent on pig iron; 40 to 50 percent on machinery; and 600 to 800 percent on woolen blankets. Transport costs between the two countries provided further protection.

A triumph for liberal ideas was the Anglo-French trade agreement of 1860, which provided that French protective duties were to be reduced to a maximum of 25 percent within five years, with free entry of all French products except wines into Britain. This agreement was followed by other European trade pacts.

Resurgence of protectionism. A reaction in favour of protection spread throughout the Western world in the latter part of the 19th century. Germany adopted a systematically protectionist policy and was soon followed by most other nations. Shortly after 1860, during the Civil War, the United States raised its duties sharply; the McKinley Tariff Act of 1890 was ultraprotectionist. England was the only country to remain faithful to the principles of free trade.

The decline of free trade

But the protectionism of the last quarter of the 19th century was mild by comparison with the mercantilist policies that had been common in the 17th century and were to be revived between the two world wars. Extensive economic liberty prevailed by 1913. Quantitative restrictions were unheard of, and customs duties were low and stable. Currencies were freely convertible into gold, which in effect was a common international money. Balance-of-payments problems were few. People who wished to settle and work in a country could go where they wished with few restrictions; they could open businesses, enter trade, or export capital freely. Equal opportunity to compete was the general rule, the sole exception being the existence of limited customs preferences between certain countries, most usually between a home country and its colonies. Trade was freer throughout the Western world in 1913 than it was in Europe in 1970.

Post-World War I trends

The "new" mercantilism. World War I wrought havoc with these orderly trading conditions. By the end of hostilities, world trade was in a straitjacket that made recovery very difficult. The first five years of the postwar period were marked by the dismantling of wartime controls proper. The 1920 crisis and the commercial advantages accruing to countries whose currencies had depreciated, as had Germany's, rapidly led to fresh measures in restraint of trade. The protectionist tide engulfed the world economy, not because policymakers consciously adhered to any specific theory but because of nationalist ideologies and the pressure of economic conditions. In an attempt to end the continual raising of customs barriers, the League of Nations organized the first World Economic Conference in May 1927. Twenty-nine states, including the main industrial countries, subscribed to an international convention that was the most minutely detailed and balanced multilateral trade agreement ever approved until that time. It was a precursor of the arrangements made under the General Agreement on Tariffs and Trade of 1947.

The 1927 agreement remained practically without effect. During the Great Depression of the 1930s, unemployment in major countries reached unprecedented levels and engendered an epidemic of protectionist measures. Countries attempted to shore up their balance of payments by raising their customs duties and introducing a range of import quotas or even import prohibitions, accompanied by exchange controls.

From 1933 onward, the recommendations of all the postwar economic conferences based on the fundamental postulates of economic liberalism were ignored. The planning of foreign trade came to be considered a normal function of the state. Mercantilist policies dominated the world scene until after World War II.

TRADE AGREEMENTS

The term trade agreement or commercial agreement can be used to describe any contractual arrangement between states concerning their trade relationships. Trade agreements may be bilateral or multilateral—that is, between two states or between more than two states.

Provisions of bilateral agreements

Bilateral trade agreements. A bilateral trade agreement usually includes a broad range of provisions regulating the conditions of trade between the contracting parties. These include stipulations governing customs duties and other levies on imports and exports, commercial and fiscal regulations, transit arrangements for merchandise, customs valuation bases, administrative formalities, quotas, and various legal provisions. Most bilateral trade agreements, either explicitly or implicitly, provide for (1) reciprocity, (2) most-favoured-nation treatment, and (3) "national treatment" of nontariff restrictions on trade.

Reciprocity. In a trade agreement, the parties make reciprocal concessions to put their trade relationships on a basis deemed equitable by each. The principle of reciprocity is extremely old, and in one form or another it is to be found, implicitly at least, in all trade agreements. The concessions may, however, be in different areas. In the Anglo-French Agreement of 1860, for example, France pledged itself to reduce its duties to 20 percent by 1864. In return, Britain granted duty-free imports of all French products except wines and spirits. The principle of reciprocity implies only that the gains arising out of foreign trade are distributed fairly.

The most-favoured-nation clause. The most-favoured-nation clause binds a country to apply to its partner country any lower rate of import duties that it may later grant to imports from some other country. The clause may cover a list of specified products only, or specific concessions yielded to certain foreign countries. Alternatively, it may cover all advantages, privileges, immunities, or other favourable treatment granted to any third country whatever. The clause is intended to provide each signatory with the assurance that the advantages obtained will not be attenuated or wiped out by a subsequent agreement concluded between one of the partners and a third country. It guarantees the parties against discriminatory treatment in favour of a competitor.

Protection against discriminating tariffs

The effect of the most-favoured-nation clause on customs duties is to amalgamate the successive trade agreements concluded by a state. If the rates in different agreements are fixed at varying levels, the clause reduces them to the lowest rate specified in any agreement. Thus, goods imported from a country benefiting from most-favoured-

nation treatment are charged the rate of duty applicable to imports from another country which, in a subsequent trade agreement, has negotiated a lower rate of duty.

The coverage of the most-favoured-nation clause can be considerably reduced by a minute definition of a particular item so that a concession, while general in form, applies in practice only to one country. The best-known illustration of this technique is to be found in the German Tariff of 1902, which admitted at a special rate

> large dappled mountain cattle, reared at a spot at least 300 metres above sea level, and which have at least one month's grazing each year at a spot at least 800 metres above sea level.

The advantages granted under the most-favoured-nation clause may be conditional or unconditional. If unconditional, the clause operates automatically whenever appropriate circumstances arise. The country drawing benefit from it is not called on to make any fresh concession. By contrast, the partner invoking a conditional most-favoured-nation clause must make concessions equivalent to those extended by the third country. A typical wording was that of the 1911 treaty between the United States and Japan, which stated that

> in all that concerns commerce and navigation, any privilege, favour or immunity . . . to the citizens or subjects of any other State shall be extended to the citizens or subjects of the other Contracting Party gratuitously, if the concession in favour of that other State shall have been gratuitous, and on the same or equivalent conditions, if the concession shall have been conditional.

The conditional form of the clause may at first sight seem more equitable. But it has the major drawback of being liable to raise dispute each time it is invoked, for it is by no means easy for a country to evaluate the compensation it is being offered as in fact being equivalent to the concession made by the third country.

The effect of the unconditional form of the most-favoured-nation clause is, finally, to wipe out any relevance the principle of reciprocity may have had to the purely bilateral preoccupations of the negotiating parties, since the results of the bargaining process, instead of being limited to the participants, influence their relationships with other states. In practice, therefore, a country negotiating a trade agreement must measure the advantages it is willing to concede in terms of the benefits these concessions will provide collaterally to that third country which is the most competitive. In other words, the concessions that may be granted are determined by the minimum protection that the negotiating state deems indispensable to protect its home producers. This sets a major limitation on the scope of bilateral negotiations.

Determining the advantage of trade agreements

Protagonists of free trade consider that the unconditional most-favoured-nation clause is the only practical way by which to obtain the progressive reduction of customs duties. Apologists for protectionism are resolutely against it, preferring the conditional form of the clause or some equivalent mechanism.

The conditional most-favoured-nation clause was generally in use in Europe until 1860, when the so-called Cobden–Chevalier Treaty between Great Britain and France established the unconditional form as the pattern for most European treaties. The United States used the conditional most-favoured-nation clause from its first trade agreement, signed with France in 1778, until the passage of the Tariff Act of 1922, which terminated the practice. (The Trade Reform Bill of 1974, however, in effect restored to the U.S. president the authority to designate preferential tariff treatment, subject to approval by Congress.)

The Genoa conference in May 1922 and the World Economic Conference in May 1927 both recommended that trade agreements should include the most-favoured-nation clause whenever possible. But the Depression of the 1930s led instead to a rise of restrictions in world trade. Imperial or regional systems of preference came into being: the Ottawa Agreement of 1932 for the British Commonwealth; similar arrangements for the French empire; and a series of tariff and preference agreements negotiated in eastern and central Europe from 1931 on.

The "national treatment" clause. The "national treatment" clause in trade agreements is designed to ensure that internal fiscal or administrative regulations are not used to introduce discrimination of a nontariff nature. It forbids discriminatory use of the following: taxes or other internal levies; laws, regulations, and decrees affecting the sale, offer for sale, purchase, transport, distribution, or use of products on the domestic market; valuation of products for purposes of assessment of duty; legislation on prices of imported goods; warehousing and transit regulations; and the organization and operation of state trading corporations.

Multilateral agreements since World War II. When World War II ended, the lessons learned from the growth of protectionism since 1871, and most of all from the resurgence of trade restrictions in the interwar years, spurred the development of multilateral trade agreements and other forms of international economic cooperation. These developments culminated in the General Agreement on Tariffs and Trade (GATT).

Attempts at trade liberalization

The General Agreement on Tariffs and Trade. The General Agreement on Tariffs and Trade was signed at Geneva on Oct. 30, 1947, by 23 countries, which among them accounted for four-fifths of world trade. On the same day 10 of them, including the United States, the United Kingdom, France, Belgium, and The Netherlands, signed a protocol bringing the agreement into force on Jan. 1, 1948.

GATT takes the form of a multilateral trade agreement setting forth the principles under which the signatories, on a basis of "reciprocity and mutual advantage," shall negotiate "a substantial reduction in customs tariffs and other impediments to trade, and the elimination of discriminatory practices in international trade." With the adherence of additional countries GATT has become a charter governing almost all world trade except for that of the Communist countries.

Principles of GATT

The main principles underlying GATT are as follows: (1) There shall be no trade discrimination. The unconditional most-favoured-nation clause is regarded as fundamental. (2) As a rule, there is to be no protection other than that provided by the customs tariff (the "national treatment" principle). (3) Customs unions and free trade groupings are considered legitimate means of trade liberalization, provided that, taken as a whole, such arrangements do not discriminate against third countries. (4) Members of GATT are entitled to levy the following charges on imports: (*a*) an import tax equal in amount to internal taxes on the product concerned, subject only to the general principle embodied in (2) above; (*b*) "antidumping" duties in the case of imported products that are being sold at less than the price in the home market, or at less than cost; (*c*) countervailing duties to offset the effects of export subsidies; (*d*) fees and other proper charges for services rendered.

These, however, are only the basic principles. The agreement also contains a variety of clauses providing exceptions from the rules in special situations. These include balance-of-payments disequilibrium; serious and unexpected damage to domestic production; the requirements of economic development or, subject to very broad reservations, of agricultural policy; the need to protect domestic raw material production; and the interests of national security. In addition the GATT rules permit countries entering a customs union to depart from the most-favoured-nation principle. For example, within the European Economic Community, France can permit duty-free entry of goods from its fellow members—such as West Germany and Italy—without extending such duty-free treatment to the products of outside nations.

Periodically, major multilateral trade conferences, called rounds, are held by GATT countries to work out trade problems. Most of these have been held at Geneva, site of GATT headquarters. The formula for multilateral tariff bargaining applied in negotiations held under GATT auspices is a major innovation in intergovernmental cooperation. In appraising the concessions that they could afford to make, governments have been able to take account of the indirect advantages that they could expect to accrue to them from the full set of bilateral negotiations. Over the years since GATT's inception, it has been successful far

beyond its creators' expectations; its contribution to the growth of world trade has been major. Two GATT sessions of particular historic importance were the Kennedy Round and the Tokyo Round.

The Kennedy Round

As the economic integration of western Europe progressed, opinion in the United States became concerned at the prospect of remaining outside. Pres. John F. Kennedy pursued the goal of an Atlantic partnership and secured special negotiating powers under the Trade Expansion Act of 1962. The act authorized tariff reductions of up to 50 percent, subject to reciprocal concessions from the European partners. This marked a fundamental shift away from the traditional protectionist posture of the United States and led to the so-called Kennedy Round negotiations in the GATT, held in Geneva from May 1964 to June 1967.

The Kennedy Round negotiations concerned four types of problems: (1) progressive reduction, to amount finally to 50 percent, in the duties on all but a few products, in place of the item-by-item bargaining that had prevailed in earlier GATT conferences; (2) inclusion of agricultural as well as industrial products in the scope of the negotiations; (3) discussion of nontariff obstacles as well as of customs duties; and (4) nonreciprocity for economically less-developed countries. Fifty-four countries participated in the negotiations, which covered 400,000 tariff headings.

The final result was an average reduction of 35 percent in the duties levied on industrial products, to be implemented over a five-year period. This was less than the 50 percent originally envisaged. Further, the reductions were not geographically uniform: U.S., European, and Japanese duties were to fall by an average of 35 percent, British duties by 38 percent, and Canadian by 24 percent. Little change was made in steel and textile tariffs, since the participants felt that reductions in those industries would create intolerable political and social tensions in most of the industrial countries. Problems arose with regard to chemicals because of a so-called American Selling Price that was used for appraising the dutiable value of some products, based on prices ruling in the U.S. market; in return for a reduction of 50 percent in the rates of duty charged by the United States, Great Britain and the European countries agreed to lower their duties by 22 percent, with a further 24 percent reduction to become effective upon abrogation of the American Selling Price. Rather less spectacular results were achieved for agricultural commodities. These included the setting of a minimum price for wheat and a weighted reduction of between 15 and 18 percent in the duties charged on other agricultural and food products. In the area of nontariff barriers to trade the most significant result was the adoption of a uniform antidumping code.

The American Selling Price

The Kennedy Round continued the process of tariff reduction begun two decades earlier by the industrial countries. While developing countries drew little immediate advantage from the Kennedy Round negotiations, they were able to obtain the addition of a new part titled "Trade and Development" to the GATT charter, calling for stabilization, as far as possible, of raw material prices; reduction or abolition of customs duties or other restrictions that differentiate unreasonably between products in their primary state and the same products in finished form; and renunciation by the advanced countries of the principle of reciprocity in their relations with less-developed countries. (Ma.A./P.Wo.)

The next ministerial meeting of GATT opened in Tokyo on Sept. 12, 1973, and was attended by representatives of ministerial or comparable level from 102 countries. On September 14 the meeting closed with the adoption of what came to be called the Tokyo Declaration.

The declaration differed markedly from previous GATT documents in the inordinately large portion of its language devoted to strengthening the negotiating position of the less-developed countries. One of the general aims of the negotiations was said to be the securing of additional trade benefits for less-developed countries. Specifically, the trade negotiations would aim at improving the conditions of access for products of interest to such countries and ensuring stable, equitable, and remunerative prices for primary products. Tropical products would be given special and priority treatment. The principle of nonreciprocity in negotiations between developed and less-developed countries, an established principle in GATT, was reaffirmed: the importance of maintaining and improving the Generalized System of Preferences granted by developed countries to less-developed countries, as well as the need for special measures and the importance of providing special, differential, and more favourable treatment for less-developed countries, were recognized. Special attention was to be given to the trade interests of the least-developed countries.

Trade benefits for less-developed countries

The Tokyo Round

The Tokyo Declaration was followed by several years of multinational trade negotiations that came to be called the Tokyo Round, concluding in 1979 with the adoption of a series of tariff reductions to be implemented generally over an eight-year period beginning in 1980. Further progress was also made in dealing with nontariff issues. Most notably, a Code on Subsidies and Countervailing Duties was negotiated. This code had two main features: it listed a number of unacceptable subsidy practices, and it introduced a requirement that formal procedures be followed before the imposition of countervailing duties on imports subsidized by foreign nations. Specifically, before the imposition of a countervailing duty, an investigation had to establish that competing domestic firms were being injured. The code was not signed by all of the members, however, and the signing nations agreed only to follow the prescribed rules before applying countervailing duties to the exports of other signatories. Thus, while the code represented progress in dealing with a new topic, it also represented a departure from the most-favoured-nation principle: signatories were not required to extend the benefits of the code to GATT members who did not sign the code.

A new set of negotiations was initiated at a conference in Uruguay in 1986. Because traditional tariffs were becoming much less important, most of the attention was focused on other impediments to international transactions. The United States was particularly eager that the new negotiations include international transactions in services and an attack on the problems of international agricultural trade, which had been severely distorted by domestic price support programs and subsidized exports of surplus production. (P.Wo./Ed.)

The stimulus of the European Recovery Program

The Organisation for Economic Co-operation and Development. On April 16, 1948, 16 European countries responded to a U.S. offer of economic aid under the European Recovery Program by setting up the Organisation for European Economic Co-operation (OEEC). Although the immediate aim was to coordinate the distribution of U.S. credits, the OEEC convention was also designed to foster free trade between the members and allow their participation in customs unions or similar institutions. The members by 1955 consisted of Britain, France, West Germany, Italy, Spain, the Benelux countries, Austria, Denmark, Sweden, Norway, Switzerland, Portugal, Greece, Ireland, Turkey, and Iceland.

The Organisation for European Economic Co-operation did much to facilitate the recovery of intra-European trade and particularly to abolish most of the quantitative restrictions on imports within the area. On Sept. 30, 1961, it was converted into a new institution, the Organisation for Economic Co-operation and Development (OECD), and its membership expanded to include the United States and Canada. Japan became a member in 1964, Finland joined in 1969, Australia in 1971, and New Zealand in 1973.

The three fundamental aims of the OECD are to promote the economic growth of member countries in conditions of financial stability, to contribute to the economic growth of less-developed countries, and to foster the growth of world trade on a multilateral, nondiscriminatory basis. Having little power to enforce its decisions, the OECD has served mostly as a consultative body, influencing trade through its studies and communications.

ECONOMIC INTEGRATION

The economic integration of several countries or states may take a variety of forms. The term covers preferential tariffs, free-trade associations, customs unions, common

Forms of economic integration

markets, economic unions, and full economic integration. The parties to a system of preferential tariffs levy lower rates of duty on imports from one another than they do on imports from third countries. For example, Great Britain and its Commonwealth countries operated a system of reciprocal tariff preferences after 1919. In free-trade associations no duty is levied on imports from other member states, but different rates of duty may be charged by each member on its imports from the rest of the world. The European Free Trade Association is an example. A further stage is the customs union, in which free trade among the members is sheltered behind a unified schedule of customs duties charged on imports from the rest of the world. The 19th-century German Zollverein (see below) was a customs union. A common market is an extension of the customs union concept, the additional feature being that it provides for the free movement of labour and capital among the members; an example was the Benelux common market until it was converted into an economic union in 1959. The term economic union denotes a common market in which the members agree to harmonize their economic policies generally, as is the case with the European Economic Community (often referred to as the Common Market). Finally, total economic integration implies the pursuit of a common economic policy by the political units involved; examples are the states of the United States or the cantons of the Swiss confederation.

Economic integration may be brought about by the political will of a state powerful enough to impose it, as under the Roman Empire or the European colonial systems of the 19th century, or it may result from freely negotiated agreement between sovereign states, as has been more common in the 20th century.

The attempts at economic integration made after World War II can be appraised only by reviewing them against the background of the long process through which, over the centuries, the nations of the world have progressively achieved economic integration. Thus, for instance, the world's greatest power in the 17th century, France, was divided into a number of provinces separated from one another by various customs barriers involving a multitude of duties, tolls, and prohibitions. Trade regulations and fiscal charges differed from one region to the next; there was not even a single system of weights and measures. Not until after the Revolution did the economic integration of France really get under way.

Integration between the states of federations. *The United States.* The economic integration of the United States was not achieved all at once, but as the result of a long process during which the powers of the federal authorities were constantly reinforced. The Constitution empowered the federal government to regulate the conditions of trade with other countries and to set up a single system of duties. It also abolished the right of individual states to maintain separate customs legislation and to issue their own currencies. It authorized the federal government alone to issue currency and established the principle of free movement of persons, merchandise, and capital between the federated states. But the conflict of interest between North and South was settled only by the American Civil War. Almost 200 years were to elapse before the economies of the states could be considered as integrated for practical purposes, and even today many economic and fiscal disparities still exist among them.

Constitutional basis for trade regulation

The difficulties faced by the 13 original states should not be underestimated. During the years prior to the adoption of the Constitution there were bitter trade disputes among the states, which imposed tariffs against each other and refused to accept each other's currencies. Everything seemed to justify the words of a contemporary liberal philosopher, Josiah Tucker, Dean of Gloucester (England):

> As to the future grandeur of America, and its being a rising empire under one head, whether republican or monarchical, it is one of the idlest and most visionary notions that ever was conceived even by writers of romance. The mutual antipathies and clashing interests of the Americans, their differences of governments, habitudes, and manners, indicate that they will have no centre of union and no common interest. They never can be united into one compact empire under any species of government whatever; a disunited people till the end of time, suspicious and distrustful of each other, they will be divided and sub-divided into little commonwealths or principalities, according to natural boundaries, by great bays of the sea, and by vast rivers, lakes, and ridges of mountains.

Switzerland. The Swiss example is no less instructive. Although the Helvetic Confederation emerged as a political entity in the 14th century, its economic integration was achieved, only after many vicissitudes, with the constitution of 1848. The terms of this document established a common currency, set forth the principle of a common protective system for the cantons, and provided for free movement of goods and Swiss citizens throughout the national territory. Swiss economic integration is all the more remarkable in that it comprises peoples who speak four different languages.

The economic integration of colonial empires. When the great colonial powers of Europe founded their empires from the 16th century onward, they attempted to monopolize trade with the colonies and to turn it to their own profit. This policy involved four main restrictions: (1) The colonies were to trade exclusively with the mother country. (2) They were not to undertake manufacturing; transformation of raw materials into finished goods remained a monopoly right of the mother country. (3) Imports and exports of the colonies were to be carried only in ships flying the mother country's flag. (4) The mother country exempted colonial products from duty, or imposed lower rates.

This system, although progressively attenuated, applied in various forms from the 16th to the 19th century. Based on force, it was to the benefit of the home countries and detrimental to the economic growth of their colonies.

The German customs union

The Zollverein. The best known of the early customs unions is the German Zollverein (literally, "customs union"). Even though Napoleon had reduced the number of German states from 300 to 40 at the beginning of the 19th century, those that remained were isolated from each other by their own customs systems. In addition, numerous internal customs barriers hampered trade within each state. At the same time there was no single external tariff, and the German industries that had sprung up during the Napoleonic Wars were being crushed by English competition. These difficulties were at the root of the creation of the Zollverein.

The starting point was Prussia's abolition of all internal duties and its adoption of an external tariff in 1818. In the next few years a number of other German states followed the Prussian example. Bavaria and Württemberg set up a customs union in 1828, and by 1830 four separate customs unions were in existence. Prussia then sought to break up the local customs unions and attach them to a general customs union, the Zollverein. The coverage of the Zollverein increased until, by 1871, it included all the German states.

In its first phase, from 1834 to 1867, the Zollverein was administered by a central authority, the Customs Congress, in which each state had a single vote. A common tariff, the Prussian Tariff of 1818, shielded the member states from foreign competition, but free trade was the rule internally.

During a second phase, from 1867 to 1871 (following Prussia's victory over Austria at Sadowa), executive power was wielded by a federal council (Bundesrat) composed of governmental delegates, in which decisions were taken by an absolute majority. Prussia was entitled to 17 of the 58 votes and held the chair of the council. Legislative power lay with a "customs parliament" (*Zollparlament*) composed of deputies directly elected by popular vote, and, like the council, taking decisions by a majority vote. This arrangement transformed what had been a confederation into a federal state.

After the victory over France and the proclamation of the German empire in 1871, the customs parliament and the federal council were replaced by the parliament and the executive council of the empire. The federal state had become a nation.

The progressive destruction of a tangled maze of regulations, prohibitions, and controls set the stage for the

subsequent rapid development of the German economy. Although economic integration occurred before political unification, it would not have been possible had not many difficulties been swept away by irresistible pressure from Prussia with its military victories.

Belgium–Luxembourg Economic Union

The Benelux Economic Union. In 1921 Luxembourg, a former member of the Zollverein, signed the Convention of Brussels with Belgium, creating the Belgium–Luxembourg Economic Union. Since 1921 Belgium and Luxembourg have, then, had the same customs tariff and a single balance of payments.

The union was expanded after World War II to include The Netherlands. At the beginning of 1948 most import duties within the Benelux area were abolished, and a common external tariff was put into operation. Exceptions were made, nevertheless, for a few agricultural products, and it was also felt necessary to introduce a system of quotas.

It was rapidly perceived that a simple customs union was inadequate, and a treaty on Oct. 15, 1949, set as its target the progressive and complete liberalization of trade between the partners, systematic coordination of their international commercial and monetary policies, and the adoption of a joint bargaining position in negotiations with other countries. Though the experiment was optimistically viewed everywhere as the precursor of a wider European economic integration, it faced difficulties arising from the very different postwar situations of Belgium and The Netherlands. The two economies were competitive rather than complementary. Other problems arose in connection with the free access of Dutch agricultural products to the Belgian market. Moreover, the Belgian economic system was more liberal than the Dutch, where rigorous price control had long been a standard practice.

The development of Benelux received strong impetus from the formation of the European Economic Community in the 1950s. The Treaty of Rome in 1957 creating the EEC, or Common Market, spurred the members of Benelux to confirm and strengthen their own integration in the Benelux Treaty of Economic Union signed at The Hague on Feb. 3, 1958. The Hague treaty, however, contained little that was new, and in outline it was no more than the codification of results already achieved.

The Benelux Economic Union makes all of its decisions unanimously. The union has executive organs (the Committee of Ministers, the Council of the Economic Union, a number of commissions, and a Secretariat General); consultative organs (the Inter-Parliamentary Council and the Economic and Social Advisory Council); and legal organs (the Court of Justice).

Efforts toward further integration in Europe

The European Coal and Steel Community. An important step in European integration was taken in May 1950 when the French foreign minister, Robert Schuman, proposed that a common market for coal and steel be set up by countries willing to delegate powers over these sectors of their economies to an independent authority. The motive behind the plan was the belief that a new economic and political framework was needed if European unity was to be achieved and if the threat of a future Franco-German conflict was to be avoided. In April 1951 France, West Germany, Italy, and the three Benelux countries signed a treaty in Paris setting up the European Coal and Steel Community (ECSC).

The signatories bound themselves to abolish all customs barriers and other restrictions on the movement of coal and steel between their countries; to renounce all discriminatory practices among producers, purchasers, or users (with respect to price and delivery conditions, transport charges, selection of suppliers, etc.); to end government subsidies or grants-in-aid; and to eliminate all practices interfering with the operation of markets.

The constitution of the community. When first promulgated, the constitution of the Coal and Steel Community allowed that it be governed by a High Authority, assisted by a Consultative Committee, a Common Assembly, a Special Council of Ministers, and a Court of Justice.

The High Authority was the permanent executive organ of the community. Its decisions, taken by a majority vote, were fully binding on all member countries, each of which was pledged to respect the "supranational character" of the High Authority. The authority was to refer important substantive matters to the Consultative Committee before taking a decision. The latter was composed of coal and steel industry representatives, including producers, workers, users, and traders.

The assembly was empowered to exercise only parliamentary control but could overrule the High Authority by a two-thirds majority. Its delegates were composed of deputies of national parliaments.

The ECSC Council of Ministers

The function of the Council of Ministers was to "harmonise the actions of the High Authority and the governments responsible for the economic policy of respective countries." It was composed of representatives of member countries, each of which delegated a member of its government. Most decisions of the council were valid if voted by a majority of representatives. Unanimous agreement was required only on decisions concerning production questions and shortages. Taken as a whole, the treaty was similar to a federal constitution, embodied in a long and complex document.

There is, however, a basic incompatibility between the community's provenance, limited to the coal and steel industries, and the sovereignty of the member countries, each of which is responsible for its own general economic policy. As a practical matter, during the first 17 years of the community's existence, authority on all substantive issues remained vested in the national governments. The High Authority was autonomous only in matters of secondary importance. Thus, the coal crisis of 1958—when West German, Belgian, and French stocks of unsold coal rose to unmanageable proportions—was resolved at the national level. All the High Authority could do was to confirm the measures taken, even when they were contrary to the provision of the treaty. Similarly, the reduction of the labour force in coal mining from 650,000 persons at the end of 1957 to 300,000 10 years later was effected by individual countries; there was no community-wide action.

The treaty reserved for member countries responsibility for their own trade policies toward third countries. This hindered the establishment of an effective common market since a common market requires a unified system of protection from foreign competition. At the height of the coal crisis, for example, when stocks of coal rose in Belgium, West Germany, and France, Italy nonetheless continued to buy cheap supplies from the United States.

Later developments. Despite such difficulties much has been accomplished by the community. The markets for steel and coal have been liberalized to a considerable degree; the community has been a useful forum in which questions of common interest could be examined; and it has fostered the growth of an international spirit, which did much to facilitate the negotiation of the Treaty of Rome, establishing the European Economic Community and the European Atomic Energy Community (Euratom).

In 1957 the Coal and Steel Community's assembly and Court of Justice were replaced by parallel institutions established by the European Economic Community. In 1967 its executive organs were merged with those of the European Economic Community and Euratom in what is now called the European Communities. The other provisions of the treaty remained unchanged.

The Common Market

The European Economic Community. The European Coal and Steel Community was only the initial step in the movement for European integration. On March 25, 1957, its six member governments signed the Treaty of Rome, under which they agreed to establish the European Economic Community, or Common Market, which came into being on Jan. 18, 1958. In 1973 it was enlarged with the entry of Great Britain, Ireland, and Denmark. Since that time, Greece (1981), Spain (1986), and Portugal (1986) have become members. The EEC is the most far-reaching attempt at economic integration among sovereign countries. Its founding treaty has been the model, in whole or part, for all subsequent attempts at economic integration.

The Treaty of Rome aimed to "establish a common market" and "progressively bring the economic policies of members into alignment" so as to

promote the harmonious growth of economic activity in the Community as a whole, regular and balanced expansion, augmented stability, a more rapidly rising standard of living, and closer relations between the participating states.

The treaty pledged the signatories to

abolish customs duties and quantitative restrictions on the entry and outflow of merchandise, to abrogate all other measures having an equivalent effect, and to fix a common customs tariff for imports from non-member states.

They also agreed to "abolish, as between members, all barriers to the free movement of persons, services and capital." This was to be accomplished during a transition period of 12 years. The transition period ended on Jan. 1, 1970, and the community then entered into its definitive phase.

Formation of a customs union. The treaty set a timetable for the abolition of customs duties between member states. On balance, this timetable was met and in some areas exceeded so that, by the middle of 1968, tariff barriers had been abolished for agricultural as well as industrial products. By that date also, most quota restrictions had been lifted. The customs posts had not disappeared, however; they were still needed for such tasks as assessing and collecting the compensatory taxes that equalized the differences in taxes between member countries.

Tariffs on imports from outside the community were gradually brought closer, and on July 1, 1970, a common community tariff was put into effect.

Development of a common agricultural policy. When the treaty took effect at the beginning of 1958, agriculture was subsidized in all six member countries. The various price-support mechanisms differed substantially, as did foreign-trade policies and tariff levels. The cumulative impact of governmental intervention of various kinds over the years had led to major differences among the members in agricultural price levels. With the average price of wheat in the six countries in 1959 indexed at 100, the relative price levels in individual countries were as follows: Germany, 108; France, 78; Italy, 108; Belgium, 101; Luxembourg, 119; and The Netherlands, 86. The achievement of common policies in agriculture appeared to be so difficult that the treaty limited itself to setting forth a number of general provisions on which agreement seemed feasible. Despite this, a common agricultural policy was achieved: all tariff and quota restrictions on trade in farm products among member countries were abolished; a common set of tariffs on agricultural imports from non-EEC countries was established; and a common system of price supports took the place of the former national systems.

'he need
o stabilize
gricultural
rices

The price supports required difficult compromises among the member governments because of the differences in their domestic price levels for farm products. The EEC wheat price, for example, was set roughly halfway between the prices of the lowest-cost suppliers in the community, France and The Netherlands, and those of West Germany, which was the highest. France exerted considerable political pressure to persuade West Germany to accept a substantial lowering of the returns to its wheat producers. The community prices are supported by purchases from a common fund. The fund begins buying in any area where the price of a crop drops to a fixed intervention level. The cost of the price support program is financed by contributions from the members.

Since its inception, the common agricultural policy has experienced several fundamental problems, especially recurrent surpluses and conflicts of interest between large- and small-scale producers. Surpluses have originated as a result of the price support system, which helps marginal farmers stay in business, but often encourages more productive farmers to overproduce, creating surpluses that must be purchased with EEC funds. Recurrent surpluses of butter and sugar have become a particular problem. Conflicts of interest have arisen between countries that are net food importers (*e.g.*, the United Kingdom and West Germany), which make large contributions to the common policy but receive little return in export subsidies and price supports, and those that are net food-exporting countries (like France and Italy), which receive greater relative support.

Toward a harmonization of policies. Another fundamental aim of the Treaty of Rome was to achieve a general harmonization of national economic policies. The treaty envisaged the working out of common rules covering such matters as competition, taxation, and other economic legislation. It also called for the development of common policies in such areas as foreign trade and transportation. Members were asked to concert their economic policies in the fields of fiscal and monetary policy, balance-of-payments policy, and social welfare.

Relations with other countries. The treaty provides that overseas countries and territories having a special relationship with Belgium, France, The Netherlands, and Italy may be granted associate membership. The purpose of this was to eliminate preferential tariff arrangements by any member with outside countries by extending the same preferential terms to all of the members. The treaty also provides for a European Development Fund to assist the economic development of these countries.

EEC associate membership

New members. Any European state may request membership in the EEC. Acceptance requires a unanimous decision by the present members after "the conditions for entry and the modifications to be made to the Treaty as a result" have been agreed upon by the member states and the would-be entrant.

Associates. The EEC may also conclude an agreement of association with a nonmember country, a union of states, or an international organization. The associate is entitled to special terms in its trade with the EEC and can send representatives to its meetings. Associate status providing for a custom union after a 12-year transitional period was granted to Greece in 1962, prior to its accession as a full member, and to Turkey in 1964. Partial association agreements have been signed with other countries.

Constitution of the EEC. As mentioned above, since 1967 the governing bodies for the EEC, the ECSC, and Euratom have been integrated. These bodies include the Council of Ministers, the Commission, the European Parliament, and the Court of Justice. The Council of Ministers and Commission are assisted by the Economic and Social Committee with advisory functions. Headquarters are in Brussels, Luxembourg, and Strasbourg.

The council has decision-making power in all matters falling within the sphere of competence of the communities. Each member sends one delegate. During the transition period before 1970, all important decisions had to be taken unanimously; they may now be taken by either absolute majority or "qualified" majority (*i.e.*, with weighted voting rights) or unanimously. Either a qualified majority or unanimity are required on certain matters by the treaty. Decisions must be taken unanimously on a very large number of issues.

The Commission members are "selected by reason of their general competence and offering the utmost guarantees of independence." They are appointed by agreement among the member countries, not more than two being of the same nationality. Members serve a four-year term and may be reappointed. They are expected to act in the general interest of the communities, without deference to any government or other organization.

The EEC Commission

The Commission's basic function is to watch over the application of the Treaty of Rome and other treaties and to assist the council with recommendations or advice. Its powers for the most part are delegated to it by the council.

The European Parliament is formed of delegates selected from the parliaments of the members. In principle the parliament meets once a year. It passes recommendations and resolutions and discusses the Commission's annual report. It may pass a resolution of no confidence in the Commission by a two-thirds majority, resulting in immediate dismissal of the Commission members.

The Court of Justice is charged with the interpretation and application of the treaties. It is composed of 11 judges appointed by mutual agreement of the governments of the member states for six-year terms. Actions may be brought in the court by any member state or by any physical or legal person. The court's powers are considerable since community law takes precedence over the national laws of each member country.

The Economic and Social Committee is a consultative body composed of representatives of various economic and social strata, including manufacturing, agriculture, transportation, trade, handicrafts, the liberal professions, wage earners, and the public. Its members are appointed for a four-year term by unanimous decision of the council and may be reappointed.

Britain's reaction to the Common Market

The European Free Trade Association. When the European Economic Community was being organized, Great Britain sought to organize a free-trade area that would include 17 member countries of the Organization for European Economic Co-operation. This would have given Britain access to the benefits of the industrial common market on the Continent while avoiding possible infringements of British sovereignty. The effort failed, mainly because of French opposition. Britain then undertook the formation of a free-trade area in association with Austria, Denmark, Norway, Portugal, Sweden, and Switzerland.

The convention setting up the European Free Trade Association (EFTA) was signed at Stockholm on Jan. 4, 1960. The preamble stated that one of the main purposes of the organization was to "facilitate the future establishment of a wider multilateral association for abolition of customs barriers." More specifically, it was intended as a mechanism for freeing trade with the six Common Market countries without subscribing to the commitments of political character embodied in the Treaty of Rome. In the meantime, EFTA gave its seven members a stronger bargaining position vis-à-vis the other six, as well as the means of creating a large market of their own.

Operation of the EFTA. The European Free Trade Association is governed by a council composed of one member from each participating state. The council found it useful to set up a joint consultative committee composed of representatives of industry, business, and labour; a set of six permanent technical committees (on customs, trade, economic development, agriculture, economics, and budget); and working parties dealing with special topics.

The EFTA treaty, like that of the EEC, provided for a transitional period, set forth rules governing competition, and called for the abolition of all indirect protection and trade discrimination.

EFTA's special problem

The EFTA had one special problem arising from its nature as a free-trade area. Since the duties charged on imports from outside countries might differ from one member to another, traders could take advantage of the differences by channeling imports through the country levying the lowest rates and delivering them to customers in another member country. Rules were established to prevent this by classifying merchandise according to whether it was produced or fabricated in one of the member countries. In the case of goods made from imported raw materials, the rules required that the import content not exceed 50 percent of the export price of the finished product.

The EFTA's record. Although a 10-year transitional period was originally envisaged, internal customs barriers on industrial goods were eliminated on Jan. 1, 1967, three years ahead of schedule. Bilateral trade agreements have also been negotiated to increase trade in agricultural products.

The EFTA passed through two grave crises in the 1960s. The first was in 1961 when Britain, acting unilaterally, informed its partners that it had applied for membership in the EEC. The upshot was a joint declaration in which the EFTA members committed themselves to "coordinate their action and remain united throughout the negotiations." The second crisis occurred in October 1964, when, to shore up the pound sterling, Britain suddenly introduced a surcharge of 15 percent on all its industrial imports—an act that was in violation of the treaty.

Finland became an associate member of EFTA in July 1961, and Iceland was admitted to full membership in March 1970. In 1973 Britain and Denmark left the association when they were accepted as members in the European Economic Community—Britain, after two previous unsuccessful tries. Since that time agreements have been reached between EFTA and EEC promoting trade between the two groups.

Economic integration in Latin America. Progress toward economic integration in Europe encouraged the Latin-American republics to make similar attempts. By the late 20th century several organizations had been established to work toward such integration; they include the Central American Common Market; the Latin American Free Trade Association; the Andean Group; and the Caribbean Community and Common Market. (Ma.A./Ed.)

The Central American Common Market. On June 10, 1958, El Salvador, Guatemala, Honduras, Nicaragua, and Costa Rica signed a multilateral treaty aiming at free trade and economic integration. It provided for the establishment of a free-trade area within 10 years. The participating countries also agreed to the industrial integration of the region. These arrangements were completed by the signature on Dec. 13, 1960, of the Treaty of Managua. Its aims were similar to those of the EEC in Europe, namely, the establishment of a common market within five years and the organization of integrated industrial development. Most barriers on the region's internal trade have since been removed or reduced, and a single customs tariff has been introduced for most products imported from outside the area.

The Treaty of Managua

Economic integration in Central America has been hampered by disagreements and military conflicts in the area. Following a dispute with El Salvador in 1970, Honduras in effect withdrew from common market membership by implementing tariffs on imports from other member countries. In 1980, however, Honduras signed a treaty with El Salvador, settling their dispute and restoring Honduran participation in the common market trade agreements. During the 1980s, tensions between the revolutionary government of Nicaragua and its neighbours, as well as other disorders, have disrupted trade among the nations of Central America.

The Latin American Integration Association. On Feb. 18, 1960, Argentina, Brazil, Chile, Mexico, Paraguay, Peru, and Uruguay signed a treaty setting up the Latin American Free Trade Association (LAFTA), predecessor to the Latin American Integration Association. By 1970 the seven signatories had been joined by Ecuador, Colombia, Venezuela, and Bolivia. The treaty provided for a 12-year transition period during which all obstacles to trade were to be eliminated. It was based on the principle of reciprocity and most-favoured-nation treatment. Member states also committed themselves to progressive coordination of their industrialization policies. Special treatment was provided for agriculture and for the relatively least-developed member countries.

Liberalization of trade between the member countries was carried out initially through negotiation of product-by-product concessions. In 1967, however, the negotiations failed; they were postponed to 1968, when agreement was reached on a system of across-the-board automatic tariff reductions similar to those of the European Economic Community. The eventual aim was that LAFTA be the first step in a process that would to lead to a common Latin-American market; but during the 1970s it became apparent that the geographic diversity and varying levels of economic development exhibited by the member countries were handicapping the formation of a true common market within the association's existing framework. In the late 1970s negotiations were begun to establish a new framework for economic integration, and in 1980, 20 years after the creation of LAFTA, the Latin American Integration Association (LAIA; Asociación Latino-Americana de Integración) was formed. Unlike its predecessor, LAIA adopted an alternative to the concept of a free-trade area in that it opted for the establishment of bilateral preference agreements that would take into account the varying stages of economic development of the member countries. In order to best negotiate bilateral preference agreements the member nations were divided into three categories: most developed countries (Argentina, Brazil, and Mexico); intermediate developed countries (Chile, Colombia, Peru, Uruguay, and Venezuela); and least-developed countries (Bolivia, Ecuador, and Paraguay). Cuba was admitted to LAIA in 1986 with observer status.

The goal of LAFTA

The Andean Group. In 1966 Bolivia, Chile, Colombia, Ecuador, Peru, and Venezuela, all members of the Latin

American Free Trade Association, agreed to form a regional subgroup. The Andean Group finally began its official existence in June 1969 without Venezuela, which had withdrawn. By 1973, Venezuela had decided to join, but Chile withdrew in 1976. (The Andean Judicial Tribunal was established in 1980 to monitor, interpret, and resolve problems resulting from Andean Group decisions. The tribunal is composed of one judge from each country, serving a six-year term.) Among the group's aims are the acceleration of economic integration between member countries, the coordination of regional industrial development, the regulation of foreign investment in member countries, and the maintenance of a common external tariff.

The Caribbean Community and Common Market. Established in 1973 by 12 Caribbean countries, the Caribbean Community and Common Market (Caricom) is the successor to the Caribbean Free Trade Association, which was founded in 1968 by five former British colonies (Antigua, Barbados, Guyana, Jamaica, and Trinidad and Tobago), all of whom joined the new organization. The organization attempts to encourage economic integration in the Caribbean region and achieved partial agreement to a common external tariff and protective policy for the community in 1978.

Aims of Caricom

Caribbean economic integration was curtailed between 1976 and 1978, in part because of import restrictions imposed by Jamaica and Guyana, and in part because of dissatisfaction among the less-developed countries, who claimed that they were not receiving their fair share of trading revenues. By 1980 Jamaica and Guyana had removed their import restrictions and the Caricom Council had endorsed several measures to improve the status of the less-developed countries within Caricom. The less-developed countries, however, remained dissatisfied, and in 1981 the seven former members of the West Indies Associated States (Antigua and Barbuda, Dominica, Grenada, Montserrat, Saint Kitts-Nevis, Saint Lucia, and Saint Vincent) formed a subregional economic integration organization, the Organization of Eastern Caribbean States. They retained their Caricom membership. (Ed.)

The Council for Mutual Economic Assistance. A Soviet-sponsored effort to integrate the economies of eastern Europe began as early as Jan. 25, 1949, in response to the Marshall Plan. The founding states were Bulgaria, Hungary, Poland, Romania, Czechoslovakia, and the Soviet Union. Albania joined in 1949, the German Democratic Republic in 1950, and the Mongolian People's Republic in 1962. Albania ceased to participate after 1961. In its early years the activities of the Council for Mutual Economic Assistance (or Comecon) were limited mainly to the registration of bilateral trade and credit agreements among the member countries. After Stalin's death in 1953 it made efforts to promote industrial specialization and to reduce "parallelism" in the economies of its members. In 1956 and 1957, when most of its standing commissions began to operate, attempts were made to harmonize the long-term plans of the members.

Integration in eastern Europe

The establishment of the European Economic Community in 1958, together with pressures from the eastern European countries for a greater degree of independence, induced the Soviet leadership to rethink the organization. A new charter was signed by the members in Sofia on Dec. 14, 1959.

Economic objectives. The economic aims of the Council for Mutual Economic Assistance are

> through unification and coordination of the efforts of its member countries to contribute to the balanced growth of their economies, more rapid economic and technical progress, increased industrialization of those countries which are least advanced in this area, the continued growth of labour productivity, and the regular improvement of the welfare of the peoples of member countries.

Its pursuit of these objectives has been hindered by certain political and economic constraints. One of the most serious is the absence of flexible and realistic price systems in the member countries. This has made it impossible to base trade on relative prices; instead it has been conducted mainly on a barter basis through bilateral agreements between governments. In negotiating such agreements, the parties have been led to use "world prices"; *i.e.,* prices prevailing in the trade of countries outside the Council for Mutual Economic Assistance. Another hindrance to economic integration is the highly centralized economic planning in the member countries, which so far have had only limited success in coordinating their plans.

There have also been serious nationalistic tensions within the council. The Romanian government, for example, announced its intention to pursue all-around industrialization, including the development of its heavy industries, in opposition to the policy of specialization in raw materials and agricultural products that was said to have been the Council for Mutual Economic Assistance's policy for Romania.

Among the practical achievements of the Council for Mutual Economic Assistance, however, have been: the organization of railroad coordination (1956); construction of a high-voltage electricity grid (1962); creation of the International Bank for Economic Cooperation (1963); the pooling of 93,000 railway freight cars (1964); and construction of the "Friendship" oil pipeline from the Soviet Union's Volga region to the eastern European countries. Comecon initially was composed of the Soviet Union's eastern Europe satellites, but in 1972 Cuba became a member and in 1978 Vietnam joined.

Achievements of Comecon

The constitution of the Council for Mutual Economic Assistance. The highest authority of the organization is the council-in-session. It is composed of delegations from all member countries, the composition of each delegation being fixed by the government concerned. The conference of representatives of member countries, composed of one representative of each country, may issue recommendations and decisions. It may also submit proposals for examination by the council-in-session.

Various permanent commissions are composed of experts and official representatives of member countries. Some are general economic commissions; others deal with issues affecting specific industries. The headquarters of the various commissions are located in the capital cities of member countries. The central secretariat is in Moscow.

The Council for Mutual Economic Assistance is often called the eastern European counterpart of western Europe's EEC. Although the general aims are indeed the same, the two organizations differ radically in their approach to the problems involved. The EEC aims to achieve integration on a decentralized basis by means of an economic market in which goods, services, capital, and persons have full freedom of movement—a market regulated by uniform economic legislation. The Council for Mutual Economic Assistance seeks to achieve cooperation among national economies each of which is centrally planned and administered. (Ma.A.)

GATT rules regarding regional arrangements. When countries join regional trading groups such as the European Economic Community, they provide preferences to one another. For example, German producers can export duty-free to France, whereas U.S. or Japanese exporters still have to pay duties on products shipped to France. Thus, German producers are preferred over U.S. or Japanese suppliers; a customs union represents a departure from most-favoured-nation treatment. Nevertheless, countries entering a customs union or free-trade association are not in violation of their commitments under the General Agreement on Tariffs and Trade; customs unions and free-trade associations are permitted under GATT article XXIV.

GATT article XXIV

Specifically, this article allows countries to grant special treatment to one another by establishing a customs union or free-trade association, provided that (1) duties and other trade restrictions are "eliminated on substantially all the trade" among the participants; (2) the elimination of internal barriers occurs "within a reasonable length of time" (commonly interpreted as permitting a phase-in period of not more than 10 years); and (3) duties and other barriers to imports from nonmember countries "shall not on the whole be higher or more restrictive" than those preceding the establishment of the customs union or free-trade association. The third condition was explicitly aimed at protecting the rights of outside countries.

The first condition disapproves partial preferential arrangements covering only some products, while accepting broad arrangements covering (substantially) all products. It was supported on the ground that large, unrestricted markets provide substantial benefits—most notably, the large internal market within the United States. Such benefits should also be available to others. For example, when the GATT articles were being drafted, consideration was being given to an integration of the nations of western Europe.

Jacob Viner's reaction to article XXIV

Shortly after article XXIV was written, it received substantial support in the classic study of Jacob Viner, *The Customs Union Issue* (1950). Viner, a Canadian-born U.S. economist, saw efficiency as the main gain from international trade, since trade encourages production in a less costly location. He contended that a customs union works to increase efficiency in one way, but to decrease it in another. To explain why, Viner drew a distinction between two forces at work when a customs union is established. As two (or more) countries cut tariffs on each other's products, new trade is created. Some goods previously bought from domestic producers are now bought from lower-cost producers in the trading partner, whose goods now come in duty-free. Such trade creation improves efficiency.

When, however, a country removes tariffs on its partner's goods but not on the goods of outside countries, the partner has preferred access. As a result, some purchases are switched—goods are bought from the partner nation, rather than from the outside world. Such trade diversion reduces efficiency; purchases are switched from the efficient outside country to the less efficient partner nation. A customs union (or free-trade area) may be predominantly trade-creating, which is desirable, or it may be predominantly trade-diverting, which is not.

Viner's book thus introduced a skeptical note into the discussion of customs unions, which had previously been given broad approval. Viner's work also supported the distinctions made in article XXIV of the General Agreement on Tariffs and Trade. Clearly, if barriers on imports from nonmember countries are kept down, then trade diversion is less likely. Furthermore, the provision to disapprove partial preferential arrangements covering only some products, while accepting broad arrangements covering virtually all products found support within Viner's framework. Because of the political dynamics of trade negotiations, partial preferential arrangements generally cause more trade diversion than trade creation.

This can be illustrated in a hypothetical situation where countries (say, France and West Germany) are permitted to get together to make whatever preferential agreements they wish. A natural way for France to open negotiations would be to say to West Germany: "We'll cut tariffs on your automobiles and buy from you rather than Japan, if you will cut tariffs on our sugar and buy from us rather than from the Caribbean nations." In other words, negotiators tend to pick and choose those items previously imported from outside countries; they tend to cut tariffs where trade diversion is greatest. By requiring a comprehensive approach, article XXIV ensures that trade-creating tariff cuts will be made, too. (P.Wo.)

Patterns of trade

DEGREES OF NATIONAL PARTICIPATION

Nations vary considerably in the extent of their foreign trade. As a very rough generalization, it may be said that the larger a country is in physical size and population, the less its involvement in foreign trade, mainly because of the greater diversity of raw materials available within its borders and the greater size of its internal market. Thus, the participation of the United States is relatively low, as measured by percentage of gross national product, and that of the Soviet Union even lower. The U.S. gross national product, however, is so immense by world standards that the United States still ranks as one of the world's most important trading countries. Some of the smaller countries of western Europe (such as The Netherlands) have export and import totals that approximate half of their gross national products.

TRADE AMONG DEVELOPED COUNTRIES

The greatest volume of trade occurs among the developed, capital-rich countries of the world, especially among industrial leaders such as Australia, Belgium, Canada, France, Italy, Japan, The Netherlands, Spain, Sweden, the United Kingdom, the United States, and West Germany. Generally, as a country matures economically its participation in foreign trade grows more rapidly than its gross national product.

Trade within the EEC

The European Economic Community (EEC) affords an impressive instance of the gains to be derived from freer trade. A major part of the increases in real income in EEC countries is almost certainly attributable to the removal of trade barriers. The EEC's formation cannot, however, be interpreted as reflecting an unqualified dedication to the free-trade principle, since EEC countries maintain tariffs against goods from outside the Community.

TRADE BETWEEN DEVELOPED AND DEVELOPING COUNTRIES

Difficult problems frequently arise out of trade between developed and developing countries. Many less-developed countries are tropical, frequently relying heavily for income upon the proceeds from export of one or two crops, such as coffee, cacao, or sugar. Markets for such goods are highly competitive (in the sense in which economists use the term competitive)—that is, prices are extremely sensitive to every change in demand or in supply. Prices of manufactured goods, the typical exports of developed countries, are commonly much more stable. Hence, as the price of its export commodity fluctuates, the tropical country experiences large fluctuations in its "terms of trade," the ratio of export prices to import prices, often with painful effects on the domestic economy. With respect to almost all important primary commodities, efforts have been made at price stabilization and output control. These efforts have met with varied success.

Comparable problems arise when the developing country exports a mineral such as petroleum or copper. The initiative in developing such a resource has often been taken by a foreign company from a developed country that owns (in part if not in full) the extracting capital facilities. Particularly since the mineral resource is exhaustible, charges of exploitation are common. The matter is a continuing source of political strife and may on occasion lead to expropriation of the mineral properties.

Developing country trade organizations

Several groups of developing countries have joined in creating organizations for the promotion of trade between themselves. Notable examples include the Central American Common Market (1961), the Organisation Commune Africaine et Mauricienne (1965), the Association of Southeast Asian Nations (1967), the Caribbean Community and Common Market (1973), and the Latin American Integration Association (1980).

SOVIET FOREIGN TRADE

The Soviet Union's involvement in foreign trade, measured as before by comparing exports or imports against gross national product, is even lower than that of the United States. The U.S.S.R. is not in general an important trader in world markets, but it has made occasional large-scale entries, such as its grain purchases during years of domestic crop shortage in the 1960s and 1970s.

Soviet trend toward expanded trade

Since the 1917 Revolution, Soviet policy has clearly been influenced by the desire for self-sufficiency, further reinforced by its suspicions of the capitalist world and by its strong desire for centrally directed planning. Today, the U.S.S.R. appears to be showing a greater interest in expanded trade connections, doubtless because of growing industrial maturity.

The attitude of the U.S.S.R. toward foreign trade was forced to undergo some change when that country acquired influence over the eastern European countries after World War II. It established the Council for Mutual Economic Assistance (Comecon) to direct trade with and among its satellite neighbours. Comecon, however, appears to be primarily a device for implementing and integrating central development plans among member countries—unlike the EEC, which was planned to facilitate the unrestricted movement of goods and productive inputs. The patterns

of Comecon trade, however, are often difficult for an outsider to interpret. (R.R./Ed.)

International commodity trade

Goods that are traded internationally fall into two broad categories—primary goods and manufactured products. Manufactured products, such as machinery and clothing, comprise products whose value reflects largely the cost of manufacturing processes. Such manufacturing processes contribute relatively little to the value of primary goods, such as crude petroleum and cotton, which undergo but little processing before they are traded. Commodities and commodity markets are terms used as synonyms for primary goods and the markets in such goods.

PRIMARY COMMODITY MARKETS

Actual and futures markets

Trade in primary goods may take the form of a normal exchange of goods for money as in any everyday transaction (referred to technically as trade in "actuals"), or it may be conducted by means of futures contracts. A futures contract is an agreement to deliver or receive a certain quantity of a commodity at an agreed price at some stated time in the future. Trade in actuals has declined considerably and in many cases (such as the Liverpool markets in cotton and grain) has even come to a halt.

Operation of the market. The great bulk of commodity trading is now in contracts for future delivery. The purpose of trading in futures is either to insure against the risk of price changes (hedging) or to make a profit by speculating on the price trend. If a speculator believes, for example, that prices will rise, he buys a futures contract and sells it when he wishes (*e.g.*, at a more distant delivery date). The speculator either gains (if prices have risen) or loses (if they have fallen), the difference being due to the change in price.

"Hedging" means the offsetting of commitments in the market in actuals by futures contracts. A producer who buys a commodity at spot (current) prices but does not normally resell until three months later can insure himself against a decline in prices by selling futures: if prices fall he loses on his inventories but can purchase at a lower price; if prices rise he gains on his inventories but loses on his futures sales. Since price movements in the actuals market and the futures market are closely related, the loss (or gain) in actual transactions will normally be offset by a comparable gain (or loss) in the futures market.

The operation of futures markets requires commodities of uniform quality grades in order that transactions may take place without the buyer having to inspect the commodities themselves. This explains why there is no futures market, for example, in tobacco, which varies too much in quality. A steady, unfluctuating supply also is needed; this is referred to technically as "low elasticity of supply," meaning that the amount of a commodity that producers supply to the market is not much affected by the price at which they are able to sell the commodity. If supply could be adjusted relatively quickly to changes in demand, speculation would become too difficult and risky because exceptionally high or low prices, from which speculators are able to profit, are eliminated as soon as supply is adjusted. Monopolistic control of demand and supply is also unfavourable to the operation of a futures market because price is subject to a large extent to the control of the monopolist and is thus unlikely to fluctuate sufficiently to provide the speculator with an opportunity for making profits. There is, for example, no market in diamonds, since there is only one marketing cooperative. In 1966 the London market in shellac ceased to function after the Indian government applied control of exporters' prices at the source.

Major commodity markets

Before World War II London was the centre of international trade in primary goods, but New York City has become at least as important. It is in these two cities that the international prices of many primary products are determined. Although New York often has the bigger market, many producers prefer the London market because of the large fluctuations in local demand in the United States that influence New York market prices. In some cases international commodity agreements have reduced the significance of certain commodity markets.

There are markets in both New York and London for numerous primary goods, including cotton, copper, cocoa, sugar, rubber, coffee, wool and wooltops, tin, silver, and wheat. Tea, wool, and furs are auctioned in London, but in the case of many other commodities, auctions have been superseded by private sales. In London the metal market is much more a "spot" or delivery market than other futures markets. Many countries have their own markets: Australia for wool, Sri Lanka (Ceylon) and India for tea, and Malaysia for rubber and tin.

The terms of trade. The relation between the price of primary goods and that of manufactures has long intrigued economists. The relationship is known as the "terms of trade" and may be defined as the ratio of the average price of a country's or a group of countries' exports to the average price of its imports. The long-range trend of the terms of trade between primary products and manufactures has been the subject of diametrically opposed conclusions: some theorists hold that the trend is favourable to the less-developed countries, others that it is unfavourable. Faulty statistical material and methods in various countries are responsible for this lack of agreement.

Any comparison of the terms of trade over a long period of time is very difficult and may be misleading because the structure of trade changes, as does the quality of the groups of goods studied. Many economists believe that the terms of trade were adverse for less-developed countries from 1870 to 1938. They point to the fact that as developed countries become more technologically advanced there is a tendency for them to require relatively less in the way of primary products. A downward influence is thus exercised on primary product prices. Another factor is that in the industrial countries the benefits of progress find expression not in lower prices but in higher wages. This, together with inflationary pressures, means that prices of manufactured goods produced by the developed countries tend to rise steadily. There is thus a tendency, it is argued, for the less-developed countries to receive relatively less for what they have to sell and to have to pay more for what they need to buy. But the statistical problems posed by any attempt to verify this hypothesis are considerable. The countries selected, the relative weight assigned to the various goods, changes in transport costs, and the fact that the quality of manufactured goods has improved much more than that of primary goods make the statistics unreliable. There is also the problem that the terms of trade between primary commodities and manufactures do not necessarily coincide with the terms of trade between less-developed and industrial areas.

Effects of changes in the terms of trade

Even if it were established that the terms of trade have moved against the less-developed, largely primary-producing countries, this would not necessarily mean that their balance-of-payments situation has been adversely affected. A decline in the terms of trade may in fact improve a country's balance-of-payments, because, although the prices of that country's exports have fallen, it may, as a consequence of this fall in price, be able to sell a far larger quantity. Total revenue from exports may thus increase. Similarly, although imports may become more expensive, the result may be that the country's demand for imports drops very steeply, so that less is spent on them than when they were cheaper.

These problems make it extremely difficult to generalize about the effects of commodity price changes on the economic situation of one or a group of countries.

PRICE MOVEMENTS

Prices usually vary widely in commodity markets, not only in the short run but also in the long run. In the short run there are frequent changes in supply because of varying climatic conditions (for agricultural products) and because of political and other events on the international scene (such as the closure of the Suez Canal) and in individual countries (such as strikes). As a rule, price changes do not give rise in the short run to substantial changes in the supply of or demand for primary goods (low elasticity of supply and demand). Business cycles in the importing

countries, however, have an influence on demand. Market conditions differ, of course, from product to product. In the case of sugar and wheat, demand is fairly stable, but supply is not; as regards tin, and, indeed, the majority of metals, the converse is true. In the case of industrial commodities, such as cotton, there are fluctuations in both supply and demand.

Long-run effect of price change

In the long run the extent of changes in demand and supply is usually greater. A considerable and sustained price increase, for example, may result in a fall in demand and the appearance of substitute products. After a number of years, supply may increase in response to a higher level of demand reflected by higher prices. The length of time required to adjust supply to demand varies from commodity to commodity. Tree crops, for example, need a long growth period, and mineral reserves are tapped only if expectations about the price trend are favourable.

Effect on economic development. Through their repercussion on export earnings, price fluctuations are often held responsible for the variations in the growth rate of countries producing primary goods, especially since exports of a single primary good account for a large part of the total exports of many countries. But apart from the fact that, as described above, quantities exported influence export earnings as much as prices, there are many other factors that determine export earnings. Such factors include the type and destination of exports and, above all, the economic policies of the countries concerned.

It is thus difficult to generalize about the relation of foreign trade to economic growth. Many countries with very unstable exports have relatively stable national incomes; others whose exports are stable have highly unstable national incomes. The stimulus from exports will usually be stronger, for example, if the rate of demand for these exports is growing rapidly. Often, however, the transmission of growth to the nonexporting sector of the economy is impeded in less-developed countries by the economic, social, and political organization of the economy. It is important, for example, for some countries to try to decrease exports of goods that have a slowly growing demand and at the same time to try to increase exports of goods, such as minerals, for which world demand is growing more rapidly.

Efforts to stabilize prices. The uncertainty both for private producers and for governments resulting from sharp and sudden commodity price changes has resulted in many efforts to achieve greater stability on the market in primary goods.

Action in individual countries. In theory a country could insulate domestic producers against international price fluctuations through variable charges and subsidies, but politically it is difficult to tax away producers' profits during a period of rising prices and to hold the resulting revenue in order to redistribute it should prices and profits fall.

Price fluctuation in Africa

In Nigeria, Ghana, Sierra Leone, and The Gambia, for instance, national marketing boards that attempted to even out price fluctuations of cocoa, cotton, and peanuts (groundnuts) were in operation before those countries became independent. In the former French territories in Africa, stabilization funds fixed producer prices and controlled margins and profits. The main dangers inherent in national stabilization schemes are inconsistent government policies and the excessive operating costs of the public bodies concerned. These factors explain the unsatisfactory results of many national price agencies.

International cooperation. In the 1920s international cartels were created for rubber, sugar, tin, and tea, but they yielded no lasting results. Nor did cooperation between the governments of exporting and importing countries (such as in the International Wheat Agreement of 1933 and the International Sugar Agreement of 1937) serve to attain the desired goals during the Great Depression. Of special significance among more recent attempts to raise and stabilize a commodity price has been the one made by the Organization of Petroleum Exporting Countries (OPEC). (The special features of the oil market are considered below.) Other attempts to stabilize commodity prices since World War II have mainly assumed three forms—the multilateral contract agreement, the quota agreement, and the buffer-stock agreement. Transactions are effected at world market prices. When a minimum or a maximum price is reached or approached, efforts are made to ensure that prices remain within the two limits. Each of the three systems achieves this in a different way.

Three approaches to price stabilization

In the multilateral contract system, consumers and producers undertake to buy or sell a specified quantity of the commodity at agreed minimum and maximum prices, or at a price within the agreed range.

In the quota method, the quantity negotiated is determined by a previously fixed quota when a minimum or maximum price is exceeded. When there is a surplus, the producers restrict their exports or production; when there is a shortage, quotas are allotted to the consumer countries. With the buffer-stock method, stability is ensured by a combination of an export control arrangement and a buffer-stock arrangement. In certain circumstances exports are restricted by the controlling body. The buffer-stock agency buys when the market price is in the lower sector or at the floor price set out in the agreement; the buffer-stock agency sells when the market price is in the upper sector or at the ceiling price.

Results. The utility of commodity agreements in general can hardly be judged on past experience. Experience with wheat, sugar, and tin agreements, which cover a comparatively long period, is not conducive to generalization. Some degree of stability, though at a high price level, was achieved in the case of wheat, but this was due to the dominant influence of U.S. and Canadian policies. In the case of tin, too, transactions for the U.S. strategic stockpile exerted an influence. Political factors (including the Cuban revolution) underlay the de facto suspension from 1962 to 1969 of the sugar agreement, which had covered, and still covers, only a limited share of the world market.

The value of world transactions in tin, wheat, coffee, and sugar amounts to only a small part of the value of the world's entire commodity trade. Furthermore, the agreements in question do not cover all transactions. It is, in a way, understandable that only a few such agreements have been concluded; during a boom the producer countries are not inclined to conclude them, and during a depression there is little incentive for consumer countries to enter into them.

Conditions for success. A prerequisite for the success of commodity agreements is that they should embrace the vast majority of producers and especially the largest of them. No transactions should be excluded, and substitute commodities should be covered by the agreements.

Fixing the price range

The most intractable of the difficulties in concluding commodity agreements lies in the fixing of the price range. Neither unduly high nor unduly low price scales are tenable. Future market conditions are not easily foreseeable, so the possibility of errors cannot be ruled out; regular adjustment of the price ranges is necessary.

When it comes to determining the price range, the importing and exporting countries, respectively, do not systematically advocate low and high prices. Certain importing countries are not opposed to a relatively high price because the difference between the international price and the tariff-protected price of domestic producers is thereby reduced; exporting countries in a favourable competitive position are often in favour of lower prices so they will be able to increase their share of the market at the expense of less-competitive countries.

In concluding an agreement, the parties have to bear in mind that complete price stabilization is impossible. It would in fact be undesirable, because in the long run supply and demand need to remain in equilibrium, and the necessary adjustments in the economies concerned must not be precluded. Price fluctuations do not necessarily imply failure, because the fluctuations might well have been larger had the agreement not been concluded.

The method of stabilization needs to be chosen carefully, with due regard for the characteristics of the commodities concerned. The multilateral purchase contract and buffer-stock systems offer the advantage of not requiring any restrictions on production; new producers with improved technical equipment may participate.

A buffer stock needs to be sufficiently large if it is to achieve its purpose. Wider financing facilities are necessary; this is something to which the importing countries could contribute. Even then the buffer stock is better used together with other methods of stabilization. Because of the perishable nature of certain commodities or their bulk and high storage costs, however, a buffer stock is not always feasible. Buffer stocks alone often are not sufficient for the control of prices, and it is sometimes necessary for producers to restrict exports in order to reduce supply, thus pushing prices up.

INTERESTS OF THE LESS-DEVELOPED COUNTRIES

So far as the producer countries are concerned, stabilization of incomes, rather than of prices, is the most important factor. Although commodity agreements may contribute to this, their relatively limited success has caused other proposals to be advanced.

Compensatory financing refers to international financial assistance to a country whose export earnings have suffered as a result of a decline in primary commodity prices. Such a system was instituted in 1963 by the International Monetary Fund (IMF). In 1969 the IMF also began making loans available to countries having a balance-of-payments need in relation to the financing of buffer stocks under international commodity agreements.

EEC stabilization fund

The European Economic Community has established a stabilization fund for its associated overseas countries; prices must fall by a specified percentage before the mechanism of the fund goes into effect, and the richer beneficiary countries must repay the aid received.

Other proposals involve the introduction of simultaneous negotiations for a whole range of commodities. These discussions, however, and more particularly the administration of the resulting multicommodity agreement, would be highly complex. It may also be argued that the significance of export instability has been exaggerated and that most of the economies involved have suffered no serious damage. Thus, the resources devoted to countering price fluctuations and compensatory financing might be better employed in investments or technical assistance.

As to the possibility of the less-developed countries themselves influencing prices, circumstances vary from commodity to commodity. In the case of primary goods, such as coffee, that are produced only in the less-developed countries and for which practically no substitutes exist, action to increase prices can easily be taken if demand is not too much affected by price increases. A simple way to raise prices would be for the governments of producing countries to levy a duty on exports. Attempts by some developing countries to raise prices, however, can induce other developing countries to increase their output. For example, African coffee production was stimulated when Latin-American countries took steps to raise the price of their coffee.

Limitations on pricing

The fact that there are substitutes for a few primary goods (such as cotton, wool, and rubber) limits the extent to which primary-goods producers can raise their prices. Also, most commodities produced by less-developed countries face competition from the developed countries, which may produce the same commodities (such as petroleum, sugar, rice, and tobacco) or goods substitutable in varying degrees (such as soybean oil for peanut oil).

Many agricultural commodities are protected in the developed countries by tariffs, which means that the latter's requirements are often met entirely from domestic production. Some developed countries produce surpluses that are sold abroad at low, subsidized prices. Such commodities are therefore traded to a relatively small extent on world markets. The sales of the less-developed countries are thus influenced by the developed countries' national policies and by the price at which these countries sell their surpluses on the residual markets. The less-developed countries that produce minerals and metals seemingly have the most favourable export prospects because demand for these commodities is expanding among the developed countries, many of which are concerned over the depletion of their domestic resources.

(M.A.v.M./Ed.)

OPEC AND OIL

Of the multinational organizations aimed at affecting the price of a commodity, one of the most significant is the Organization of Petroleum Exporting Countries (OPEC). It was founded in 1960 by Middle Eastern countries and Venezuela, although its membership has come to include developing nations in other parts of the world. Some major oil-exporting nations have remained outside the organization, notably Mexico and the Soviet Union.

OPEC's effect on oil prices

The principal objective of OPEC has been to raise the price received by the oil-exporting countries. During its early years, it was notably unsuccessful: plentiful supplies of oil kept the price low throughout the 1960s. In the early 1970s, however, major changes took place. The rapid economic expansion, which was simultaneously occurring in many countries, put upward pressure on the demand for oil. At the same time, the production of oil was leveling out and beginning to decline in the United States, with the result that U.S. demand for imported oil was rising rapidly.

In 1973, OPEC seized the opportunities offered by the changing market conditions—and by the political and economic disruptions associated with the war between Israel and its Arab neighbours—to raise prices sharply, from about $3 to more than $12 per barrel. Between 1974 and 1979 the international price of oil remained quite stable, but then OPEC was once again successful in pushing the price up sharply—to more than $30 per barrel in 1980. These price increases caused a huge transfer in revenues from the oil-importing nations to the oil-exporting countries. They also contributed to a major increase in inflation in the importing countries.

World economic reaction to higher oil prices

The large increase in revenues in the OPEC nations allowed many of them to embark on major development programs. On the other side, the loss of revenues, combined with the inflationary impact, precipitated major recessions in many of the oil-importing countries in 1974–75 and 1980–82. The higher oil price has also been suggested as a cause of a decline in productivity in many countries after 1973, although the causes of the decline are not well understood.

OPEC has often been called an international cartel, but it lacked the standard enforcement mechanism of a cartel during the two periods (1973 and 1979–80) when prices rose spectacularly. That is, it did not have a mechanism for sharing the market among the oil-exporting nations. Saudi Arabia played a key role in enforcing the organization's price increases.

In the 1970s Saudi Arabia had proven reserves in excess of 150 billion barrels, more than twice as much as any other nation, and five or six times the proven reserves of such major non-OPEC producers as the United States and Mexico. Because of its huge reserves and productive capacity, Saudi Arabia was able to act as the residual supplier, cutting back on production when demand slackened, thus reducing downward pressures on prices. Saudi Arabia's willingness to act as the residual supplier was partly the result of its limited population; even when producing at much less than capacity, it had a very large oil income per capita.

Oil glut of the 1980s

During the early and middle 1980s, the oil market softened markedly. Oil consumption grew much more slowly, partly as a result of the major U.S. recession of 1982 and sluggish growth in western Europe, and partly as a result of increased conservation measures, a reaction to the upward spiral of fuel prices in the 1970s. At the same time, oil output increased in a number of non-OPEC areas such as the North Sea.

The result was downward pressure on prices through the mid-1980s. In order to maintain sales and revenues, OPEC members had an incentive to undercut the posted price. As its oil production fell sharply and the bills from its ambitious development projects continued to increase, Saudi Arabia became less willing to act as the residual supplier. In order to relieve the downward pressure on prices, OPEC members attempted to transform the organization into a more formal cartel, with production quotas for each member. But these efforts faced the classic problem of cartels: each member had an incentive to cheat on the

organization by producing more than its quota and by offering secret price concessions to buyers. (P.Wo.)

International payments and exchange

Economic life does not stop at national boundaries but flows back and forth across them. The money of one country, however, cannot as a rule be used in another country; the flow of payments must be interrupted at national boundaries by exchange transactions in which one national money is converted into another. These transactions serve to cover payments so long as there is a balance between them: local money can be exchanged against foreign money only insofar as there is a counterbalancing offer of foreign money in exchange.

In the Soviet Union and other countries with the Soviet kind of economic planning, there are no legal private markets for foreign exchange; in those countries the state has a monopoly of the business of foreign trade, which is generally conducted through formal agreements on a country-by-country basis. While the currencies of the Communist countries have official par values, these bear no particular relationship to their purchasing power or to the prices at which goods are exchanged. The international economic relationships of those countries therefore fall outside the scope of this discussion.

BALANCE-OF-PAYMENTS ACCOUNTING

The balance-of-payments accounts provide a record of transactions between the residents of one country and the residents of foreign nations. The two types of accounts used are the current account and the capital account.

The current account. When using balance of payments statistics it is important to understand their basic concepts. The balance of payments includes, among other things, payments for goods and services; these are often referred to as the balance of trade, but the expression has been used in a variety of ways. In order to be more specific, some authorities have taken to using the expression "merchandise balance," which unmistakably refers to trade in goods and excludes services and other occasions of international payment.

The various kinds of payments

Figures for the merchandise balance often quote exports valued on an FOB basis (free on board) and imports valued on a CIF basis (including cost, insurance, and freight to the point of destination). This swells the import figures relative to the export figures by the amount of the insurance and freight included. The reason for this practice has been that in many countries the trade statistics have been based on customs house data, which naturally include insurance and freight costs for imports but not for exports. The authorities have more recently made a point of providing estimates of imports valued on an FOB basis.

Another expression, "balance of goods and services," is often used. The British, however, continue to use the term invisibles for current services entering into international transactions. For many years the "visible" balance was taken to be equivalent to exports quoted FOB and imports CIF as explained above. The British authorities have more recently instituted another linguistic usage by which the "visible" balance is equivalent to the true merchandise balance. The old usage still lingers on in the less-expert literature.

And so the total current account is the balance of goods (merchandise) and services. The United Kingdom includes unilateral transfers among "invisibles" and in the current account. The United States statistics, more correctly, show them under a separate heading.

Services include such items as payments for shipping and civil aviation, travel, expenditures (including military) by the home government abroad and expenditures by foreign governments at home, interest and profits and dividends on investments, payments in respect of insurance, earnings of banking, merchanting, brokerage, telecommunications and postal services, films and television, royalties payable by branches, subsidiaries and associated companies, agency expenses in regard to advertising and other commercial services, expenditures by journalists and students, construction work abroad for which local payment is made and, conversely, earnings of temporary workers such as entertainers and domestic workers, and professional consultants' fees. This list contains the more important items but is not comprehensive.

Among unilateral transfers the more important are outright aid by governments, subscriptions to international agencies, grants by charitable foundations, and remittances by immigrants to their former home countries.

The capital account. There is also the capital account, which includes both long-term and short-term capital movements.

Long-term flows. Long-term capital movement divides into direct investments (in plant and equipment) and portfolio investments (in securities). In the 19th century direct investment in plant and equipment was preponderant. The United Kingdom was much the most important contributor to direct investment overseas. In the early part of the century it even contributed to the industrial development of the United States; later its attention shifted to South America, Russia, other European countries, and India. Investment in what came to be called the "Commonwealth" and "Empire," not prominent at that time, became very important in the 20th century. The other countries of western Europe also made important contributions to direct investment overseas.

British leadership in direct investments

The most important items of direct investment were railways and other basic installations. In early stages direct investment may help developing countries to balance their payments, but in later stages there will have to be a flow of interest and profit in the opposite direction back to the investing country. The United Kingdom is frequently cited as the country whose overseas investments were most helpful for developing countries because her rapidly growing population and small cultivable land area permitted her to develop large net imports of food and to run corresponding deficits on her merchandise account. The complementary surplus this generated in the developing countries from which the imports came enabled them to pay the interest and profit on British capital without straining their balances of payments.

Between World War I and World War II the United States began to take a more active interest in overseas investment, but this was not always well-advised. After the great world slump, which started in 1929, international investment almost ceased for lack of profit opportunities.

The international capital market

After World War II the United States began to build up a leading position as overseas investor. The process accelerated in 1956 and afterward, both on direct investment and on portfolio investment accounts. This may have been partly due to the desire of U.S. firms to have plants inside the European Economic Community. Other countries also found more opportunities for capital export than there had been in the interwar period. The United Kingdom gave special attention to the Commonwealth. During the 1970s and 1980s Japan became a major overseas investor, financing its foreign investments with the funds accumulated with its large current account surpluses. The U.S. international position changed sharply in the 1980s. As a result of its large current account deficits, the United States accumulated large overseas debts. Its position changed from that of major net creditor (it had larger investments abroad than foreign nations had in the United States) to that of the largest debtor nation. Its liabilities to foreign nations came to exceed its foreign assets by hundreds of billions of dollars.

Short-term flows. A very important distinction must be drawn between the short-term capital that flows in the normal course of industrial and commercial development and that which flows because of exchange-rate movements. The first class of short-term capital may be thought of as going in the train of direct long-term investment. A parent company may desire from time to time to supply its branch or affiliate with working capital. There may also be repayments from time to time. The second type of short-term capital flow occurs because of expectations of changes in exchange rates. For example, if people expect that the price of the dollar will fall in terms of the Japanese yen, they have an incentive to sell dollars and buy yen.

An international capital market developed in the 1960s

The Eurodollar

dealing in what are known as Eurocurrencies, of which much the most important was the Eurodollar. The prefix Euro is used because initially the market largely centred on the countries of Europe, but it has by no means been confined to them. Japan and the Middle Eastern oil states have been important dealers. While these short-term lendings normally move across national frontiers, they do not directly involve foreign exchange transactions. They may, however, indirectly cause such transactions to take place.

The nature of the market is as follows: In the ordinary course of affairs, an Italian, for example, acquiring dollars—say from exports or from a legacy—would sell these dollars for his own currency. But he may decide to deposit the dollars at his bank instead, with an instruction not to sell them for cash but to repay him in dollars at a later date. Thus the bank has dollars in hand and a commitment to pay them out in, say, three months. It may then proceed to lend these dollars to another bank, anywhere in the world. Since the lending and borrowing is done in dollars, no foreign exchange transaction is directly involved. The sum total of all operations of this sort is the Eurodollar market. It is not centred on any particular place and has no formal rules of procedure or constitution. It consists of a network of deals conducted by telephone and telex around the world. U.S. residents themselves lend to and borrow from this market.

One may ask why lenders and borrowers use this market in preference to more conventional methods of lending and borrowing. Ordinarily the answer is because they can get more favourable terms, since the market works on very narrow margins between lending and borrowing rates. This involves expertise; London has played the most important part in the creation of the market. The lender hopes to get a better rate of interest than he would on a time deposit in the United States (restrictions limiting interest payable on U.S. time deposits are said to have been a contributing cause of the growth of the market during the 1960s). At the same time, normally, the borrower will find that he has to pay a lower rate than he would on a loan from a commercial bank in the United States.

Rise in Eurodollar interest rates

This has not always been the case. In 1969 Eurodollar interest rates went to very high levels. One reason for this was the set of restrictions imposed by the United States on its commercial banks lending abroad. The second was that although the prime lending rates of the principal U.S. banks might be below Eurodollar rates, many individuals, including U.S. citizens, found that they could not get loans from their banks because of the "credit squeeze."

Because this form of international lending does not involve the sale of one currency for another, it does not enter into balance-of-payments accounts. Nonetheless it may have a causal effect on the course of the exchanges. For instance, the Italian cited above might have chosen to sell his dollars had he not been tempted by the more attractive Eurodollar rate of interest. In this case, the market causes dollars not to be sold that otherwise would have been. Others who have liquid cash at their disposal for a time may even buy dollars in order to invest them in the market at short term. That would be helpful to the dollar. There are countercases. An individual who has to make a payment in dollars but lacks cash may borrow the dollars in the Eurodollar market, when otherwise he would have got credit in his own country and used that to buy dollars; in this case the market is damaging to the dollar because its existence prevents someone from buying dollars in the regular way.

Assessing the balance. To summarize, the overall balance of payments comprises the current account (merchandise and services), unilateral transfers (gifts, grants, remittances, and so on), and the capital account (long-term and short-term capital movements). If payments due in exceed those due out, a country is said to be in overall surplus; and when payments due out exceed payments due in, it is in overall deficit. The surplus or deficit must be balanced by a monetary movement in the opposite direction, and consequently the overall balance including monetary movements must always be equal.

Deficit or surplus

In practice, great difficulties have been found in assessing whether a country is in deficit or in surplus. It is often important to establish this with a view to possible corrective measures. The United Kingdom stresses the combined balance of current and long-term capital account—*i.e.*, excluding short-term capital. Such a balance, however, omits short-term movements that occur in the ordinary course of business, which may be called "normal" and which ought in principle to be included. On the other hand it is not desirable to include equilibrating or disequilibrating capital movements. These occur in consequence of a deficit (or surplus), actual or anticipated. But there may be great statistical difficulty in distinguishing between the normal short-term capital flows and those that are consequential on a surplus or deficit.

It has been noted that the overall balance, including monetary movements, must be equal, but it usually happens that the figures do not in fact balance. U.S. statisticians call the residual figure that has to be inserted to square the account "errors and omissions." If the average value of this figure over a substantial period, such as 10 years—an even longer period may have to be taken if a country is in persistent surplus or deficit—has a positive or negative value of substantial amount, then it may be taken to constitute genuine items that have escaped the statistical net. These may legitimately be included in assessing whether a country is in genuine surplus or deficit and whether corrective measures are needed.

"Errors and omissions"

The "errors and omissions" item is extremely volatile from year to year and often very large. Such movements up and down are probably caused by precautionary short-term capital movements. There have been periods when a minus item in the U.S. account was rather strikingly associated with a plus item in the U.K. account, and conversely. Accordingly, in the short term, the "errors and omissions" item should not be included in assessing whether a country is in surplus or deficit.

It has been noted that the United Kingdom stresses the balance of current and long-term capital accounts (which include unilateral transfers). The U.S. position is less clear. It traditionally published two overall balance-of-payments measures: the "Liquidity Balance" and the "Official Settlements Balance."

In distinguishing between monetary and nonmonetary items, the Liquidity Balance included any increase in the holding of short-term dollar securities abroad as part of the U.S. deficit during the period; but it did not include as counterweight any increase in short-term foreign claims held by U.S. resident banks or others (apart from official holdings). Thus, in this respect the treatment was asymmetrical. The rationale for this was precautionary. The argument was that short-term dollar assets held abroad outside the central banks might at any time be sold in the market or turned in to the central banks of the respective countries and thus constitute a drain, or the threat of a drain, on U.S. reserves. On the other hand the corresponding foreign short-term assets held by U.S. resident banks or others were not readily mobilizable by U.S. authorities for making payments. Thus by this reckoning, if during a period non-central-bank foreign holdings of short-term dollar securities and resident non-central-bank U.S. holdings of short-term foreign securities went up by an equal amount, the situation would be shown as having deteriorated, since the former class (liabilities) were a threat to U.S. reserves, while the latter class (assets) could not be mobilized by U.S. authorities to meet such a threat. Thus, though the motive for this asymmetrical treatment may have been understandable, it was statistically unsatisfactory and also unsatisfactory as a guide to corrective action. This balance is thus mainly of historical interest, and it has not been commonly used since 1971.

The U.S. Official Settlements Balance

The U.S. Official Settlements Balance reckoned an increase in non-central-bank foreign holdings of short-term dollar assets as an inflow of short-term capital into the United States; similarly an increase in U.S. resident holdings of short-term foreign assets was an outflow of short-term capital. This was a logical treatment. But the balance thus defined proved in the 1960s to be extremely volatile. This was due to large movements of funds between foreign central banks and non-central-bank foreign holders, associated with the rise of the Eurodollar market. Oscillations

of this kind do not represent changes in the fundamental balance that are needed in order to determine whether corrective measures are required. It may well be that the British method of omitting short-term capital movements altogether in the assessment of surplus or deficit is, although imperfect, the most practical available. Since exchange rates began to float in the early 1970s, the major industrial countries have paid much less attention to overall balance-of-payments measures. The current account and the trade account are the two measures that are now most commonly used in developing countries.

(R.F.H./P.Wo.)

ADJUSTING FOR FUNDAMENTAL DISEQUILIBRIUM

A "fundamental disequilibrium" exists when outward payments have a continuing tendency not to balance inward payments. A disequilibrium may occur for various reasons. Some may be grouped under the head of structural change (resulting from changes in tastes, habits, institutions, technology, etc.). A fundamental imbalance may occur if wages and other costs rise faster in relation to productivity in one country than they do in others. Imbalance may also result when aggregate demand runs above the supply potential of a country, forcing prices up or raising imports. A war may have a profoundly disturbing effect on a country's economy.

The classical view. In the traditional "classical" view no intervention by the authorities was necessary to maintain external equilibrium, except for their readiness to convert currency into gold (or silver) upon demand. The system was supposed to work automatically. If a country had a deficit, gold would flow out, and the consequent reduction in the domestic money supply would cause prices to move downward. This would stimulate exports and tend to reduce imports. The process would continue until the deficit was eliminated. Classical doctrine did not embody a clear-cut theory about international capital movements. It was usually assumed that the trade balance (more strictly, balance on goods and services) would be tailored to accommodate any capital movement that occurred. Thus, if the country was exporting capital, gold flows would cause prices to move to such a level that exports minus imports would be equal to the capital flow; equilibrium in the overall balance was automatically secured.

From laissez-faire to intervention

In due course the classical scheme of thought came under criticism. Some critics asked if an outflow or inflow of specie would necessarily have a sufficient effect on the price level to ensure an equal balance of payments. More important, a reduction in the money supply, it was pointed out, might have a side effect on the level of economic activity. Some critics went further and argued that this side effect would be stronger than the effect on prices to such a degree as to cause unemployment to rise to an undesirable level.

Contemporary views. *Monetary and fiscal measures.* The belief grew that positive action by governments might be required as well. The doctrine was first related to monetary policy in particular. The idea was that interest-rate adjustments should be combined with open-market operations by a central bank to ensure that the domestic money supply and borrowing facilities were conducive to external long-period equilibrium. After World War II the idea came to be widely held that government budget policy (usually called fiscal policy) should be brought in to assist monetary policy. For instance, if aggregate domestic demand was running so high as to cause rising prices, this should be reduced both by having a tight monetary policy and by increasing taxation more than expenditure or reducing expenditure without reducing taxation. The correct apportionment of this task between the monetary and fiscal arms is still a subject of discussion.

The question of reducing demand-deflationary policy

Nor is there yet agreement about the scope of these policies or their ability to secure fundamental equilibrium in all cases. There is probably agreement that when overall demand is running in excess of the supply potential of the economy, it should be reduced by monetary and fiscal policies. There is difference of opinion, however, as to whether the reduction of aggregate demand will bring external payments into balance in all cases. For instance, a country may have a deficit owing to some underlying economic change (such as a shift in the pattern of world trade), even if domestic demand is not above the supply potential and prices are not rising. In this case, policies designed to reduce domestic demand (commonly called deflationary policies) would cause unemployment. Some hold that, if there is an external deficit, deflationary policies should be pursued to whatever extent may be needed to eliminate the deficit. Others hold that such a policy is socially unacceptable.

Opinions differ also about how deflationary measures work to improve the external balance. Some hold that they work mainly by reducing domestic activity and thereby the amount of imported materials that a country needs and the amount of income that people can afford to spend on imported goods. If this were the whole effect of a deflationary policy, it would improve the external balance only in proportion to the amount by which it increases unemployment. Those who hold that this is the only manner in which deflation affects the external balance are especially opposed to relying on deflationary policies alone to eliminate a deficit in conditions in which aggregate domestic demand is not running above the supply potential. Some hold that a reduction of home demand also helps because it makes producers look around more eagerly for export markets (and increase their selling efforts in the home market). This appears to be doubtful, however. There is further disagreement on the extent to which deflationary policies influence the course of prices. If aggregate demand is running above the supply potential of the economy, it is highly probable that deflationary policies will slow the increase of prices and thus make a country more competitive with foreign suppliers. There is not the same agreement about the effects when demand is initially running below the supply potential of the economy. Some hold that a deflationary policy, if pushed hard enough, will infallibly slow up price increases and so help the country's external balance. Others hold that it will not, and some even argue that higher interest rates and higher taxes (weapons of deflation) can cause prices to rise. Thus, it is not absolutely clear that monetary and fiscal policies will in all cases suffice to cure an external deficit, at least without socially unacceptable results.

There is also the opposite case of countries with a trade surplus. It is clear that these countries will be unwilling to encourage policies that cause domestic prices to rise. Price inflation is a social evil and politically unpopular.

Supply and demand in surplus countries

In the case of surplus countries, the same distinction must be made between the situation in which aggregate demand is fully up to or above the supply potential of the economy and that in which it is not. In the former case a further increase in demand would almost certainly have an inflationary effect; accordingly, surplus countries in this condition will be unwilling to use monetary and fiscal policies to eliminate their external surpluses. On the other hand, if aggregate demand is running below supply potential, then a surplus country might reasonably be asked to increase aggregate demand by monetary and fiscal policies on the view that the increase will not cause inflation but will tend to remove the external surplus by inducing more imports and possibly causing producers to be less active in their selling efforts abroad.

Incomes policy. Prices may rise even when aggregate demand is not in excess of the supply potential. This may be due to wage increases and other factors. Some hold that this can be dealt with through efforts to discourage excessive wage increases by a direct approach, which may consist of a propaganda campaign on the evil effects of wage-price inflation, together with guidelines governing rates of wage increases. This direct attempt to deal with the problem is generally known as "incomes policy."

Changes in exchange rates. Exchange-rate movements work by making the products of a deficit country more price competitive or those of a surplus country less price competitive. Any program that seeks to rectify an imbalance by changing the level of prices will be effective only if demand is "price elastic." In other words, if the offer of an article at a lower price does not cause an increase in demand for it more than in proportion to the fall in price,

the proceeds from its export will fall rather than increase. Economists believe that price elasticities are sufficiently great for most goods so that price reductions will increase revenues in the long run. The outcome is not quite so certain in the short run.

Devaluation

A fast means of changing relative price levels is devaluation, which is likely to have a quick effect on the prices of imported goods. This will raise the cost of living and may thereby accelerate demands for higher wages. If granted, these will probably cause rises in the prices of domestically produced goods. A "wage–price spiral" may follow. If this spiral moves too quickly it may frustrate the intended effect of the devaluation, namely that of enabling the country to offer its goods at lower prices in terms of foreign currency. This means that if the beneficial effects of a devaluation are not gathered in quickly, there may be no beneficial effect at all.

The authorities of a country that has just devalued must therefore be especially active in preventing or moderating domestic price increases. They will need to use the other policy measures discussed above. Devaluation (or the downward movement of a flexible rate) is thus not a remedy that makes other forms of official policy unnecessary. Some have argued that, if exchange rates were allowed to float, nothing further would have to be done officially to bring the external balance into equilibrium, but this is a minority view.

Resistance to upward valuation

One further point must be made regarding exchange-rate movements. It has been found in practice that governments resist upward valuation more than they do devaluation. Under the IMF system prior to 1973, devaluations in fact were larger and more frequent than upward valuations. This had an unfortunate consequence. It meant that the aggregate amount of price inflation in deficit countries resorting to devaluation as a remedy was not offset by equivalent price decreases in the surplus countries. Therefore this system had a bias toward worldwide inflation.

Trade restrictions. Since World War II the major industrial countries have attempted to reduce interferences with international trade. This policy, by extending the international division of labour, should increase world economic welfare. An exception has had to be allowed in favour of the less-developed countries. In the early stages of the development of a country, the effectiveness and feasibility of the three types of adjustment mechanism discussed above, particularly monetary and fiscal policies, may be much less than in the more advanced countries. The less-developed countries may therefore be driven to protection or the control of imports, for lack of any other weapon, if they are to stay solvent. It has already been noted that, even in the case of a more advanced country, the effectiveness and appropriateness of the above-mentioned adjustment mechanisms are not always certain. Thus, there is no certainty that some limitation on foreign trade and on the international division of labour may not be a lesser evil than the consequences that might follow from a vigorous use of the other adjustment mechanisms, such as unemployment.

Predisposition to the free flow of trade

Restrictions on capital exports. Interference with capital movements is generally considered a lesser evil than interference with the free flow of trade. The theory of the optimum international movement of capital has not yet been thoroughly developed, but there may be a presumption in favour of absolutely free movement. The matter is not quite certain; for instance, it might be desirable from the point of view of the world optimum to channel the outflow of capital from a high-saving country into the less-developed countries, although the level of profit obtainable in other high-saving countries might be greater. Or it might be expedient to restrain wealthy individuals in less-developed countries, where domestic saving was in notably short supply, from sending their funds to high-saving countries.

While there may be good reasons for interfering with the free international flow of capital in certain cases, it is not obvious that the outflow of capital from, or inflow of capital into, a country should be tailored to surpluses or deficits in current external accounts. It may be that in some cases the sound remedy for a deficit (or surplus) is to adopt adjustment measures such as those discussed above, bearing upon current items, rather than taking the easier way of adjusting capital movements to the de facto balance on current account.

FOREIGN EXCHANGE MARKETS

The buying and selling of currencies

A foreign exchange market is one in which those who want to buy a certain currency in exchange for another currency and those who want to move in the opposite direction are able to do business with each other. The motives of those desiring to make such exchanges are various. Some are concerned with the import or export of goods between one country and another, some with the purchase and sale of services. Some wish to move capital from one area to the other, and some wish to make gifts (the latter including government aid and gifts by charitable foundations).

In any organized market there must be intermediaries who are prepared to "quote a price," in this case a rate of exchange between two currencies. These intermediaries must move the price quoted in such a way to permit them to make the supply of each currency equal to the demand for it and thus to balance their books. In an important foreign exchange market the price quoted is constantly on the move.

An exchange rate is the price of one currency in terms of another. For example, in the market for the British pound sterling (£) in exchange for U.S. dollars ($), the exchange rate might be £1 = $2. This price may also be quoted the other way around; that is, $1 = £0.50.

In a foreign exchange market, there may be a standard, government-determined price, or par value. This par value may be quoted in terms of another currency; for example, the par value of the pound was £1 = $2.80 between 1949 and 1967. In 1973 many governments abandoned their par values and let their exchange rates be determined by the forces of demand and supply. An exchange rate determined in this way, without being tied to an official par, is called a flexible or floating exchange rate; in contrast, an exchange rate is said to be pegged if the government ties it to par value.

Historically, countries often tied their currencies to gold, setting their official parities in terms of that metal. Under this historical gold standard, the gold equivalence of currencies determined exchange rates. For example, the British pound was worth 4.86 times as much gold as the U.S. dollar during the period prior to World War I. The exchange rate remained at or quite close to the mint parity of £1 = $4.86. Nobody would pay much more than $4.86 for a British pound, nor take much less.

Bimetallism

Historically, there were also periods of bimetallism, when the gold standard was combined with a silver standard, and currencies were fixed in terms of both gold and silver. The bimetallic standard was given up by most of its adherents (the United States, France, Italy, Switzerland, The Netherlands, and Belgium) in the 1870s.

THE GOLD STANDARD

If the demand by those holding a particular currency, say sterling, for another currency, say the dollar, exceeds the demand of dollar holders for sterling, the dollar will tend to rise in the foreign exchange market. Under the gold standard system there was a limit to the amount by which it could rise or fall. If a sterling holder wanted to make a payment in dollars, the most convenient way for him to procure the dollars would be in the foreign exchange market. But under the gold standard he had another option; *i.e.,* he had a legal right to obtain gold from the authorities in exchange for paper currency at the established par value of that currency and remit the gold to the other country, where he would have a legal right to obtain its currency in exchange for bars of gold at the official valuation. Thus, it would not be advantageous for a sterling holder to obtain dollars in the foreign exchange market if the quotation for a dollar there exceeded parity by more than the cost of remitting gold. The exchange rate at which it became cheaper to remit gold rather than use the foreign exchange market was known as the "gold-export point." There was also a "gold-import point" determined on similar lines.

Most of those seeking dollars, however, did not under-

take to remit gold even if the dollar quotation was at the gold-export point. The remission of gold was handled by arbitrageurs. These are people who buy and sell currencies simultaneously on different exchanges in order to profit by small differences in the quoted rates. Their action would reduce the supply of sterling, since they would be selling sterling for gold to the British authorities, and increase the supply of dollars, since they would acquire dollars in exchange for gold from the U.S. authorities. The arbitrageurs would carry out these operations to the extent needed to prevent the scarcity of the dollar from raising its sterling price above the gold-export point for the United Kingdom, and conversely. At the same time, the gold reserve of the British authorities would be diminished, and the gold reserve of the U.S. authorities increased.

The arbitrageurs

The international gold standard provided an automatic adjustment mechanism, that is, a mechanism that prevented any country from running large and persistent deficits or surpluses. It worked in the following manner. A country running a deficit would see its currency depreciate to the gold-export point. Arbitrage would then result in a gold flow from the deficit to the surplus country. In other words, the deficit would be settled in gold.

The gold flow had an effect on the money system. When gold flowed into the banking system of the surplus country, its money stock rose as a consequence. On the other side, when a deficit country lost gold, its money stock fell. The falling money stock caused deflation in the deficit country; the rising money stock caused inflation in the surplus country. Thus, the goods of the deficit country became more competitive on world markets. Its exports rose, and its imports declined, correcting the balance-of-payments deficit.

Although this adjustment process worked automatically, it was not problem-free. The adjustment process could be very painful, particularly for the deficit country. As its money stock automatically fell, aggregate demand fell. The result was not just deflation (a fall in prices) but also high unemployment. In other words, the deficit country could be pushed into a recession or depression by the gold standard. A related problem was one of instability. Under the gold standard, gold was the ultimate bank reserve. A withdrawal of gold from the banking system could not only have severe restrictive effects on the economy but could also lead to a run on banks by those who wanted their gold before the bank ran out.

The gold standard during the Great Depression

These twin problems materialized during the Great Depression of the 1930s; the gold standard contributed to the instability and unemployment of that decade. Because of the strains caused by the gold standard, it was gradually abandoned. In 1931, faced with a run on its gold, Britain abandoned the gold standard; the British authorities were no longer committed to redeem their currency with gold. In early 1933 the United States followed suit. Although the tie of the dollar to gold was partially restored at a later date, one very important feature of the old gold standard was omitted. The public was not permitted to exchange dollars for gold; only foreign central banks were allowed to do so. In this way the U.S. authorities avoided the risk of a run on their gold stocks by a panicky public.

THE INTERNATIONAL MONETARY FUND

The International Monetary Fund (IMF), founded at the Bretton Woods Conference in 1944, is the official organization for securing international monetary cooperation. It has done useful work in various fields, such as research and the publication of statistics and the tendering of monetary advice to less-developed countries. It has also conducted valuable consultations with the more developed countries.

Drawing Rights

Of particular interest to this discussion is the Fund's system of Drawing Rights, which permits countries in temporary deficit to draw supplies of foreign currency according to predetermined quotas. These extra supplies of currency give a country more time in which to adjust its balance of payments and so avoid taking unsound or unneighbourly measures like import restrictions for lack of enough reserves to tide it over a difficulty. The mechanism is as follows: members of the Fund are required to make initial deposits according to their quotas, which are based on the country's national income, monetary reserves, trade balance, and other economic factors. Quotas are payable partially in Special Drawing Rights (see below *Special Drawing Rights*) and partially in a country's own currency. A country's quota closely approximates its voting power, the amount of foreign exchange it may purchase (Drawing Rights), and its allotment of Special Drawing Rights. The Fund makes its stock of members' currencies available to member countries that wish to draw upon their quotas. When creditor countries are presented with their own currencies previously deposited by them with the Fund, they are obliged to take them in final discharge of debts owed by other member countries. Since they previously deposited these currencies themselves they are in effect getting nothing from the debtor countries in respect of the debts owed to them, and their willingness to accept payment in this way is their contribution to the overall liquidity of the world system. Later the creditor countries may themselves become debtors and partake of the benefits. The debtors have to repay the Fund usually in three to five years. A country with more serious financial problems may draw as much as 140 percent of its quota during a three-year period, and repayment must be made between four to 10 years afterward.

The exercise of Drawing Rights is subject to discussion and sometimes to conditions, except for drawings on what are called the reserve tranches (sums equal to the member's original deposits in its own currency and Special Drawing Rights), which are given "the overwhelming benefit of the doubt." Countries are also free to draw without discussion up to the net amount to which they have previously been drawn upon by other countries.

IMF country quotas

The quotas paid by members of the IMF are the primary source of income for the organization. Quotas for member countries are periodically reviewed and reevaluated according to the country's financial situation. General increases in quotas normally occur following the periodic reviews, although special reviews and increases sometimes occur for specific countries, such as Saudi Arabia in 1981. The IMF also borrows to supplement its quota resources. In 1981, for example, Saudi Arabia agreed to loan the Fund more than $8,000,000,000 over a two-year period, and an additional $1,300,000,000 was loaned by a group of countries. Between 1976 and 1980 about one-third of the Fund's gold holdings were sold at public auction to benefit the member developing countries. More than $4,600,000,000 was received from the gold sale; part of the revenue was made available to members according to their quotas, and part of the revenue was placed in a trust fund to dispense low-interest loans to developing countries. (R.F.H./Ed.)

The International Monetary Fund as it finally emerged from the wartime discussions was a much more modest undertaking than had originally been conceived by the British. An early British proposal would have required creditor countries to receive payment in paper money up to the total amount of all the quotas of all the debtor countries. This seemed to many to be more than it was fair to ask creditors to do. The United States claimed that for a number of years after the war it was likely to be in credit against the whole of the rest of the world, and so it was. Under the British plan they would have had to give an unconscionably large amount of credit, with no certainty of repayment. At that time it did not seem at all likely that the United States would ever go into deficit, which, of course, it eventually did.

THE IMF SYSTEM OF PARITY (PEGGED) EXCHANGE RATES

The modified gold standard

When the IMF was established toward the end of World War II, it was based on a modified form of the gold standard. The system resembled the gold standard in that each country established a legal gold valuation for its currency. This valuation was registered with the International Monetary Fund. The gold valuations served to determine parities of exchange between the different currencies. As stated above, such fixed currencies are said to be pegged to one another. It was also possible, as under the old gold standard, for the actual exchange quotation to deviate somewhat on either side of the official parity. There was

agreement with the International Monetary Fund about the range, on either side of parity, within which a currency was allowed to fluctuate.

But there was a difference in the technical mode of operation. The service of the arbitrageurs in remitting physical gold from country to country as needed was dispensed with. Instead the authorities were placed under an obligation to ensure that the actual exchange rates quoted within their own territories did not go outside the limits agreed upon with the International Monetary Fund. This they did by intervening in the foreign exchange market. If, for instance, the dollar was in short supply in London, the British authorities were bound to supply dollars to the market to whatever extent was needed to keep the sterling price of the dollar from rising above the agreed-upon limit. The same was true with the other currencies of the members of the International Monetary Fund. Thus, the obligation of the monetary authorities to supply the currency of any Fund member at a rate of exchange that was not above the agreed-upon limit took the place of the obligation under the old gold standard to give actual gold in exchange for currency.

It would be inconvenient for the monetary authorities of a country to be continually watching the exchange rates in its market of all the different currencies. Most authorities confined themselves to watching the rate of their own currency against the dollar and supplying from time to time whatever quantity of dollars might be required. At this point the arbitrageurs came into service again. They could be relied upon to operate in such a way that the exchange rates between the various currencies in the various foreign exchange markets could be kept mutually consistent. This use of the dollar by many monetary authorities caused it to be called a currency of "intervention."

The official fixing of exchange rates as limits on either side of parity, outside of which exchange-rate quotations were not allowed to fluctuate, bears a family resemblance to the gold points of the old gold standard system. The question naturally arose why, in devising a somewhat different system, it was considered desirable to keep this range of fluctuation. In the old system it arose necessarily out of the cost of remitting gold. Since there was no corresponding cost in the new system, why did the authorities decide not to have a fixed parity of exchange from which no deviation would be allowed? The answer was that there was convenience in having a range within which fluctuation was allowed. Supply and demand between each pair of currencies would not be precisely equal every day. There would always be fluctuations, and if there were one rigidly fixed rate of exchange the authorities would have to supply from their reserves various currencies to meet them. In addition to being inconvenient, this would require each country to maintain much larger reserves than would otherwise be necessary.

Under a system of pegged exchange rates, short-term capital movements are likely to be equilibrating if people are confident that parities will be maintained. That is, short-term capital flows are likely to reduce the size of overall balance-of-payments deficits or surpluses. On the other hand, if people expect a parity to be changed, short-term capital flows are likely to be disequilibrating, adding to underlying balance-of-payments deficits or surpluses.

Methods of currency dealing

Equilibrating short-term capital movements. Commercial banks and other corporations involved in dealings across currency frontiers are usually able to see some (but not necessarily all) of their needs in advance. Their foreign exchange experts will watch the course of the exchanges closely and, if a currency is weak (*i.e.,* below parity), advise their firms to take the opportunity of buying it, even if somewhat in advance of need. Conversely, if the currency is above parity but not expected to remain so indefinitely, they may recommend postponing purchases until a more favourable opportunity arises. These adjustments under the influence of common sense and self-interest have an equilibrating influence in foreign exchange markets. If a currency is temporarily weak, it is presumably because of seasonal, cyclical, or other temporary factors. If on such an occasion private enterprise takes the opportunity to buy the currency while it is cheap, that tends to bring demand up to equality with supply and relieves the authorities from the need to intervene in order to prevent their currency from falling below the lower point whenever there is a temporary deficit in the balance of payments. As previously noted, when confidence in the fixed parity exchange rate drops and market participants expect a change in parity, short-term capital movements may be disequilibrating. (See below *Disequilibrating capital movements.*)

Equilibrating by short-term rate movements

Another equilibrating influence arises from the movements of short-term interest rates. When the authorities have to supply foreign currencies in exchange for the domestic currency, this causes a decline in the money supply in domestic circulation—unless the authorities deliberately take offsetting action. This decline in the money supply, which is similar to that occurring under the gold standard, tends to raise short-term interest rates in the domestic money market. This will bring an inflow of money from abroad to take advantage of the higher rates or, what amounts to the same thing, will discourage foreigners from borrowing in that country's money market since borrowing will have become more expensive. Thus, the interest-rate differential will cause a net movement of short-term funds in the direction required to offset the temporary deficit or, in the opposite case, to reduce a temporary surplus that is embarrassing to others. It must be stressed again that this equilibrating interest-rate mechanism implies confidence that the parity will not be altered in the near future.

The helpful movement of interest rates may be reinforced by action of the monetary authorities, who by appropriate open-market operations may cause short-term interest rates to rise above the level that they would have attained under market forces and thus increase the equilibrating movement of short-term funds. The Bank of England provided the most notable example of the smooth and successful operation of this policy under the old gold standard during many decades before World War I.

Forward exchange. The transactions in which one currency is exchanged directly for another are known as spot transactions. There can also be forward transactions, consisting of contracts to exchange one currency for another at a future date, perhaps three months ahead, but at a rate determined now. For instance, a German firm may have a commitment to pay a U.S. firm in dollars in three-months' time. It may not want to take the risk that the dollar will rise relative to the mark during the three months, so that it would have to surrender more marks in order to honour its commitment. It could of course buy the dollars right away and thus obviate this risk, but it may not have any spare cash and borrowing may be inconvenient. The firm has the alternative of buying dollars at a rate agreed upon now for which it does not have to surrender marks until three months have passed. Some firms have a regular routine procedure for covering all future commitments to be paid for in a foreign currency as soon as these are entered into. Of course, even a firm that does this may combine its routine procedure with a little judgment, for instance, if there are good reasons for believing that the foreign currency will become cheaper during the relevant period. And firms with multinational commitments will vary the distribution of their assets among different currencies in accordance with changing conditions. The forward-exchange rate will, like the spot rate, be continually varying. It is not usually identical with the spot rate but in normal times has a regular relation to it. This relation is determined as follows:

Forward rates

Dealers in forward exchange usually balance their commitments; for instance, a contract to deliver forward marks can be offset against one to deliver forward dollars, and nothing more has to be done about it. If a particular dealer cannot manage this he will be in communication with another who may be in the opposite position. It may not, however, always be possible to offset every transaction. If this is not done, the dealer must make a spot purchase of the currency—say marks—in excess demand in the forward market. If he did not do this he would risk an exchange loss on some of his forward transactions. For the purpose of evaluating the forward-exchange rate to be asked in a particular deal, it is always correct to suppose that the deal is one that cannot be offset. If the dealer

has to purchase marks on the spot, he can earn the rate of interest prevailing in Frankfurt until the time comes when he has to deliver the marks. Whether this is advantageous or not depends on whether the rate of interest in Frankfurt is higher or lower than that in New York City. If it is higher in Frankfurt, the dealer will normally quote a rate per forward mark that is lower than the spot rate; but if the rate of interest in Frankfurt is lower, then the forward mark will normally stand above the spot mark to compensate the dealer for having to employ his liquid funds in a less remunerative market. When the relation of the forward rate to the spot rate is determined by a comparison of the short-term interest rates in the two centres in the manner just described, the forward rate is said to be at "interest parity."

The question arises as to what particular interest rates are used to calculate the interest parity. There is a variety of practice. In previous times the rate of interest on U.S. Treasury bills and the rate of interest on British Treasury bills were used to determine the interest parity of the sterling price for forward dollars. More recently the interest rates on Eurodollars and Eurosterling have been used—that is, the interest on dollar and sterling accounts held by European banks.

In normal times arbitrage may be expected to hold forward rates to their interest parities. There have been times, and even rather prolonged periods, in which the forward rate for a currency has fallen below (or risen above) its interest parity. This may happen when there is a large one-way movement of funds (such as when there is a lack of confidence in a particular currency). In some cases, such as a simultaneous multiple swapping of currencies, the arbitrager does not have to commit any funds, but in forward arbitrage funds have to be committed for a period of three months. It is true that an arbitrageur who had bought three-months' sterling could resell the sterling before the three months had elapsed, but if he did so he might have to accept a loss. If the one-way movement is very heavy there may be a shortage of funds available for forward arbitrage. Nonetheless the demand for forward sterling has to be kept equal to the supply of it, and if there is insufficient arbitrage for this purpose then a positive profit has to appear on the purchase of forward sterling; in other words, its price has to fall below the interest parity.

Dealer problems in balancing contracts

If dealers in a forward currency cannot offset contracts for sale with contracts for purchase and find an excess of customers wishing to sell, the excess supply causes immediate pressure on the spot market, since arbitrageurs and others who supplement the forward demand for the weak currency must cover their positions by selling an equivalent amount spot. The only way in which the authorities can prevent an excess offer of their currency forward from causing an immediate drain on their reserves is by offering to buy it forward themselves, without simultaneously selling it spot. British authorities engaged in such operations during periods when sterling was weak, and similar operations have been conducted by other central banks in connection with swap agreements for mutual accommodation.

The foregoing descriptions of the equilibrating movements of short-term funds have not applied when there has been a serious lack of confidence that a given parity will be maintained. Occasions of lack of confidence occurred much more frequently under the modified gold standard (International Monetary Fund) than they did under the old gold standard. The reason for this is simple. Under the old gold standard it was not expected that a country of good standing would alter the gold valuation of its currency (although in much earlier days "debasement" was common enough). A devaluation of the official gold content was regarded as not far removed from a declaration of bankruptcy, and it was assumed that a country would avoid it at all costs and in all times short of a major war or revolution.

The IMF solution to disequilibrium

Under the International Monetary Fund this position was altered quite deliberately to allow a country whose payments were in "fundamental disequilibrium," to propose a change of parity. This remedy was proposed at the Bretton Woods Conference (1944), which set up the International Monetary Fund, because it was thought to be better than alternative remedies, such as domestic deflation.

Disequilibrating capital movements. Whatever its merits from a long-term point of view, the idea that it is quite respectable for a country to alter the par value of its currency in certain circumstances had disturbing effects on the movements of short-term funds—effects that may not have been clearly foreseen at the time of Bretton Woods. Such movements of funds were sometimes very large indeed. These movements were not equilibrating, like those described in relation to a parity in which there is confidence; on the contrary, they were disequilibrating. If a currency became weak—if the demand for it fell below the supply—this could give rise to the idea that the authorities having the weak currency might in due course decide to devalue it, as they were perfectly entitled, under International Monetary Fund principles, to do.

Covering. Foreign exchange advisers to corporations had to watch for such possibilities and propose a readjustment of assets entailing a movement out of the weak currency. It was not necessary that there be, on an objective assessment, a probability (more than a 50 percent chance) that the currency in question had to be devalued. To provoke a disequilibrating movement of funds it was enough that there should be a small chance (much less than 50 percent) that it would be devalued. In strict theory, funds should be moved out of a given currency whenever the probability that it will be devalued outweighs the cost of moving the funds.

Forms of covering

If a firm or its affiliate has foreseeable commitments to make payments in a currency other than that of the area in which it operates, it may think it wise to "cover" its position by buying the currency at once, in either the spot or the forward market. Covering may take other forms also. If a contract to pay abroad is in the currency of the home, or paying, country, then the prospective foreign receiver of these funds will have to consider whether he should not cover his own position by selling the currency of the paying country forward. Payments in the opposite direction have also to be considered. If these are in the currency of the home country, the foreigner due to make the payment will consider whether he should cover his position by buying the currency of the home country forward. If the payment is in the foreign currency, then the firm in the home country due to receive it will consider whether to cover itself by selling the foreign currency forward. Thus, there are four main classes of covering. In normal times it is probable that not all positions are covered in these four ways, although it is not impossible that they should be.

Effect of possible devaluation

If a suspicion arises that a particular currency, say that of the home country, may be devalued, then the position is radically changed. The following arguments apply in reverse to the case when it is believed that a particular currency may be valued upward. It is necessary to go through the four classes of cases. Members of the home country who normally cover their commitments to make payments in a foreign currency would clearly continue to do so. And those, if any, who do not habitually do so would be strongly advised to do so when there is a possibility that the home currency may be devalued. To take the second case—that of outward payments to be made in the home currency—the same applies: foreigners who normally sell it forward should continue to do so, and those who do not normally sell it forward would be strongly advised to do so lest the currency be devalued before the payment is made. Coming to the payments due to the home country, in the case of those to be made in the home currency, the foreigners who normally cover themselves by buying forward or spot should be advised to cease doing so immediately, since they may get the currency cheaper before the payment has to be made. Thus, in this case the fear of devaluation causes those concerned to stop covering their positions. The same applies to inward payments to be made in foreign currencies; residents of the home country would be advised to cease from such covering, since in the interval their currency may be devalued, and therefore it would be foolish to sell the foreign currency due to come, in advance of payment.

Thus, the prospect of devaluation may cause both addi-

tional covering and uncovering. Both types of change are adverse to the currency under suspicion. It is notable that the total value of the appropriate covering plus that of the uncovering when a currency becomes suspect is independent of the proportion of positions that are normally covered. If all positions are normally covered then the adverse effect will consist of an uncovering of about half of all positions. If all positions are not normally covered, then the adverse effect will be equal to the sum of the amount of extra covering and the amount of uncovering. The movement of funds under these heads can be very large in relation to a country's normal balance of incoming and outgoing payments. It makes no difference whether the changed action by the firms relates to the spot or to the forward markets. This is because, when there is a big one-way movement in the forward market, the whole of it is thrown, through the actions of the dealers, arbitrageurs, and the like, onto the spot market.

Hedging. Whereas the word "covering" relates to payments foreseen or possible, the term hedging is used for operations related not to prospective payments but to existing assets. Thus, a non-British firm may need to have a sterling balance for an indefinite period ahead. It may think it desirable in this case to protect its position against the possibility of sterling being devalued in the near future by selling sterling forward at the existing quoted rate. If sterling is devalued before the forward contract matures, the operator will get a foreign currency—say the franc—at the old rate and can rebuy sterling at a cheaper rate. The profit that he makes recoups him for the loss in the franc value of his sterling due to the devaluation. If there is no devaluation he can renew his hedge at the date due, if sterling is still suspect, or he can terminate it without loss except for the actual cost, or service charge, of the hedging transaction.

Hedging as protection of real assets

An even more important use of hedging is to protect the international value of real assets such as securities, real estate, and industrial buildings and plants. If a non-British person conducts business and has assets in Britain, he may think it wise to protect the international value of these assets by selling a certain amount of sterling forward. A devaluation, if it occurs, will reduce the foreign exchange value of the sterling assets; but the profit that the owner makes from selling sterling forward and buying it back at a cheaper rate will be an offset to this loss.

Speculation. The movements so far considered are of a precautionary nature. It is sometimes suggested, when there is a big movement of funds out of a currency, that those prompting it are actuated by some motive hostile to the suspect currency. This is usually quite wrong. Such large movements of funds are often referred to incorrectly as "speculative." This gives a false impression of what is happening. Speculation can, and often does, occur when a currency becomes suspect; but the word speculative should be confined to movements of funds made not to protect positions but purely in the hope of gain. A person may believe that the Deutsche Mark is likely to be valued upward and decide to buy Deutsche Marks, not because he has any commitments denominated in Deutsche Marks but because he wants to resell them afterward at a profit. He will probably buy the Deutsche Marks forward. Such speculation plays only a minor role in the early movements of funds in anticipation of a change of parity. It may, however, mount up very strongly in the last stages when an upward or downward revaluation has become almost certain.

A big outward movement of funds may precipitate a change of parity, desirable or undesirable in itself, simply because there are not enough reserves to finance the withdrawals. Even if the country in trouble is assisted by international credits, in certain cases these may not be large enough to avert the need for devaluation. A great movement of funds from a particular country may occur because it is thought likely that it will have to devalue. There may also be a great movement into a country thought likely to value upward. The latter kind of movement will cause difficulties for other countries, since the funds must come from somewhere. This adverse effect may be concentrated on one other currency, as in the classic crisis centred on a possible upward valuation of the Deutsche Mark in November 1968, where the drain was mainly from the French franc; or it may be more widely diffused, as in the crisis of the mark in September 1969.

Stresses in the IMF system. The International Monetary Fund system of pegged-but-adjustable exchange rates came under increasing pressures during the 1960s. The system suffered from three major, interrelated problems: inadequate adjustment, confidence, and liquidity. Changes actually made in exchange rates were inadequate to deal with the major disturbances occurring in international payments. Because the adjustment mechanisms in the system were inadequate, a number of countries ran large and persistent imbalances in their international payments. This led to a lack of confidence that existing par values could be maintained and to periodic speculative rushes into strong currencies and away from weak ones. Deficit countries were not in a position to meet large speculative attacks because of their limited quantities of liquid reserves.

Creating international reserves

Traditionally, there had been two major methods of international reserve creation: the mining of gold and the acquisition of reserves in the form of key currencies (mainly dollars). Gold mining did not keep up with the rapid increase in international trade; gold reserves became less and less adequate as a means for covering balance-of-payments deficits. The alternative method for acquiring reserves—the accumulation of U.S. dollars by central banks—had one major disadvantage. For countries such as the United Kingdom, West Germany, or Brazil to accumulate dollars, the United States had to run a balance-of-payments deficit. But when the United States ran large deficits, doubts arose regarding the ability of the United States to maintain the convertibility of the dollar into gold. In other words, there was a fundamental inconsistency in the design of the IMF system, which created something of a paradox: if the United States did run large deficits, the dollar would sooner or later be subject to a crisis of confidence; if it did not run large deficits, the rest of the world would be starved for dollar reserves.

Special Drawing Rights. To deal with the inability of the existing system to create an adequate quantity of reserves without requiring the United States to run large deficits, a new kind of reserve called Special Drawing Rights (SDRs) was devised by the International Monetary Fund. Members of the Fund were to be allocated SDRs, year by year, in prearranged quantities to be used for the discharge of international indebtedness. At the IMF meeting in 1969, agreement was reached for an issue extending over three years. These Special Drawing Rights differed from ordinary Drawing Rights in three important respects: (1) The use of Special Drawing Rights was not to be subject to negotiations or conditions. (2) There was to be only a very much modified form of repayment obligation. A member who used more than 70 percent of all the Special Drawing Rights allotted in a given period had to repay to the extent needed to reduce its average use of the rights during that period to 70 percent of the total. Thus, 70 percent of all Special Drawing Rights issued could be thought of as reserves in the fullest sense, since a member who limited its use to this amount would have no repayment obligation. (3) In the case of Drawing Rights, the Fund uses currencies as subscribed by members to provide the medium of payment. By contrast, the Special Drawing Rights were to be accepted in final discharge of debt without being translated into any particular currency. Though currencies would still have to be subscribed by members receiving Special Drawing Rights, these would be in the background and would not be used, except in the case of a member in net credit on Special Drawing Rights account who wished to withdraw from the scheme.

SDRs versus Drawing Rights

Initially, the total amount of Special Drawing Rights allocated was equivalent to more than U.S. $9,000,000,000, but additional allocations to IMF members during the 1970s more than doubled the total. The value of the Special Drawing Rights is based on the currencies of the largest exporting IMF members. The use of SDRs was altered and expanded in 1978, allowing agencies other than the IMF to use SDRs in monetary exchange. Subsequently SDRs have been used by the Andes Reserve Fund, the

Arab Monetary Fund, the Bank for International Settlements, and others.

Other efforts at financial cooperation. *The Group of Ten.* As early as 1961 there were signs of a crisis in the IMF system. The United States had been running a heavy deficit since 1958, and the United Kingdom plunged into one in 1960. It looked as if these two countries might need to draw upon continental European currencies in excess of the amounts available. Per Jacobssen, then managing director of the IMF, persuaded a group of countries to provide standby credits amounting to $6,000,000,000 in all, so that supplementary supplies of their currencies would be available. The plan was not confined to the countries that happened to be in credit at that time but was extended to other important countries, the currencies of which might run short at some future time. This plan was known as the "General Arrangements to Borrow." The adhering countries were 10 in number: the United States, the United Kingdom, Canada, France, West Germany, Italy, The Netherlands, Belgium, Sweden, and Japan. They became known as the "Group of Ten."

Problems of the 1960s

The arrangement was subject to the agreement that countries actually supplying additional currency would have the right to take cognizance of how the Fund used it. This put them in a power position as against the International Monetary Fund itself. Since then the Group of Ten has worked together in deliberating on international monetary problems.

The dominant position gained by the Group of Ten has been due not only to their provision of standby credit but also to the manner in which they do their business. The ultimate authority of the Group resides in the finance ministers of the countries concerned, who meet from time to time. Their deputies meet more frequently for detailed work on particular problems. These deputies consist of high-ranking persons in their respective treasuries and central banks; they are resident in their own countries and have day-to-day knowledge of their problems and of what is politically feasible. In this respect they are in a much more advantageous position than the executive directors of the International Monetary Fund, who live in Washington, D.C., and have less contact with their home governments; they also tend to be persons of higher standing and authority.

The Basel Group. In 1930 a Bank for International Settlements was established at Basel, Switz.; its main duty was to supervise and organize the transfer of German reparations to the recipient countries. This "transfer problem" had caused much trouble during the 1920s. There may also have been a hope in the minds of some that this institution might one day develop into something like a world central bank.

Not long after it was set up the Germans gained a moratorium on their reparations payments. By then, however, the Bank for International Settlements had become a convenient place for the heads of the European central banks to meet together and discuss current problems. This practice was resumed after the war, and the United States, although not a member, was invited to join in the deliberations.

The European Payments Union

When Marshall Plan aid was furnished by the United States to help European countries in their postwar reconstruction, a European Payments Union was established to facilitate multilateral trade and settlements in advance of the time when it might be possible to reestablish full multilateralism on a world scale. The war had left a jumble of trade restrictions that could not be quickly abolished. The European Payments Union also contained a plan for the provision of credit to European debtors. The United Kingdom was a member, and with it was associated the whole sterling area. Responsibility for working the machinery of the European Payments Union was assigned to the Bank for International Settlements. The European Payments Union was ultimately wound up after the countries of Europe were able to eliminate the last restrictions and make their currencies fully convertible in 1958.

In January and February 1961 there was a serious sterling crisis, due partly to the British deficit of 1960 and partly to a large movement of funds in anticipation of an upward valuation of the West German mark, which happened, and thereafter in anticipation of a second upward valuation, which did not happen at that time. To help the British, the Basel Group of central banks provided substantial credits. These were liquidated when the United Kingdom transferred its indebtedness to the International Monetary Fund the following July. The Basel Group has provided further credits from time to time. The problems involved have continued to be discussed at the monthly meetings.

Sterling devaluation of 1967

The arrangement made for the support of the sterling area in 1968 is noteworthy. After the devaluation of sterling in 1967 it was feared that the monetary authorities of the countries composing the sterling area might wish to reduce their holdings of sterling. Because there was a continuing problem of world liquidity and sterling played an important part as a reserve currency, the international consensus was that any substantial reduction in the holding of sterling as a reserve currency would be damaging to the international monetary system. Under the arrangement made in 1968 the United Kingdom on its side agreed to give a dollar guarantee to the value of the greater part of the sterling-area reserves; there were slightly different arrangements with each monetary authority. On its side the Bank for International Settlement agreed to organize credits to finance payments deficits for some countries of the sterling area, should these occur at times when the United Kingdom might find it difficult to handle them.

The OECD. The Organisation for European Economic Co-operation (OEEC) was set up in 1948 to make arrangements for the distribution of Marshall Aid among the countries of Europe. When its tasks in this connection were accomplished, it remained in existence, was broadened to include the United States, Canada, and Japan, and it was renamed the Organisation for Economic Co-operation and Development (OECD). It has a permanent staff and headquarters in Paris. It undertakes research on a substantial scale and affords a forum for the discussion of international economic problems. The Working Party No. 3 of the organization's Economic Committee, which is concerned with problems of money and exchange, has made significant contributions; it issued a very important report on balance-of-payments adjustment problems in 1966. At times the personnel of the Working Party has been much the same as that of the deputies of the Group of Ten. The Organisation for Economic Co-operation and Development has also set up an organization called the Development Assistance Committee, concerned with problems of assistance to the developing countries.

Working Party No. 3

Swap agreements. The informal system of swap agreements provides a mutual arrangement between central banks for standby credits designed to see countries through difficulties on the occasions of large movements of funds. These are intended only to offset private international flows of capital on precautionary or speculative account, not to finance even temporary deficits in countries' balance of payments. Arranged ad hoc and informally, they depend on the mutual goodwill and trust of the central banks involved. The system of credits, although informal, must be reckoned as important, because they are of large amount. (R.F.H./P.Wo.)

The end of pegged exchange rates. *The crisis of the dollar.* The monetary system established by the IMF in 1944 underwent profound changes in the 1970s. This system had assumed that the dollar was the strongest currency in the world because the United States was the strongest economic power. Other countries were expected to have difficulty from time to time in stabilizing their exchange rates and would need assistance in the form of credits from the International Monetary Fund, but the dollar was expected to remain stable enough to function as a substitute for gold in international transactions. In the second half of the 1960s these assumptions came into question. The war in Vietnam led to inflation. The flood of dollars into other countries caused difficulty for the European central banks, which were forced to increase their dollar holdings in order to maintain their currencies at the established exchange rates. As the flood continued in 1971, the West German and Dutch governments decided to let their currencies float—that is, to let their exchange

End of Bretton Woods system

rates fluctuate beyond their assigned parities. Austria and Switzerland revalued their currencies upward in relation to the dollar. These measures helped for a time, but in August the outflow of dollars resumed, and on August 15 Pres. Richard M. Nixon formally suspended the U.S. commitment made in 1934 to convert dollars into gold, effectively ending the postwar monetary system established by the IMF. Most of the major trading countries decided to abandon fixed exchange rates temporarily and temporarily let their currencies find their own values in relation to the dollar.

The Smithsonian Agreement and after. On Dec. 17 and 18, 1971, representatives of the Group of Ten met at the Smithsonian Institution in Washington, D.C., and agreed on a realignment of currencies and a new set of pegged exchange rates. The dollar was devalued in terms of gold, while other currencies were appreciated in terms of the dollar. On the whole, the dollar was devalued by nearly 10 percent in relation to the other Group of Ten currencies (those of the United Kingdom, Canada, France, West Germany, Italy, The Netherlands, Belgium, Sweden, and Japan). Several months after the Smithsonian Agreement, the six members of the European Economic Community (EEC) agreed to maintain their exchange rates within a range of 2.25 percent of parity with each other.

The Smithsonian Agreement proved to be only a temporary solution to the international currency crisis. A second devaluation of the dollar (by 10 percent) was announced in February 1973, and not long afterward Japan and the EEC countries decided to let their currencies float. At the time, these were thought of as temporary measures to cope with speculation and capital shifts; it was, however, the end of the system of established par values.

(R.F.H./F.S.P./Ed.)

FLOATING EXCHANGE RATES

Inadequacy of the adjustable pegs system

The floating exchange-rate system emerged when the old IMF system of pegged exchange rates collapsed. The case for the pegged exchange rate is based partly on the deficiencies of alternative systems. The IMF system of adjustable pegs proved unworkable in a world in which there were huge volumes of internationally mobile financial capital that could be shifted out of countries in balance-of-payments difficulties and into the stronger nations. The earlier gold standard system had likewise contained substantial defects. Under some circumstances, it required countries to go through a painful deflation. The gold standard, it is widely held, made the Great Depression of the 1930s even deeper than it might otherwise have been.

Three major, interrelated hopes were expressed when flexible exchange rates replaced the collapsing IMF system of pegged exchange rates in the early 1970s. First, flexible exchange rates would allow currencies to hold at or near their fundamental equilibrium values; national authorities would not feel obliged to defend exchange rates that were severely out of line. Second, deficit countries would be able to reestablish their international competitiveness without going through the painful deflationary process required by the old gold standard and without facing the political embarrassment of abandoning an established par value. Finally, the national monetary authorities would have a substantial degree of independence to pursue the most appropriate domestic monetary and fiscal policies, without being severely constrained by balance-of-payments pressures. In practice, exchange-rate flexibility turned out to be more complicated than its proponents had anticipated.

Exchange-rate fluctuations. The pegged exchange-rate system collapsed in two speculative flurries against the U.S. dollar in 1971 and 1973. In each case, the dollar depreciated about 10 percent in terms of an average of other currencies. (In calculating an average exchange rate for the dollar, the currencies of each other nation is weighted according to the volume of trade of that nation with the United States.) After these initial adjustments, exchange rates of the major trading nations were generally quite stable for the next four years (late 1973–77), although there were some fluctuations. The dollar strengthened following the first oil shock, which occurred in 1973–74; because the United States still produced most of the oil it consumed, it was expected to be less severely shaken by high oil prices than would its major trading partners, especially West Germany and Japan. In the 1973–77 period, the major exchange-rate change was a fall in the British pound sterling by about 30 percent when measured in terms of dollars.

Period of dollar decline

In late 1977 the dollar entered a period of instability. As the U.S. economy expanded and inflation increased, U.S. goods became less competitive on world markets. In response, the dollar began to slide downward. This raised the price of imported goods in the United States, adding to inflationary pressures. The United States seemed in danger of entering a wage–price–exchange rate spiral. Anticipating worse to come, speculators began to unload dollars, moving the exchange value of the dollar even lower and adding to the inflationary spiral. During the 1977–79 period, the average exchange value of the dollar declined by about 15 percent.

Faced with a rapidly deteriorating situation, the United States tightened its domestic policies sharply. In particular, monetary policy was tightened in order to combat the rapid inflation. This experience provided one early, important lesson about flexible exchange rates. Even though flexible exchange rates provide some independence for domestic monetary policies, domestic policies cannot be made without concern for international complications. This is true even for a large, prosperous economy, such as that of the United States.

Dollar surge in the 1980s

During the late 1970s, the U.S. dollar was threatened with a collapse. By the mid-1980s the opposite had occurred: the dollar had soared—rising about 80 percent. A number of forces contributed to this rise. One was U.S. fiscal policy: tax rates were cut sharply, and budgetary deficits ballooned. Large-scale government borrowing added to the demands on financial markets, leading to high interest rates. This encouraged foreign asset holders to buy U.S. bonds. To do so, they bought dollars, creating upward pressure on the exchange value of the dollar. In turn, the high dollar made it difficult for U.S. producers to compete on world markets. U.S. imports rose briskly; exports were relatively sluggish, and the U.S. trade deficit soared.

Because of strong competition from imports, U.S. producers of automobiles, textiles, and a number of other products lobbied for protection. Under the threat of unilateral U.S. actions, the government of Japan was persuaded to impose "voluntary" limits on exports of cars to the United States. There were concerns—both in the United States and in its trading partners—that the United States might adopt a much more protectionist policy because the high exchange value of the dollar was making it so difficult for U.S. producers to compete.

The "Plaza Agreement"

Faced with this unwelcome prospect, senior officials of the "Group of Five" (France, West Germany, Japan, the United Kingdom, and the United States) met at the Plaza Hotel in New York City in 1985. In the "Plaza Agreement," they declared their intention to bring the dollar down to a more competitive level, if necessary by official sales of dollars on exchange markets.

This episode raised fundamental questions about flexible exchange rates, leading some financial experts to suggest an intermediate system between freely flexible exchange rates and the old IMF system of adjustable pegs. With sizable exchange-market interventions by governments and central banks, exchange rates were not freely flexible. They were being managed by the authorities. (Such a managed floating rate is sometimes called a "dirty" float.)

Some experts supported more active exchange-rate management in order to prevent currencies from becoming severely misaligned. Governments were advised to declare "target zones" for exchange rates and to buy or sell currencies whenever needed to keep exchange rates within these zones—moving the target zones as fundamental economic conditions changed. The concept was to avoid large exchange-rate swings.

The European Monetary System. In the early 1970s, when the IMF system of adjustable pegs broke down, the currencies of the western European countries began to float, as did most other currencies.

However, the members of the European Economic Community wanted an exchange-rate agreement to comple-

ment their customs union. An early step was taken in this direction when the nations instituted the so-called "snake in a tunnel." Exchange-rate fluctuations between EEC members were limited, and the currencies moved in a narrow, undulating, snakelike pattern against the U.S. dollar and other outside currencies.

In 1979 most of the members of the EEC (with the important exception of the United Kingdom) entered a more formal agreement, the European Monetary System (EMS), which had some characteristics of the old IMF system.

The European Currency Unit

Exchange rates were to be pegged to a European Currency Unit (ECU), made up of a basket of European currencies. However, there were three important differences from the old IMF system: (1) the flexibility around the official rate was as much as 6 percent, substantially wider than the 1 percent under the IMF system; (2) official rates were to be adjusted more quickly and frequently than the IMF par rates; and (3) the U.S. dollar was not included in the EMS system; thus, the EMS currencies fluctuated as a group against the U.S. dollar.

THE INTERNATIONAL DEBT CRISIS

Developing nations have traditionally borrowed from the developed nations to support their economies. In the 1970s such borrowing became quite heavy among certain developing countries, and their external debt expanded at a very rapid, unsustainable rate. The result was an international financial crisis. Countries such as Mexico and Brazil declared that they could not keep up with the schedule of interest and principal payments, causing severe reactions in the financial world. Cooperating with creditor nations and the IMF, these countries were able to reschedule their debts—that is, delay payments to remove financial pressure. But the underlying problem remained—developing countries were saddled with staggering debts that totaled more than $800,000,000,000 by the mid-1980s. For the less-developed countries as a whole (excluding the major oil exporters), debt service payments were claiming more than 20 percent of their total export earnings.

Rescheduling of debts

The large debts created huge problems for the developing countries and for the banks that faced the risk of substantial losses on their loan portfolios. Such debts increased the difficulty of finding funds to finance development. In addition, the need to acquire foreign currencies to service the debt contributed to a rapid depreciation of the currencies and to rapid inflation in Mexico, Brazil, and a number of other developing nations.

The wide fluctuations in the price of oil were one of the factors contributing to the debt problem. When the price of oil rose rapidly in the 1970s, most countries felt unable to reduce their oil consumption quickly. In order to pay for expensive oil imports, many went deeply into debt. They borrowed to finance current consumption—something that could not go on indefinitely. As a major oil importer, Brazil was one of the nations adversely affected by rising oil prices.

Paradoxically, however, the oil-importing countries were not the only ones to borrow more when the price of oil rose rapidly. Some of the oil exporters—such as Mexico—also contracted large new debts. They thought that the price of oil would move continually upward, at least for the foreseeable future. They therefore felt safe in borrowing large amounts, expecting that rapidly increasing oil revenues would provide the funds to service their debts. The price of oil drifted downward, however, making payments much more difficult.

The debt reschedulings, and the accompanying policies of demand restraint, were built on the premise that a few years of tough adjustment would be sufficient to get out of such crises and to provide the basis for renewed, vigorous growth. To the contrary, however, some authorities believed that huge foreign debts would act as a continuing drag on growth and could have catastrophic results.

BIBLIOGRAPHY

General texts: The classic works in the field of international trade are ADAM SMITH, *An Inquiry into the Nature and Causes of the Wealth of Nations,* 2 vol. (1776), available also in many later editions, both complete and in selections; DAVID RICARDO, *On the Principles of Political Economy and Taxation* (1817), available also in modern editions; JOHN STUART MILL, *Principles of Political Economy,* 2 vol. (1848, reissued in 1 vol., 1909; reprinted 1987); GOTTFRIED VON HABERLER, *The Theory of International Trade with Its Applications to Commercial Policy* (1936, reprinted 1968; originally published in German, 1933); and JACOB VINER, *Studies in the Theory of International Trade* (1937, reprinted 1975). Useful textbooks include RICHARD E. CAVES and RONALD W. JONES, *World Trade and Payments: An Introduction,* 4th ed. (1985); WILFRED J. ETHIER, *Modern International Economics* (1983); PETER B. KENEN, *The International Economy* (1985); and PETER H. LINDERT, *International Economics,* 8th ed. (1986). Surveys of recent international financial developments may be found in the annual compiled by the INTERNATIONAL MONETARY FUND, *World Economic Outlook.* For surveys of recent advanced work in international economics, see RONALD W. JONES and PETER B. KENEN (eds.), *Handbook of International Economics,* vol. 1 (1984); and DAVID GREENAWAY (ed.), *Current Issues in International Trade: Theory and Policy* (1985).

Theories of international trade: The Heckscher-Ohlin theory has been the most scrutinized explanation of trade patterns. The classic work is BERTIL OHLIN, *Interregional and International Trade,* rev. ed. (1967). Other surveys are NICHOLAS OWEN, *Economies of Scale, Competitiveness, and Trade Patterns Within the European Community* (1983); and EDWARD E. LEAMER, *Sources of International Comparative Advantage: Theory and Evidence* (1984).

State interference: Elementary treatments of the theory and practice of tariffs may be found in standard texts, such as those listed above. Institutional and historical studies of tariffs include ASHER ISAACS, *International Trade, Tariff, and Commercial Policies* (1948); and F.W. TAUSSIG, *The Tariff History of the United States,* 8th ed. (1931, reprinted 1967). An analysis of the economic and political issues in trade policy may be found in BELA BALASSA, *Trade Liberalization Among Industrial Countries: Objectives and Alternatives* (1967); ROBERT E. BALDWIN and ANNE O. KRUEGER (eds.), *The Structure and Evolution of Recent U.S. Trade Policy* (1984); JAGDISH N. BHAGWATI (ed.), *Import Competition and Response* (1982); WILLIAM R. CLINE *et al., Trade Negotiations in the Tokyo Round: A Quantitative Assessment* (1978); WILLIAM R. CLINE (ed.), *Trade Policy in the 1980s* (1983); I.M. DESTLER, *American Trade Politics: System Under Stress* (1986); GARY CLYDE HUFBAUER and HOWARD F. ROSEN, *Trade Policy for Troubled Industries* (1986); and ROBERT Z. LAWRENCE and ROBERT E. LITAN, *Saving Free Trade: A Pragmatic Approach* (1986). Accounts of the General Agreement on Tariffs and Trade (GATT) may be found in GERARD CURZON, *Multilateral Commercial Diplomacy* (1965); GARDNER PATTERSON, *Discrimination in International Trade: The Policy Issues, 1945–1965* (1966); and GILBERT R. WINHAM, *International Trade and the Tokyo Round Negotiations* (1986). The effects of trade policies on the developing countries are studied in HARRY G. JOHNSON, *Economic Policies Toward Less Developed Countries* (1967).

International trade arrangements: General analyses include F.W. TAUSSIG, *Free Trade, the Tariff, and Reciprocity* (1920); LEAGUE OF NATIONS, *Commercial Policy in the Interwar Period: International Proposals and National Policies* (1942); JACQUES LACOUR-GAYET, *Histoire du commerce,* 6 vol. (1950–55); and ROBERT SCHNERB, *Libre-échange et protectionnisme,* 4th rev. ed., edited by MADELEINE SCHNERB (1977). Economic integration is treated in MAURICE ALLAIS, *L'Europe unie: route de la prospérité* (1960); BELA BALASSA, *The Theory of Economic Integration* (1961, reprinted 1982); BELA BALASSA *et al.* (eds.), *European Economic Integration* (1975); EUROPEAN FREE TRADE ASSOCIATION, *The Effects of EFTA on the Economies of Member States* (1969); JOSEPH GRUNWALD, MIGUEL S. WIONCZEK, and MARTIN CARNOY, *Latin American Economic Integration and U.S. Policy* (1972); MICHAEL HODGES and WILLIAM WALLACE (eds.), *Economic Divergence in the European Community* (1981); JAMES E. MEADE, *The Theory of Customs Unions* (1955, reprinted 1980); SCOTT R. PEARSON and WILLIAM D. INGRAM, "Economies of Scale, Domestic Divergences, and Potential Gains from Economic Integration in Ghana and the Ivory Coast," *The Journal of Political Economy,* 88:994–1009 (October 1980); DENNIS SWANN, *The Economics of the Common Market,* 5th ed. (1984); JACOB VINER, *The Customs Union Issue* (1950, reprinted 1983); WILLIAM WALLACE (ed.), *Britain in Europe* (1980); PAUL WONNACOTT, *The United States and Canada: The Quest for Free Trade* (1987); and BRUCE PARROTT (ed.), *Trade, Technology, and Soviet-American Relations* (1985).

International commodity markets: BELA BALASSA, *Trade Prospects for Developing Countries* (1964); L. BARANYAI and J.C. MILLS, *International Commodity Agreements* (1963; originally published in Spanish, 1962); C.P. BROWN, *Primary Commodity Control* (1975); JOSEPH D. COPPOCK, *International Economic Instability: The Experience After World War II* (1962), and

International Trade Instability (1977); SIMON KUZNETS, *Modern Economic Growth: Rate, Structure, and Spread* (1966); M.A.G. VAN MEERHAEGHE, *International Economic Institutions,* 4th ed. (1985; originally published in Dutch, 1964); J.W.F. ROWE, *Primary Commodities in International Trade* (1965); and UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT, *Proceedings* (irregular).

International exchange and payments: BENJAMIN J. COHEN, *Organizing the World's Money: The Political Economy of International Monetary Relations* (1977); RICHARD N. COOPER *et al.* (eds.), *The International Monetary System Under Flexible Exchange Rates: Global, Regional, and National* (1982); RICHARD N. COOPER, *Macroeconomic Policy Adjustment in Interdependent Economies* (1969); PAUL EINZIG and BRIAN SCOTT QUINN, *The Euro-Dollar System: Practice and Theory of International Interest Rates,* 6th ed. (1977); J. MARCUS FLEMING, *Guidelines for Balance-of-Payments Adjustment Under the Par-Value System* (1968); ANNE O. KRUEGER, *Exchange-Rate Determination* (1983); GERALD M. MEIER, *Problems of Trade Policy* (1973), *Problems of a World Monetary Order,* 2nd ed. (1982), and *Problems of Cooperation for Development* (1974); ROBERT A. MUNDELL, *International Economics* (1968); ROBERT SOLOMON, *The International Monetary System, 1945–1981,* new ed. (1982); and ROBERT TRIFFIN, *Gold and the Dollar Crisis: The Future of Convertibility* (1960, reprinted 1983), in which the author points out the contradictions in the IMF system. A suggestion for target zones for exchange rates is presented in JOHN WILLIAMSON, *The Exchange Rate System,* 2nd ed. (1985). International debt is specifically discussed in BELA BALASSA *et al., Toward Renewed Economic Growth in Latin America* (1986); WILLIAM R. CLINE, *International Debt: Systemic Risk and Policy Response,* new ed. (1984); THOMAS O. ENDERS and RICHARD P. MATTIONE, *Latin America: The Crisis of Debt and Growth* (1984); and CAROL LANCASTER and JOHN WILLIAMSON (eds.), *African Debt and Financing* (1986).

(P.Wo.)

Iran

The Islāmic republic of Iran (Jomhūrī-ye Eslāmī-ye Īrān) covers 636,300 square miles (1,648,000 square kilometres) in southwestern Asia. It is bounded on the north by the Soviet Union and the Caspian Sea, on the east by Pakistan and Afghanistan, on the south by the Persian Gulf and the Gulf of Oman, and on the west by Turkey and Iraq. Iran also controls about a dozen islands in the Persian Gulf. More than 30 percent of its 4,770-mile (7,680-kilometre) boundary is seacoast. The capital is Tehrān (Teheran).

Among Middle Eastern countries, Iran is unique in many ways. Its official language, Fārsī (Persian), is Indo-European. While Iran adopted a modified version of the Arabic alphabet, it refused to lose its separate identity after the establishment of the Arab Empire in the 7th century. From the beginning of the Islāmic era there was strong support for the house of ʿAlī, Muḥammad's son-in-law. ʿAlī's claim of succession to the Prophet was disputed by the Sunnite majority, but his supporters proclaimed him first *imām* of their party. The party came to be called Shīʿah (Shīʿism), from *shīʿat ʿAlī,* "party of ʿAlī." When conversion to Islām began, Shīʿism was adopted by many, even though it was considered to be a heresy at the time. Despite persecution the Shīʿite branch continued to grow, and in the 16th century it became the official religion of Iran. The concept of divine or sacred kingship, which originated in the Mesopotamian Valley, was an integral part of pre-Islāmic Persian political culture, and the concept persists in the religious beliefs of the Shīʿites, who hold that sovereignty evolves from God and is expressed through his spokesmen on earth.

Iran is the only country in the Middle East that uses the Islāmic solar calendar, which originated before the beginning of the Persian Empire (550 BC). After the conversion of the vast majority of Iranians to Islām, the ancient Iranian calendar was adjusted to begin with the year of the Prophet Muḥammad's immigration to Medina (the *hijrah*), equivalent to AD 622 in the Gregorian calendar. To convert from the Islāmic solar calendar to the Gregorian, 621 or 622 years, depending on the time of year, are added to the Iranian year. (For example, the Iranian year 1361 began on March 21, 1982, and ended on March 20, 1983.) Nō-rūz (New Year's Day) falls on March 21, the vernal equinox. The Arabic lunar calendar is used for religious observances.

Iran has played an important role in the Middle East, as an imperial power and as a factor in rivalries between East and West. Its strategic position and its vast resources, including petroleum and natural gas, make it a nation to be reckoned with in the modern world.

This article is divided into the following sections:

Physical and human geography 855
- The land 855
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 858
- The economy 858
  - Resources
  - Agriculture, forestry, and fishing
  - Industry
  - Finance
  - Trade
  - Transportation
- Government and social conditions 860
  - Government
  - Armed forces
  - Education
  - Health and welfare
- Cultural life 861

History 862
- The Elamites, Medians, and Achaemenids 862
  - The prehistoric period
  - The Elamites
  - The protohistoric period and the kingdom of the Medes
  - The Achaemenid dynasty
  - Achaemenid society and culture
  - The organization and achievement of the empire
- The Hellenistic and Parthian periods 875
  - Alexander and his successors
  - The rise of the Parthians
  - The "phil-Hellenistic" period (*c.* 171 BC–AD 10)
  - The "anti-Hellenistic" period (AD 12–162)
  - The end of the Parthian Empire (AD 162–226)
- The Sāsānian period 882
  - Foundation of the empire
  - Religious developments
  - Art and literature
  - Foreign policy
- Iran from 640 to the present 888
  - The advent of Islām (640–829)
  - The "Iranian intermezzo" (821–1055)
  - The Seljuqs and the Mongols
  - The Timurids and Turkmen
  - The Ṣafavids (1502–1736)
  - Nāder Shāh (1736–47)
  - The Zand dynasty (1750–79)
  - The Qājārs (1779–1925)
  - The Pahlavi dynasty (1925–79)
  - The Islāmic republic

Bibliography 897

Physical and human geography

THE LAND

Relief. A series of massive, heavily eroded mountain ranges surrounds Iran's high interior basin. Most of the country is above 1,500 feet (460 metres), with one-sixth of it over 6,500 feet. In sharp contrast are the coastal regions outside the mountain ring. In the north, the 400-mile strip along the Caspian Sea, never more than 70 miles wide and frequently narrowing to 10, falls sharply from the 10,000-foot summits to the marshy lake's edge, 90 feet below sea level. Along the southern coast, the land drops away from a 2,000-foot plateau, backed by a rugged escarpment three times as high, to meet the Persian Gulf and the Gulf of Oman.

The Zagros Mountains

The Zagros range stretches from the border with the Armenian Soviet Socialist Republic in the northwest to the Persian Gulf, and thence eastward into Baluchistan (Balūchestān). As it moves southward, it broadens into a 125-mile-wide band of parallel ridges lying between the plains of Mesopotamia and the great central plateau of Iran. It is drained on the west by streams that cut deep, narrow gorges and water fertile valleys. The land is extremely rugged and difficult of access and is populated largely by pastoral nomads.

The Elburz Mountains (Reshteh-ye Kūhhā-ye Alborz) run along the south shore of the Caspian Sea to meet the border ranges of Khorāsān (Khurasan) to the east. The highest of the chain's many volcanic peaks, some of which are still active, is snow-clad Qolleh-ye Damāvand (Mt. Demavend); estimates of its height range from about 18,400 feet (5,612 metres) to more than 19,000 feet.

Volcanic activity. Other volcanic peaks include Kūh-e Taftān (Mt. Taftān), a massive cone more than 13,000 feet high in southeastern Iran, which emits gas and mud at sporadic intervals. In the northwest Mt. Sabalān (more than 15,700 feet) and Mt. Sahand (more than 12,000 feet) have been inactive in historical times. Earthquake activity is frequent and violent. The 1,200-mile-long volcanic belt extends from the Soviet border in the northwest to Baluchistan in the southeast. In addition, in the northwestern section of the country lava and ashes cover a 200-mile stretch of land from Jolfā on the Soviet border eastward to the Caspian Sea. A third volcanic region, 250 miles long and 40 miles wide, runs between Lake Urmia (Daryācheh-ye Orūmīyeh) and the city of Qazvīn (Kazvin).

The interior plateau. The arid interior plateau, which extends into Central Asia, is cut by two smaller mountain ranges. In the flatlands lie the plateau's most remarkable features, the Dasht-e Kavīr (Kavīr Desert) and the Dasht-e Lūt (Lūt Desert), also called Kavīr-e Lūt. At the lowest elevations, series of basins in the poorly drained soil remain dry for months at a time; the evaporation of any accumulated water produces the salt wastes known as *kavīrs*. As elevation rises, surfaces of sand and gravelly soil gradually merge into fertile soil on the hillsides and mountain slopes.

Dasht-e Kavīr and Dasht-e Lūt

Drainage. The few streams emptying into the desiccated central plateau dissipate in saline marshes. The general drainage pattern is down the outward slopes of the mountains, terminating in the sea. There are three large rivers, but only one—the Rūd-e Kārūn (Kārūn River)—is navigable. It originates in the Zagros Mountains and flows south to the Shatt al-Arab (Arvand Rūd), which empties into the Persian Gulf. The Sefīd Rūd (Safid River) originates in the Elburz Mountains in the north. It runs as a mountain stream for most of its length but flows rapidly into the Gīlān plain and then to the Caspian Sea. The Safid River Dam, completed in the early 1960s at Manjīl, generates hydroelectric power and provides water for irrigation.

The Zāyandeh Rūd, the lifeline of Eṣfahān (Isfahan), also originates in the Zagros Mountains, flowing southeastward to Bāṭlāq-e Gāv Khūnī (Gāvkhāneh Marsh), a swamp northwest of the city of Yazd. The completion of the Kūhrang Dam in 1971 diverted water from the upper Kārūn through a two-mile-long tunnel into the Zāyandeh Rūd for irrigation purposes.

Other streams are seasonal and variable: spring floods do enormous damage, while in summer many streams disappear. Water is, however, stored naturally underground, finding its outlet in springs and in tap wells.

The largest inland body of water, Lake Urmia, in northwestern Iran, covers an area that varies from 2,000 to 2,300 square miles. Other lakes are principally seasonal, and all have a high salt content.

Soils. Soil patterns vary widely. The abundant subtropical vegetation of the Caspian coastal region is supported by rich brown forest soils. Mountain soils are shallow layers over bedrock, with a high proportion of unweathered fragments. Natural erosion moves the finer textured soils into the valleys. The alluvial deposits are mostly chalky, and many are used for pottery.

The semiarid plateaus lying above 3,000 feet are covered by brown or chestnut-coloured soil that supports grassy vegetation. The soil is slightly alkaline and contains 3 to 4 percent of organic material. The saline and alkaline soils in the arid regions are light in colour and infertile. The sand dunes are composed of loose quartz and fragments of other minerals. Except where anchored by vegetation, they are in almost constant motion, driven by high winds.

Climate. Iran's complex climate ranges from subtropical to subpolar. In winter a high-pressure belt, centred in Siberia, slashes west and south to the interior of the Iranian Plateau, while low pressures develop over the warm waters of the Caspian Sea, the Persian Gulf, and the Mediterranean Sea. In summer one of the lowest pressure centres in the world prevails in the south. Low pressures in Pakistan generate two regular wind patterns: the *shamāl,* which blows from February to October northwesterly through the Tigris–Euphrates Valley, and the "120-day" summer wind, which sometimes reaches velocities of 70 miles per hour in the Sīstān (Seistan) region near the Pakistan frontier. Warm Arabian winds bring heavy moisture from the Persian Gulf.

Altitude, latitude, maritime influences, seasonal winds, and proximity to mountain ranges or deserts play a significant role in the fluctuation of temperature, which varies from a high of 131° F (55° C) in Khūzestān (Khuzistan) at the head of the Persian Gulf to a low of −35° F (−37° C) in Azerbaijan (Āẕarbāyjān) in the northwest. Precipitation also varies widely, from less than two inches (50 millimetres) in the southeast to about 78 inches in the Caspian region. The annual average is about 16 inches. Winter is normally the rainy season for the country as a whole: more than 50 percent of the annual precipitation occurs in the three-month period. The northern coastal region presents a sharp contrast. The high Elburz Mountains, which seal off the narrow Caspian plain, wring moisture from the clouds, trap humidity from the air, and create a fertile semitropical region with luxuriant forests, swamps, and rice paddies. Temperatures may soar to 100° F (38° C) and the humidity to 98 percent. Frost is extremely rare. Except for this region, summer is a dry season. The northern and western parts of Iran have four distinct seasons. Moving south and east, spring and autumn become shorter and ultimately merge in an area of mild winters and hot summers.

Temperature and rainfall

Plant and animal life. Topography, altitude, water supply, and soil determine the character of the vegetation. Approximately one-tenth of Iran is forested, most extensively in the Caspian region. In the area are found broad-leafed, deciduous trees—oak, beech, linden, elm, walnut, ash, and hornbeam—and a few broad-leafed evergreens. Thorny shrubs and ferns also abound.

The Zagros Mountains are covered by semihumid oak forests, together with elm, maple, hackberry, walnut, pear, and pistachio trees. Willow, poplar, and plane trees grow in the ravines, as do many species of creepers. Thin stands of juniper, almond, berberis, cotoneaster, and wild fruit trees grow on the intermediate dry plateau. Thorny shrubs form the ground cover of the steppes, while artemisia (a kind of wormwood) grows at medium elevations of the desert plains and the rolling country. There are acacia, dwarf palm, kunar trees (Jerusalem thorn), and scattered shrubs below 3,000 feet. Desert sand dunes, which hold water, support thickets of brush. Forests follow the courses of surface or subterranean waters. Oases support vines



IRAN

MAP INDEX

Political subdivisions

Āẕarbāyjān-e Bākhtarī . . . . . 37 40 N 45 00 E
Āẕarbāyjān-e Khāvarī . . . . . 38 00 N 47 00 E
Balūchestān va Sīstān . . . . . 28 30 N 60 30 E
Bovīr Aḥmadī va Kohkīlūyeh . . . . . 30 50 N 50 40 E
Chahār Maḥāll va Bakhtīārī . . . . . 32 10 N 50 45 E
Eṣfahān . . . . . 33 00 N 52 10 E
Fārs . . . . . 29 00 N 53 00 E
Gīlān . . . . . 37 00 N 49 00 E
Hamadān . . . . . 34 50 N 48 35 E
Hormozgān . . . . . 27 45 N 56 00 E
Īlām . . . . . 33 10 N 47 00 E
Kermān . . . . . 29 00 N 57 00 E
Kermānshāhān . . . 34 30 N 47 00 E
Khalij-e Fārs . . . . . 29 00 N 51 35 E
Khorāsān . . . . . 35 00 N 59 00 E
Khūzestān . . . . . 31 00 N 49 00 E
Kordestān . . . . . 35 40 N 47 00 E
Lorestān . . . . . 33 30 N 48 30 E
Markazī . . . . . 34 30 N 50 10 E
Māzandarān . . . . . 36 30 N 53 30 E
Semnān . . . . . 35 30 N 55 00 E
Tehrān . . . . . 35 30 N 51 40 E
Yazd . . . . . 32 00 N 54 40 E
Zanjān . . . . . 36 40 N 48 29 E

Cities and towns

Ābādān . . . . . 30 20 N 48 16 E
Ahar . . . . . 38 28 N 47 04 E
Ahvāz . . . . . 31 19 N 48 42 E
Āmol . . . . . 36 28 N 52 21 E
Arāk . . . . . 34 05 N 49 41 E
Ardabīl . . . . . 38 15 N 48 18 E
Bābol . . . . . 36 33 N 52 42 E
Bāfq . . . . . 31 36 N 55 24 E
Bajestān . . . . . 34 31 N 58 10 E
Bam . . . . . 29 06 N 58 21 E
Bampūr . . . . . 27 12 N 60 27 E
Bandar Beheshtī . . . . . 25 18 N 60 37 E
Bandar-e 'Abbās . . 27 11 N 56 17 E
Bandar-e Anzalī . . . 37 28 N 49 27 E
Bandar-e Khomeynī . . . . . 30 26 N 49 04 E
Bandar-e Lengeh . . 26 33 N 54 53 E
Bandar-e Torkeman . . . . . 36 54 N 54 04 E
Basṭām . . . . . 36 29 N 55 00 E
Behbahān . . . . . 30 35 N 50 14 E
Behshahr . . . . . 36 43 N 53 34 E
Bīrjand . . . . . 32 53 N 59 13 E
Bojnūrd . . . . . 37 28 N 57 19 E
Borūjerd . . . . . 33 54 N 48 46 E
Būshehr (Bushire) . . . . . 28 59 N 50 50 E
Dāmghān . . . . . 36 09 N 54 22 E
Dārāb . . . . . 28 45 N 54 34 E
Delījān . . . . . 33 59 N 50 40 E

Dezfūl ..... 32 23 N 48 24 E
Emāmshahr ..... 36 25 N 54 58 E
Eṣfahān ..... 32 40 N 51 38 E
Ferdows ..... 34 00 N 58 09 E
Firūzābād ..... 28 50 N 52 36 E
Gachsārān ..... 30 12 N 50 47 E
Gonābād ..... 34 20 N 58 42 E
Gonbad-e Qābūs ..37 15 N 55 09 E
Gorgān ..... 36 50 N 54 29 E
Haft Gel ..... 31 27 N 49 32 E
Hamadān ..... 34 48 N 48 30 E
Īlām ..... 33 38 N 46 26 E
Jahrom ..... 28 31 N 53 33 E
Jāsk ..... 25 38 N 57 46 E
Jolfā ..... 38 57 N 45 38 E
Karaj ..... 35 48 N 50 59 E
Kāshān ..... 33 59 N 51 29 E
Kāzerūn ..... 29 37 N 51 38 E
Kermān ..... 30 17 N 57 05 E
Kermānshāh (Bakhtaran) ..... 34 19 N 47 04 E
Khomeynishahr ...32 41 N 51 31 E
Khorramābād .... 33 30 N 48 20 E
Khorramshahr .... 30 25 N 48 11 E
Khvoy ..... 38 33 N 44 58 E
Kūhrang ..... 32 26 N 50 06 E
Lādīz ..... 28 56 N 61 19 E
Lāhījān ..... 37 12 N 50 01 E
Lār ..... 27 41 N 54 17 E
Mahābād ..... 36 45 N 45 43 E
Malāyer ..... 34 18 N 48 50 E
Manjil ..... 36 44 N 49 24 E
Marāgheh ..... 37 23 N 46 14 E
Marv Dasht ..... 29 52 N 52 48 E
Mashhad ..... 36 18 N 59 36 E
Masjed-e Soleymān ..... 31 58 N 49 18 E
Miāneh ..... 37 26 N 47 42 E
Nā'īn ..... 32 52 N 53 05 E
Najafābād ..... 32 37 N 51 21 E
Nehbandān ..... 31 32 N 60 02 E
Neyshābūr ..... 36 12 N 58 50 E
Orūmīyeh ..... 37 33 N 45 04 E
Qā'emshahr ..... 36 28 N 52 53 E
Qā'en ..... 33 44 N 59 11 E
Qazvīn ..... 36 16 N 50 00 E
Qom ..... 34 39 N 50 54 E
Qūchān ..... 37 06 N 58 30 E
Rafsanjān ..... 30 24 N 56 00 E
Rasht ..... 37 16 N 49 36 E
Sabzevār ..... 36 13 N 57 42 E
Sakht Sar ..... 36 53 N 50 41 E
Sanandaj ..... 35 19 N 47 00 E
Saqqez ..... 36 14 N 46 16 E
Sarāb ..... 37 56 N 47 32 E
Sārī ..... 36 34 N 53 04 E
Semnān ..... 35 33 N 53 24 E
Shahr-e Kord ..... 32 19 N 50 50 E
Shahr-e Rey ..... 35 35 N 51 25 E
Shīrāz ..... 29 36 N 52 32 E
Shūshtar ..... 32 03 N 48 51 E
Sūrmaq ..... 31 03 N 52 48 E
Tabrīz ..... 38 05 N 46 18 E
Tehrān ..... 35 40 N 51 26 E
Torbat-e Heydarīyeh ..... 35 16 N 59 13 E
Torbat-e Jām ..... 35 14 N 60 36 E
Yāsūj ..... 30 40 N 51 36 E
Yazd ..... 31 53 N 54 22 E
Zābol ..... 31 02 N 61 30 E
Zāhedān ..... 29 30 N 60 52 E
Zanjān ..... 36 40 N 48 29 E

Physical features and points of interest

Abū Mūsā, Jazīreh-ye, *island* ..... 25 52 N 55 03 E
Alborz, Reshteh-ye Kūhhā-ye, see Elburz Mountains
Aras (Aras Nehri), *river* ..... 39 56 N 48 20 E
Atrak, *river* ..... 37 28 N 54 03 E
Bakhtegān Wildlife Refuge ..29 28 N 53 43 E
Bakhtegān, Daryācheh-ye, *lake* ..... 29 20 N 54 05 E
Baluchistan, *geographical region* ..... 28 00 N 63 00 E
Caspian Sea ..... 39 00 N 51 20 E
Central Elburz National Park .... 36 15 N 51 24 E
Choghā Zanbīl ...32 01 N 48 31 E
Damāvand, Qolleh-ye, *mountain* ..... 35 56 N 52 08 E
Elburz Mountains (Alborz, Reshteh-ye Kūhhā-ye) ..... 36 00 N 53 00 E
Gāv Khūnī, Bāṭlāq-e, *marsh* .32 10 N 52 50 E
Gavāter, Khalīj-e, *bay* ..... 25 04 N 61 36 E
Golestān National Park .... 37 19 N 55 19 E
Harīrūd, *river* ..... 37 24 N 60 38 E
Hormuz, Strait of ..26 34 N 56 15 E
Jaz Mūrīān, Hāmūn-e, *lake* ..... 27 20 N 58 55 E
Kabīr Kūh, *mountains* ..... 33 25 N 46 45 E
Kārūn, *river* ..... 30 26 N 48 10 E
Kavīr, Dasht-e, *desert* ..... 34 40 N 54 30 E
Kavīr National Park ..... 34 47 N 51 24 E
Khārk, Jazīreh-ye (Kharg Island) ...29 15 N 50 20 E
Khersān, *river* ....31 33 N 50 22 E
Khosh Yeylagh Wildlife Refuge ..... 36 55 N 55 13 E
Kish, Jazīreh-ye, *island* ..... 26 32 N 53 58 E
Koppeh Dāgh, *mountains* ..... 37 50 N 58 00 E
Lake Urmia National Park .... 37 40 N 45 20 E
Lāvān, Jazīreh-ye, *island* ..... 26 48 N 53 15 E
Lūt, Dasht-e, *desert* ..... 33 00 N 58 00 E
Mand, *river* ..... 28 11 N 51 17 E
Meydān-e Shāh ...32 40 N 51 40 E
Moteh Wildlife Refuge ..... 33 30 N 50 47 E
Namak, Daryācheh-ye, *salt lake* ..... 34 30 N 51 50 E
Oman, Gulf of .... 25 00 N 58 00 E
Orūmīyeh, Daryācheh-ye, see Urmia, Lake
Persepolis (Takht-e Jamshīd) ..... 29 57 N 52 52 E
Persian Gulf ..... 27 00 N 51 00 E
Qeshm, Jazīreh-ye, *island* ..... 26 45 N 55 45 E
Qezel Owzan, *river* ..... 36 47 N 49 07 E
Sabalān, Reshteh-ye Jebāl-ye, *mountain* ..... 38 15 N 47 49 E
Ṣāberī, Hāmūn-e, *lake* ..... 31 30 N 61 20 E
Sahand, Kūh-e, *mountain* ..... 37 44 N 46 27 E
Sefīd, *river* ..... 37 26 N 49 55 E
Shahdād, Namakzār-e, *salt waste* ..... 30 30 N 58 30 E
Sirrī, Jazīreh-ye, *island* ..... 25 55 N 54 32 E
Sīstān, *geographical region* ..... 30 30 N 62 00 E
Sīstān, Daryācheh-ye, *lake* ..... 31 00 N 61 15 E
Taftān, Kūh-e, *mountain* ..... 28 36 N 61 08 E
Tūrān Protected Area ..... 35 10 N 55 44 E
Urmia, Lake (Orūmīyeh, Daryācheh-ye) ...37 40 N 45 30 E
Zagros Mountains (Zāgros, Kūhhā-ye) ..... 33 40 N 47 00 E
Zard Kūh, *mountain* ..... 32 22 N 50 04 E

and tamarisk, poplar, date palm, myrtle, oleander, acacia, willow, elm, plum, and mulberry trees. In swamp areas reeds and grass provide good pasture.

Wildlife includes leopards, bears, hyenas, wild boars, ibex, gazelles, and mouflon, which live in the wooded mountains. Jackals and rabbits roam the interior. Wild asses live in the *kavīr*. Cheetahs and pheasants are found in the Caspian region, and partridges live in most parts of the country. Other birds, such as seagulls, ducks, and geese, live on the shores of the Caspian Sea and the Persian Gulf, while buzzards nest in the desert. Deer, hedgehogs, foxes, and 22 species of rodents live in semidesert, high-altitude regions. Palm squirrels, Asiatic black bears, and tigers live in Baluchistan. Tigers also inhabit the forests of the Caspian region.

Studies made in Khūzestān, Baluchistan, and along the slopes of the Elburz and Zagros mountains have revealed the presence of a remarkably wide variety of amphibia and reptiles. Examples are toads, frogs, tortoises, lizards, salamanders, boas, racers, rat snakes, cat snakes, and vipers.

Some 200 varieties of fish live in the Persian Gulf, as do shrimps, lobsters, and turtles. Sturgeon, the most important commercial fish, is one of 30 species found in the Caspian Sea. It constitutes a major source of export income for the government, in the production of caviar. Mountain trout abound in small streams at high elevations and in rivers that are not seasonal.

Wildlife sanctuaries have been established in Iran, and the hunting of swans, pheasants, deer, tigers, and a number of other animals and birds is prohibited. Domestic animals include horses, donkeys, cattle, water buffalo, sheep, goats, both species of camels, dogs, and cats.

Regional characteristics

Settlement patterns. Traditionally, the country has been sharply divided along cultural, ethnic, and linguistic lines. In the popular mind, the people of each region, frequently associated with a particular city, have specific attributes. The Isfahani, for example, is known as a shrewd businessman, witty and industrious; the Yazdi as a hard-working agriculturalist. Azerbaijan produces soldiers and merchants. A Kurd is religiously independent, a fanatical fighter, while the Shīrāzī is a poet by nature and temperament.

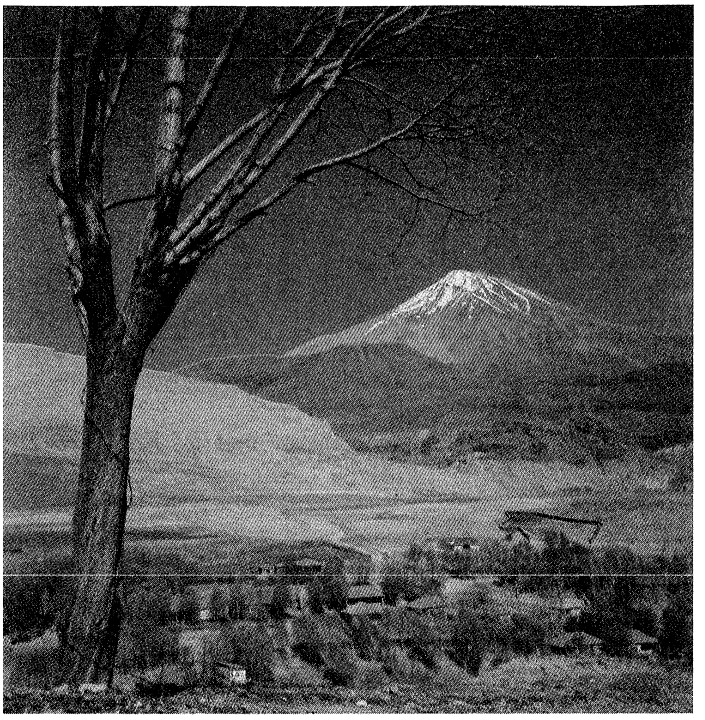
The topography and the water supply determine the regions fit for human habitation, the character of the people, and their types of dwellings. The deep gorges and defiles, the unnavigable rivers, the empty deserts, and the impenetrable *kavīr*s all result in insularity and tribalism and in a concentration of population around the periphery and in the oases. The felt yurts (circular dwellings) of the Turkmen, the black tents of the Bakhtyārī, and the osier (willow) huts of the Baluchi are typical, as the tribesmen roam from summer to winter pasture. In the vast central and southern plains there are a number of oasis settlements with scattered rudimentary hemispherical or conical huts. Since the mid-20th century the migrations have shortened and the nomads have settled in more permanent villages.

The villages on the plains follow an ancient rectangular pattern. High mud walls with corner towers form the outer face of the houses, which have flat roofs of mud and straw supported by wooden rafters. In the open centre of the village is a mosque serving also as a school.

Mountain villages are situated on the rocky slopes above the valley floor, surrounded by terraced fields, usually irrigated, of grain and lucerne (alfalfa). The houses are square, mud-brick, windowless buildings with flat or domed roofs; a roof hole provides ventilation and light. The stable is usually under the house.

Caspian villages are completely different. The scattered hamlets have two-storied wooden houses. Separate outbuildings (barns, henhouses, silkworm houses) surround an open courtyard.

Tehrān, the capital and largest city, is separated from



Mt. Demavend, the highest volcanic peak in the Elburz Mountains.
The J. Allan Cash Photolibrary

the Caspian by the Elburz Mountains. Eṣfahān, about 250 miles south of Tehrān, is the second most important city and is famed for its architecture. There are few cities in central and eastern Iran, where water is scarce, although lines of oases penetrate the desert. Most towns are supplied with water by *qanāt*, an irrigation system by which an underground mountain water source is tapped and the water channelled down through a series of tunnels, sometimes 50 miles in length, to the town level. Towns are, therefore, often located a short distance from the foot of a mountain. The essential feature of a traditional Iranian street is a small canal.

Islāmic influence has determined city layout. There is still strict division into separate quarters—the governmental, the residential, and the business sector, or bazaar. The latter, fronting on a central square, is a maze of narrow arcades lined with small individual shops, grouped according to the type of product sold. Modern business centres, however, have grown up outside the bazaars. Dwellings in the traditional style are built around closed courtyards, with a garden and a pool. Public baths are found in all sections of the cities.

Construction of broad avenues and ring roads to accommodate modern traffic has changed the appearance of the large cities. Their basic plan, however, is still that of a labyrinth of narrow, crooked streets and culs-de-sac.

THE PEOPLE

Iran is a multilingual and diverse cultural society, and the majority of the population is extremely young. Nearly one-half of the people speak Fārsī, and another one-fourth speak some other Indo-European language or dialect. These are descendants of the Aryan tribes, whose origins are lost in antiquity. The Kurds, whose language has seen scant modification over the centuries, are a fierce nomadic people dwelling in the western mountains of Iran and in Iraq and Turkey. They constitute a small percentage of Iran's population. They have resisted the Iranian government's efforts, both before and after the revolution of 1979, to assimilate them into the mainstream of national life. Also inhabiting the western mountains are seminomadic Lurs, thought to be the descendants of the aboriginal inhabitants of the country. Closely related are the Bakhtyārī tribes, who live in the Zagros Mountains west of Eṣfahān. Both speak Lurī, a language distinct from, but related to, Fārsī. The Baluchi are a smaller minority who inhabit Iranian Baluchistan, which borders on Pakistan.

The Kurds

About one-fifth of Iranians speak a variety of Turkic languages. The largest Turkic-speaking group is the Azerbaijani, a farming and herding people who inhabit two northwestern provinces bordering on Turkey and the Soviet Union. Two other Turkic ethnic groups are the Qashqā'īs in the Shīrāz area to the north of the Persian Gulf and the Turkmen of Khorāsān in the northeast.

The Armenians, with a different ethnic heritage, have maintained their Indo-European linguistic identity. They are concentrated in Tehrān, Eṣfahān, and Azerbaijan and are engaged primarily in commercial pursuits. A few isolated remnants of Dravidians are found in the Sīstān region to the southeast.

Semites—Jews, Assyrians, and Arabs—constitute only a small percentage of the population. The Jews, like the Armenians, have retained their ethnic, linguistic, and religious identity and traditionally have clustered in the largest cities. The Assyrians are concentrated in the northwest, and the Arabs live primarily in the Persian Gulf islands and in Khūzestān. The Arabs' demand for autonomy was among the factors that led to the Iran–Iraq war of the 1980s.

Wars, intermarriage, and commerce have resulted in the intermingling of both blood and vocabulary. Before the revolution of 1979, English and French, and to a lesser degree German and Russian, were widely used by the educated class. Since 1979 Arabic has received priority because of its religious significance (it is the language of the Qur'ān).

The vast majority of Iranians are Muslims, mostly of the Shī'ite branch, which is the official state religion. The Kurds and Turkmen are Sunnite Muslims, and among Iran's Arabs the Sunnites and other sects are represented. (For a discussion of Islāmic sects, see ISLĀM, MUḤAMMAD AND THE RELIGION OF: *Theology and sectarianism.*)

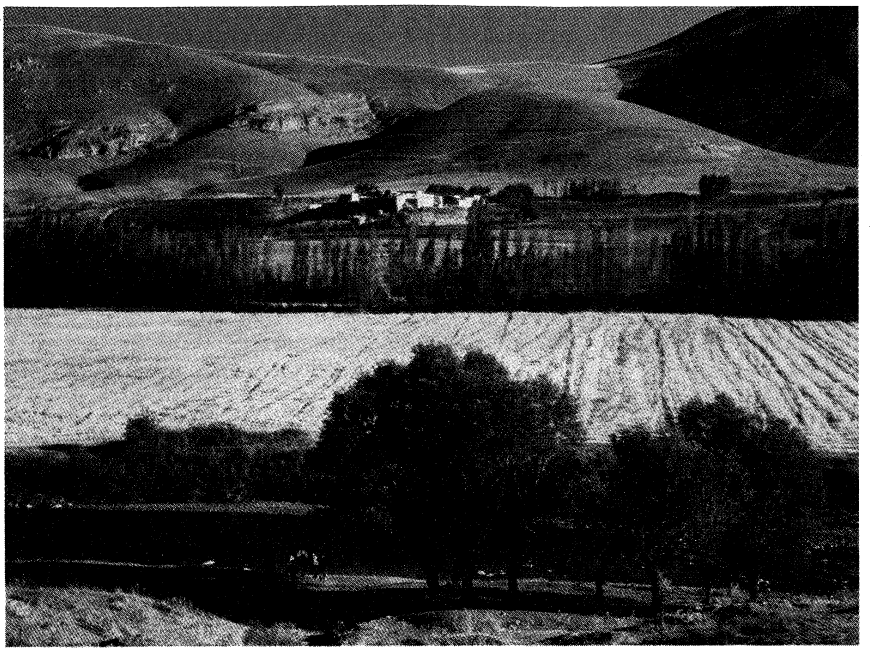
Major religious minorities are Christians, Jews, and Zoroastrians. Christians comprise the largest minority group, of whom Orthodox Armenians constitute the bulk. The Assyrians are Nestorians, Protestants, and Roman Catholics, as are a few converts from other ethnic groups. The Zoroastrians are largely concentrated in Yazd in central Iran, Kermān to the southeast, and Tehrān.

Religious toleration, one of the characteristics of Iran as a monarchy, came to an end with the Islāmic revolution in 1979. While Christians, Jews, and Zoroastrians are recognized in the constitution of 1979 as official minorities, the revolutionary atmosphere in Iran is not conducive to equal treatment of non-Muslims. Among these, Iran's Bahā'īs were the greatest victims of persecution. Emigration greatly reduced the Jewish population, which had been a significant minority before 1979.

THE ECONOMY

Planned economic development, initiated in 1949, was terminated in 1978 and revived in 1982, but Iran's involvement in war with Iraq impeded progress. The management of the economy under the republic is subject to Islāmic criteria as determined by the Council of Guardians and approved by the legislature. The constitution of 1979 establishes specific guidelines for the administration of the nation's economic and financial affairs. The ultimate objectives are economic independence, full employment, and a comfortable standard of living. The economy is divided into three sectors: public, which includes major industries, banks, insurance companies, utilities, communications, foreign trade, and mass transportation; cooperative, which includes production and distribution of goods and services; and private, which consists of all activities that supplement the first two sectors.

In general terms, all former laws or any sections of them that violated the principles of Islāmic law were declared null and void. The nationalization of private banks and insurance companies and the increased state control of foreign trade have given the government a monopoly of most income-producing activities. As to the rest of the economy, most of the items listed for nationalization were already under state control at the time that the monarchy



The rich agricultural region of northwestern Iran.
© Abbas—Magnum

was overthrown. The various departments of the central government in Tehrān, as well as their provincial and local offices, either are headed by mullahs (religious teachers who are conversant with Islāmic law) or are advised by local religious leaders in the communities concerned.

In addition to the government's mandatory taxes, Islāmic taxes are collected on a voluntary basis. These include an income tax (*khums:* "one-fifth"); a tax for charitable causes (*zakāt*), which has a variable rate and is frequently in kind; and a land tax (*kharāj*), the rate of which has also been based on the principle of *ūshr,* 10 percent of the value of the crops, unless the land is tax-exempt.

Resources. Iran's most important natural resources are oil and natural gas. Production and exploration are concentrated in the southwest, but oil has also been discovered in Qom and the Kavīr-e Lūt, as well as under the offshore waters of the Persian Gulf.

Oil is produced for export and for domestic consumption. It is moved by pipelines to the terminal at Jazīreh-ye Khārk (Khārg Island) in the Persian Gulf and from there is shipped by tankers to western Europe, Asia, Japan, and other countries. Oil was refined primarily in Ābādān until the city was destroyed in the war with Iraq. Smaller refineries in Eṣfahān, Shīrāz, Lāvān, Tehrān, and Tabrīz produce a variety of refined products.

Natural gas, found in the south as well as in the Elburz Mountains and in Khorāsān, is also a valuable asset. Gathering and distributing spur lines run to Tehrān, Kāshān, Eṣfahān, Shīrāz, Mashhad, Ahvāz, and the industrial city of Alborz, near Qazvīn. This state-owned system, known as IGAT, is one of the largest gas lines in the Middle East.

The petrochemical industry

The petrochemical industry, concentrated in the south, expanded rapidly before the revolution. The Rāzī (formerly Shāhpūr) Chemical Company at Bandar-e Khomeynī (formerly Bandar-e Shāhpūr) produces ammonia, phosphates, sulfur, and other products. The plants of Khārg Island turned out sulfur, liquid gas, and light oil. The war with Iraq led to massive destruction of the petroleum and petrochemical industries.

Iran's other mineral resources are largely undeveloped. Major coal mines are in Khorāsān, Kermān, Semnān, Māzandarān, and Gilān. Smaller mines operate north of Tehrān, in Azerbaijan, and in Eṣfahān. Lead, in combination with zinc and other minerals, is widely scattered. Copper is found in a belt extending from the Pakistani border to the Soviet border. Kermān is the centre of copper production. Cement production is centred in the provinces of Tehrān, Fārs, and Khūzestān and in the region of Qazvīn, northwest of Tehrān. Fireclay, chalk, lime, gypsum, ochre, and kaolin (china clay) are also produced in commercial quantities. Other mineral resources include magnesite, uranium, chromium, antimony, manganese, tin, mica, alum, marble, turquoise, and emeralds.

Agriculture, forestry, and fishing. Because of poor soil and lack of adequate water distribution in many areas, less than 10 percent of the agricultural land in rural areas is under cultivation. Less than one-third of the cultivated area is irrigated; the rest is devoted to dry farming. The western and northwestern portions of the country are the most fertile.

Although the economy is primarily agrarian, agricultural activities provide a relatively small part of the gross domestic product. The farms are small; most are less than 25 acres. Soil is poor, seed is of low quality, water is scarce, and farming techniques are archaic. All of these factors contribute to low crop yield and to poverty in the rural areas.

The wide range of temperature fluctuation in different parts of the country makes it possible to cultivate diverse crops. These include cereals (wheat, barley, rice, corn [maize]), fruits (dates, figs, pomegranates, melons, grapes) and vegetables, cotton, sugar beets and sugarcane, nuts, olives, spices, tea, tobacco, and medicinal herbs.

Forests cover approximately the same amount of land as agricultural crops. The largest and most valuable forested areas are in the Caspian region, where many of the forests are commercially exploitable and include both hard and soft woods. Forest products include plywood, fibreboard, and lumber for the construction and furniture industries.

Forest crops and products

Fishing is important, and fish are caught both for domestic consumption and for export. They are marketed fresh, salted, smoked, or canned. Sturgeon (yielding its roe for caviar), bream, whitefish, salmon, mullet, carp, catfish, perch, and roach are caught in the Caspian Sea. More than 200 species of fish are found in the Persian Gulf, 150 of them edible. The catch includes shrimps and prawns.

Of the country's livestock, sheep are by far the most numerous, followed by goats, cattle, asses, horses, water buffalo, and mules. Camels and poultry are also found in large numbers.

Industry. Tehrān is the largest market for domestic products. Agricultural products are shipped to the nearest local town and thence transported to Tehrān or to provincial capitals by air, truck, rail, camels, mules, and donkeys. Since crafts are localized, each city has created a market for its products in the capital and other major

cities. Major industries are scattered, and their products are distributed nationwide.

Since 1954 industrial development has transformed parts of the country. Iran produces a wide range of products, such as automobiles, electrical appliances, telecommunications equipment, industrial machinery, paper, rubber products, steel, food products, wood and leather products, textiles, and pharmaceuticals.

Textile mills are centred in Eṣfahān and the Caspian region. Carpet weaving, a traditional Iranian industry, remains important; it contributes substantially to incomes in rural areas and is one of Iran's most important export industries.

Mining industry. Mining was primarily by hand until the early 1960s, and the ore was moved to refining centres by truck, rail, donkey, or camel. Mechanized methods were introduced as new mines and quarries were opened up. The mineral industries encompass both refining and manufacturing.

Construction. Until the early 1950s the construction industry was limited for the most part to small indigenous companies. When large government contracts were involved, however, foreign bidders were attracted. Increased income from oil and gas and the availability of easy credit triggered a building boom that took major international construction firms to Iran. This continued until 1975, when, because of a sharp rise in inflation, credit was tightened and the boom collapsed.

Energy. Until the 20th century Iran's sources of energy were almost entirely limited to wood and charcoal. Oil, natural gas, and coal are now used to supply heat and energy. A system of dams supplies hydroelectric power (and also irrigates cropland).

The establishment of the Atomic Energy Organization (AEO) of Iran in 1973 was intended to implement the government's plan to construct more than 20 nuclear plants. By 1978 two 1,200-megawatt reactors near Būshehr on the Persian Gulf were near completion, and they were scheduled to begin operation early in 1980. The revolutionary government cancelled the program in 1979. The AEO has been revived, but its activities are limited to the search for uranium inside Iran.

Tourism. Until the early 1960s little attention was paid to tourism. Lack of facilities made travel in Iran a rugged experience. Provisions for paved highways and hotel construction were then included in each successive development plan. The number of tourists increased steadily in the years 1964–78, but the political turmoil of 1978, which led to the overthrow of the monarchy, practically destroyed the tourist industry as well. The Islāmic regime discourages tourism from non-Muslim countries in an effort to exclude Western influences. Service industries have been decimated by the sharp decline in the tourist industry.

Finance. Government loans and credits have been made available to industrial and agricultural projects, primarily through banks. All private banks and insurance companies were nationalized in 1979. The Islāmic Bank of Iran (later reorganized as the Islāmic Economy Organization) was established in Tehrān, with branches throughout the country. Banks are divided into categories—such as commercial, industrial, and agricultural—but all are subject to the same regulations. In lieu of interest, considered to be usury and forbidden by Islām, banks impose a service charge, a commission, or both on loans.

In 1984 the value of trade in Tehrān's stock exchange was less than $1,000,000. Foreign investment came to a halt in 1979.

Trade. Despite the government's attempts to achieve self-sufficiency, the value of Iran's imports continues to be high, with machinery and transport equipment accounting for a considerable proportion, followed by basic manufactures and foodstuffs. The huge income derived from the export of petroleum products created a favourable balance of trade in the past, but this has been reversed. Other exports include caviar, carpets, cotton, fruits, textiles, minerals, chemicals, nuts, and motor vehicles.

Transportation. Iran's large centres of population are widely scattered, and transportation is made difficult by mountainous and desert terrain. Low funding and poor maintenance have reduced the efficiency of the all-weather roads, of which less than a third are paved. Nevertheless, motor vehicles, and buses and trucks in particular, are the most important means of transportation for both passengers and goods.

The principal line of the state-owned railway system spans the distance between the Caspian Sea and the Persian Gulf, with spur lines to some of the provincial capitals. In 1971 the railway was linked through Turkey with the European system; the link stimulated trade and tourism appreciably, undercutting air fares and materially reducing sea-transportation time. The Iranian portion of a line eastward to Singapore was completed as far as Mashhad by 1971. There is also a connection with the Soviet rail system via Jolfā in the northwest.

The Kārūn is the only navigable river and is used to transport passengers and cargo. On Lake Urmia there is regular passenger and cargo ferry service between the port of Sharafkhāneh in the northeast and Golmānkhāneh in the southwest. Before the outbreak of war with Iraq, Iran was served by five major ports on the Persian Gulf. Oil terminals at Ābādān and Khārg Island were destroyed or damaged in the fighting, as were port facilities at Khorramshahr and Bandar-e Khomeynī. Although facilities at the port of Būshehr have been expanded, neither it nor Bandar-e ʿAbbās, the largest port on the Gulf, can accommodate foreign trade. Consequently, a new port was readied at Bandar Beheshtī (formerly Chāh Bahār), on the Gulf of Oman.

Persian Gulf ports

Caspian seaports are primarily used for trade with the Soviet Union. Notable among them are Bandar-e Anzalī (formerly Bandar-e Pahlavī) and Bandar-e Torkeman (formerly Bandar-e Shāh).

Major cities and provincial capitals are served by the state-owned Iran Air, and all have airports capable of handling jets. Some major European, Asian, and African airlines serve Iran; Tehrān, Eṣfahān, Shīrāz, and Bandar-e ʿAbbās have international airports.

GOVERNMENT AND SOCIAL CONDITIONS

Government. As provided in the constitution of 1979, Iran is an Islāmic republic, with a president (officially the head of state), a Cabinet, a unicameral legislature (Majles), and a separate judicial branch. The *valiy-e faqih,* the supreme spiritual leader, chosen for his knowledge of Islāmic theology, is considered to be the representative of the Twelfth Imām and as such has final authority in all executive, legislative, and judicial matters. After the *valiy-e faqih*'s death, experts (jurists) elected by the people consult and select his successor, to be designated as the new "leader." If no single individual is available to fill the position, a Council of Leadership of three to five theologians is established, subject to popular approval.

The *valiy-e faqih*

The president, elected for four years, is largely a figurehead. He appoints the prime minister, who in turn selects his Cabinet with the advice and consent of the Majles. None of these can be chosen without the prior approval of the *valiy-e faqih;* any can be dismissed upon the recommendation of either the Supreme Court or the Majles.

The 270-member Majles is elected by secret ballot for four years; recognized minorities have token representation. A 12-member Council of Guardians with staggered terms determines the constitutionality of all laws passed by the Majles, as well as their conformity to Islāmic principles; six theologians are appointed by the *valiy-e faqih,* and six other Islāmic jurists are nominated by the High Council of the Judiciary and approved by the Majles. No legislation may be enacted unless the council is in session.

The judiciary consists of a Supreme Court, a Supreme Judicial Council, and lower courts. The chief justice and the prosecutor general must be Shīʿite jurists (*mujtahids*). Individual rights—such as freedom of press, assembly, and expression—are guaranteed within the framework of the Sharīʿah (Islāmic law). Under the constitution all Islāmic judges rely on the Sharīʿah. In 1982 any portion of the law codes of the monarchy that in the opinion of the Supreme Court did not conform with Islām were declared null and void. In 1983 the Majles revised the penal code and instituted a system of *qiṣāṣ* ("retribution"). Punish-

ments, including amputation of limbs and execution, may be carried out by a member of the injured party's family or by the state.

The *ostāns* (provinces) are subdivided into *shahrestāns* (counties), *bakhshs* (districts), and *dehestāns* (villages). Governors general (for provinces) and governors (for counties) are nominated by the minister of interior and appointed by the president. At each level there is a council, and the Supreme Council of Provinces is formed from representatives of the provincial councils. The Ministry of the Interior appoints each city's mayor, and city councilmen are locally elected. Villages are administered by a village master advised by elders.

Under the constitution elections are to be held at least every four years, supervised by the Ministry of Interior. Suffrage is universal and the minimum voting age is 16. All important matters are subject to referenda. Candidates for the presidency and the legislature are nominated by political parties. The most important among these is the Islāmic Republican Party, which controls the Majles. The Muslim People's Party, which once claimed more than 3,000,000 members, became inactive after 1981. Several parties—including the Tudeh (Communist) Party, the Mujāhedīn-e Khalq (Fighters for the People) Party, and the Democratic Party of Iranian Kurdistan—have been outlawed.

Political parties

In the absence of an organized ecclesiastical hierarchy, the religious scholars (*'ulemā*, or mullahs) have always been respected throughout the Islāmic world in general and in Iran in particular. In Iran their position is symbolized by the turban, a black one signifying direct descent from the Prophet Muḥammad. At all levels of government the mullahs exercise a great deal of influence either through direct participation in the political process or because of their ability to sway popular opinion.

Early in its existence the Islāmic regime declared its neutrality between East and West and expressed its intention to strengthen its ties with the Third World, particularly with other Islāmic countries. While the value of foreign trade with this latter group has increased, Iran remains heavily dependent upon the industrial countries of the world. It has retained membership in the United Nations and its affiliated agencies.

Revolutionary Guards

Armed forces. The Revolutionary Guards (Pāsdārān-e Inqilāb), organized in the early days of the republic, are the country's most effective military force. The Ministry of Revolutionary Guards operates separately from the Ministry of Defense. Security forces involved in war or in armed internal conflicts are either accompanied by Revolutionary Guards or led by them. Law enforcement in the cities is the responsibility of the national police force, and the gendarmerie oversees rural areas. Both are under the direction of the Ministry of Interior.

Education. Compulsory education through the fifth grade was established in 1943. All schools and universities reflect the Islāmic system of beliefs, except for a small number of private and parochial schools that are maintained by minority groups. The University of Tehrān and 25 training colleges were opened in 1934. Several more universities, teachers' colleges, and technical schools have been established since then.

The public school system is controlled by the Ministry of Education and Training. Each university is either controlled by the state or has a board of regents appointed by the president of the republic.

Advances in public health

Health and welfare. Health conditions appreciably improved after World War II through the combined efforts of the government, international agencies, and philanthropic endeavour. By 1964 smallpox had been eradicated, plague had disappeared, and malaria had been practically wiped out. Cholera, believed to have been controlled, broke out in 1970 and again in 1981 but was speedily checked. Health facilities, nevertheless, are far from adequate. There is a severe shortage, especially in rural areas, of doctors, nurses, and medical supplies.

Public hospitals provide free treatment for the poor. These are supplemented by private institutions, but all are inadequate. All health services are supervised by the Ministry of Health, the branch offices of which are headed by certified physicians.

The flow of population to the cities has created housing shortages. In major cities purified water is piped into the houses, while small towns and villages rely on wells, *qanāt*s, springs, or rivers. Central heating is not common except for modern buildings in major cities. Portable kerosene heaters, iron stoves using wood and coal, and charcoal braziers are common sources of heat. Welfare is administered by the Ministry of State for Welfare and the Mostaz'afān ("Downtrodden") and Shahid ("Martyr") foundations, the latter being particularly concerned with families of war casualties.

CULTURAL LIFE

Suffering, tears, self-flagellation, and martyrdom are the essential elements of the Shī'ite culture. As members of a minority religion and a protest movement in the early years of Islām, the Shī'ites were persecuted by the Sunnite majority. The martyrdom of al-Ḥusayn ibn 'Alī in AD 680 at Karbalā', Iraq, at the hands of the Umayyads, the sufferings of al-Ḥusayn's household, and the massacre of his followers became the rallying point for a national revival after the establishment of the Ṣafavid Empire in 1501. The Shī'ite Persians were then pitted against the Sunnite Ottomans. At the beginning of each lunar Islāmic year the Karbalā' tragedy is commemorated by *ta'ziya* (passion plays) and self-flagellation with chains, scimitars, and bare hands. The participants are often clothed in bloody shrouds signifying the blood-stained soil of Karbalā' and their willingness to join the ranks of the martyrs.

The martyrdom of al-Ḥusayn

The commemoration of Karbalā' has permeated all of Persian culture and finds expression in poetry, music, and the pessimistic Shī'ite view of the world. No religious ceremony is complete without a reference to Karbalā', and no month passes without at least one day of mourning. None of the efforts of the monarchy, such as the annual festivals of art and the encouragement of musicians and native craftsmen, succeeded in changing this basic attitude, which finds laughter and joy undesirable and, in some circles, even sinful.

There are, however, several festive occasions, the most important being *Nō-rūz*, the Persian New Year, and the birthday of the Twelfth Imām, whose second return the Shī'ites expect. The New Year celebration begins on the last Wednesday of the old year, is followed by a week-long holiday and continues until the 13th day of the new year, which is a day for picnicking in the countryside. On the Twelfth Imām's birthday cities sparkle with lights, and the bazaars are decorated and teem with shoppers.

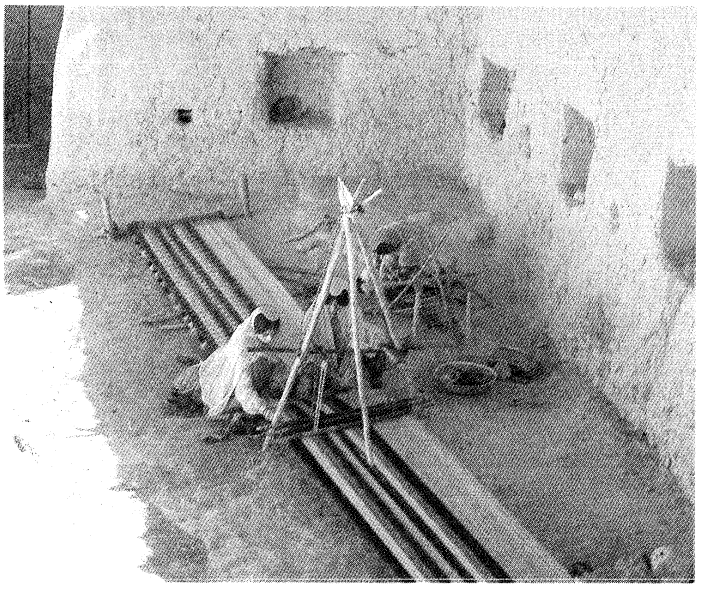
Carpet looms dot the country. Each locality prides itself on a special design and quality of carpet that bears its name, such as Kāshān (Kashan), Kermān, Eṣfahān, Shīrāz, Tabrīz, and Qom. Carpets are used locally and are exported. The handwoven-cloth industry has survived stiff competition from the country's modern textile mills. Weavers produce velvets, printed cottons, wool brocades, shawls, and cloth shoes. Felt is made in the south and sheepskin is embroidered in the northeast.

Carpet weaving and cloth manufacture

A wide range of articles, both utilitarian and decorative, are made of various metals. The most famous centres are Tehrān (gold); Shīrāz, Eṣfahān, and Zanjān (silver); and Kāshān and Eṣfahān (copper). Khorāsān is known for its turquoise industry, as is the Persian Gulf region for natural pearls. The craft techniques are as divergent as the products themselves. Articles may be cast, beaten, wrought, pierced, or drawn (stretched out). The most widespread techniques for ornamentation are engraving, embossing, chiselling, damascening, encrustation, or gilding.

Numerous decorative articles in wood are produced for both the domestic and export markets in Eṣfahān, Shīrāz, and Tehrān (inlay) and in Rasht, Orūmīyeh (formerly called Reẓā'īyeh), and Sanandaj (carved and pierced wood). Machine-made ceramic tiles are produced in Tehrān, but handmade tiles and mosaics, known for their rich designs and beautiful colours, continue to be produced.

Stone and clay are also used for the production of a wide range of household utensils, trays, dishes, and vases. Mashhad is the centre of the stone industry. Potteries are widely scattered throughout the country, Hamadān being the largest centre.



Cloth being woven by Qashqā'ī women of the Shīrāz area.
© R. & S. Michaud 1982—Woodfin Camp & Associates

Architecture

Under the monarchy two architectural trends developed—an imitation of Western styles that had little relevance to the country's climate and landscape, and an attempt to revive indigenous designs. The National Council for Iranian Architecture, founded in 1967, discouraged blind imitation of the West and promoted the use of more traditional Iranian styles that were modified to serve modern needs.

The strict Islāmic interpretation of the Bible's Second Commandment, which the Qur'ān reiterates, actually discouraged painting and sculpture. This, however, did not prevent Iranian artists from working in other media, such as calligraphy, illumination, weaving, ceramics, and metalwork. Western classical painting and sculpture were introduced in the late 19th century and were adapted to Iranian themes.

Music

For centuries musical development was inhibited because of Islāmic injunctions. Folk songs and ancient Persian classical music were preserved only through oral transmission from generation to generation. It was not until the 20th century that a music conservatory was founded in Tehrān and that Western techniques were used to record traditional melodies and encourage new compositions. This trend was reversed, however, in 1979, when the former restrictions on music were restored.

A motion-picture industry had developed before the revolution, with strong support from the Ministry of Higher Education and Culture. Production by the private sector has been discontinued, and the industry now concentrates on documentaries and on historical and religioethical themes. It operates under the supervision of the Ministry of Islāmic Guidance.

Museums are not numerous and are relatively recent, with two exceptions: the Golestān Palace Museum in Tehrān, which was opened in 1894, and the All Saviour's Cathedral Museum of Jolfā (Eṣfahān), which was built by the Armenian community in 1905. The only gallery devoted solely to art is the Tehrān Museum of Modern Art, opened in 1977. Other well-known museums include Iran Bāstān (1946) and Negarestān (1975) in Tehrān and Pārs in Shīrāz (1938). Among the learned societies, all of which are located in Tehrān, the most important are the Ancient Iran Cultural Society, the Iranian Mathematical Society, and the Iranian Society of Microbiology. There are also a number of research institutes, such as those for cultural, scientific, archaeological, anthropological, and historical research. In addition to the libraries of the various universities, there are public and private libraries in Tehrān, Mashhad, Eṣfahān, and Shīrāz.

Wrestling, horse racing, and ritualistic bodybuilding are the traditional sports of the country. Team sports were introduced from the West in the 20th century, the most popular being rugby and volleyball.

The development of municipal and amusement parks gained some momentum after World War II. Playgrounds in major cities are few, however, nor are there many public campgrounds. Several major cities have large sports stadiums.

The country's daily newspapers and periodicals are published primarily in Tehrān and must be licensed under the press law of 1979. The publication of any non-Muslim or anti-Muslim sentiments is strictly forbidden. Iran operates the Islāmic Republic News Agency (IRNA). Foreign correspondents are allowed into the country on special occasions. Despite constitutional guarantees of freedom of the press, censorship is ubiquitous.

Radio and television

Radio and television broadcasting stations serve Iran, and some radio broadcasts have international reception. In addition to Fārsī and foreign languages, transmitters broadcast in local languages and dialects. Because of widespread illiteracy, audiovisual media are much more effective than the press in the dissemination of information.

For statistical data on the land and people of Iran, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (Kh.Mo.)

History

This section covers the history of the Iranian Plateau from the beginning of recorded time until the present. It covers as well the succession of ancient empires centred there, the borders of which extended beyond those of the present country.

The early history of Iran may be divided into three phases: (1) the prehistoric period beginning with the earliest evidence of man on the Iranian Plateau (*c.* 100,000 BC) and ending roughly at the start of the 1st millennium BC; (2) the protohistoric period covering approximately the first half of the 1st millennium BC; and (3) the period of the Achaemenid dynasty (6th to 4th century BC), when Iran entered the full light of written history. The civilization of Elam, centred off the plateau in lowland Khūzestān, is an exception, for written history began there as early as it did in neighbouring Mesopotamia (*c.* 3000 BC).

Archaeological sources

The sources for the prehistoric period are entirely archaeological. Early excavation in Iran was limited to a few sites. In the 1930s archaeological exploration increased rapidly, but work was abruptly halted by the outbreak of World War II. After the war ended, interest in Iranian archaeology revived quickly, and since 1950 numerous excavations have revolutionized the study of prehistoric Iran.

For the protohistoric period the historian is still forced to rely primarily on archaeological evidence, but much information comes from written sources as well. None of these sources, however, is both local to and contemporary with the events described. Some sources are contemporary but belong to neighbouring civilizations that are only tangentially involved in events in the Iranian Plateau; for example, the Assyrian and Babylonian cuneiform records from lowland Mesopotamia. Some are local but not contemporary, such as the traditional Iranian legends and tales that supposedly speak of events in the early 1st millennium BC. And some are neither contemporary nor local but are nevertheless valuable in reconstructing events in the protohistoric period (*e.g.*, the 5th-century-BC Greek historian Herodotus).

For the study of centuries of the Achaemenid dynasty there is sufficient documentary material, so that this period is the earliest for which archaeology is not the primary source of data. Economic texts from Mesopotamia, Elam, and Iran; historical inscriptions such as that of Darius the Great at Bīsitūn; contemporary and later classical authors; and later Iranian legends and literature contribute, among other sources, to the understanding of the Achaemenid period.

THE ELAMITES, MEDIANS, AND ACHAEMENIDS

The prehistoric period. *The Paleolithic.* Enigmatic evidence of man's presence on the Iranian Plateau as early as Lower Paleolithic times comes from a surface find in the

Kermanshah Valley. The first well-documented evidence of human habitation is in deposits from several excavated cave and rock-shelter sites, mainly located in the Zagros Mountains of western Iran, dated to Middle Paleolithic or Mousterian times (*c.* 100,000 BC). There is every reason to assume, however, that future excavations will reveal Lower Paleolithic man in Iran. The Mousterian flint-tool industry found there is generally characterized by an absence of the Levallois technique of chipping flint and thus differs from the well-defined Middle Paleolithic industries known elsewhere in the Near East. The economic and social level associated with this industry is that of fairly small, peripatetic hunting and gathering groups spread out over a thinly settled landscape.

The Mousterian culture

Locally, the Mousterian is followed by an Upper Paleolithic flint industry called the Baradostian. Radiocarbon dates suggest that this is one of the earliest Upper Paleolithic complexes; it may have begun as early as 36,000 BC. Its relationship to neighbouring industries, however, remains unclear. Possibly, after some cultural and typological discontinuity, perhaps caused by the maximum cold of the last phase of the Würm glaciation, the Baradostian was replaced by a local Upper Paleolithic industry called the Zarzian. This tool tradition, probably dating to the period 12,000 to 10,000 BC, marks the end of the Iranian Paleolithic sequence.

The Neolithic. Evidence indicates that the Near East in general was one of the earliest areas in the Old World to experience what the Australian archaeologist V. Gordon Childe called the Neolithic revolution. That revolution witnessed the development of settled village agricultural life based firmly on the domestication of plants and animals. Iran has yielded much evidence on the history of these important developments. In the early Neolithic (sometimes called the Mesolithic), evidence of significant shifts in tool manufacture, settlement patterns, and subsistence methods, including the fumbling beginnings of domestication of both plants and animals, comes from such important western Iranian sites as Āsīāb, Gūrān, Ganj-e Dareh, and Ali Kosh. Similar developments in the Zagros, on the Iraqi side of the modern border, are also traceable at sites such as Karīm Shahīr and Zawi Chemi–Shanidar. This phase of early experimentation with sedentary life and domestication was soon followed by a period of fully developed village farming as defined at important Zagros sites such as Jarmo, Sarāb, upper Ali Kosh, and upper Gūrān. All of these sites date wholly or in part to the 8th and 7th millennia.

By approximately 6000 BC these patterns of village farming were widely spread over much of the Iranian Plateau and in lowland Khuzistan. Tepe Sabz in Khuzistan, Hajji Firuz in Azerbaijan, Godin Tepe VII in northeastern Luristan, Tepe Sialk I on the rim of the central salt desert, and Tepe Yahya VI C–E in the southeast have all yielded evidence of fairly sophisticated patterns of agricultural life (Roman numerals identify the level of excavation). Though distinctly different, all show general cultural connections with the beginnings of settled village life in neighbouring areas such as Afghanistan, Baluchistan, Soviet Central Asia, and Mesopotamia.

The 5th to mid-3rd millennia. Rather less is known of the cultures in this time range in Iran than of contemporary cultures elsewhere in the ancient Near East. Research has tended to concentrate on the Neolithic and protohistoric periods, and the scattered evidence for important cultural and artistic developments in the Chalcolithic and Early Bronze ages resists coherent summary. It is clear that trends that began in the late Neolithic period continued in the millennia that followed and that the rugged, broken landscape of the Iranian Plateau forced man into a variety of relatively isolated cultures. In no instance, with the important exception of Elam (see *The Elamites* below), did Iran participate in the developments that led to fully urban civilization in lowland Mesopotamia to the west or in the Indus Valley to the east. Throughout prehistory the Iranian Plateau remained at the economic and cultural level of village life achieved in the Neolithic. The separate cultural areas on the plateau are as yet barely understood by the modern archaeologist in any terms other than through the painted pottery assemblages found at several sites throughout Iran. Though developing in comparative isolation, each of these areas does yield some evidence for cultural contact with its immediate neighbours and, in some striking cases, with developments in the centres of higher civilization in Mesopotamia. Trade would appear to be the principal mechanism by which such contacts were maintained, and often Elam appears to have acted as an intermediary between Sumer and Babylon on the one hand and the plateau cultures on the other. Trade across the northern part of the plateau, through the sites of Hissar and Sialk, most probably involved the transshipment from Afghanistan to Mesopotamia of semiprecious stones such as lapis lazuli. The appearance of proto-Elamite tablets in Sialk IV may bear witness to such trade. So also may the appearance of similar proto-Elamite tablets at Tepe Yahya south of Kermān and in the great central desert provide evidence of trade connections between Mesopotamia and the east—in this case a trade that may have centred on specific items such as steatite and copper. The province of Fārs (Parsa) perhaps also participated in such trade networks, as is suggested by the appearance there, alongside strictly local ceramics, of wares that have clear Mesopotamian affinities. In the west central Zagros, outside influences from both the north and the west can be traced in the ceramic record; such is also the case for local cultures in Azerbaijan, to the northwest. In general, however, these millennia represent a major dark age in Iranian prehistory and warrant considerably more attention than they have recently received.

Trading contacts

The late 3rd and 2nd millennia. The beginning of this period is generally characterized by an even more marked isolation of the plateau than earlier, while the latter half of the period is one of major new disruptions, heretofore unique in Iranian history, that laid the groundwork for developments in the protohistoric period. In northwestern and central western Iran, local cultures, as yet barely defined beyond their ceramic parameters, developed in relative isolation from events elsewhere. All occupation had ceased at Tepe Sialk, but the painted pottery cultures characteristic of earlier Hissar and of the sites on the Gurgan Plain in the northeast continued. Little Mesopotamian influence is evident, though some contacts between Elam and the plateau remain. Perhaps beginning as early as 2400 BC but more probably somewhat later, a radical transformation occurred in the culture of the northeast: earlier painted potteries are entirely replaced by a distinctive gray or gray-black ceramic that is associated with a variety of other artifacts, primarily weapons and ornaments in copper or bronze, also unique. Whether this cultural change represents a strictly local development or testifies to an important intrusion of new peoples into the area is still under debate. In any case, none of these developments can be traced to Mesopotamia or to other areas to the west, regions which had previously been the source of outside influences on the Iranian Plateau. Somewhat later, the local cultures of central and northwestern Iran were apparently influenced by developments in northern Mesopotamia and Assyria along patterns of contact that were well established in earlier periods. Yet this contact, as it is observed at Godin III, Hasanlu VI, and Dinka Tepe, did not cause any major dislocation of local cultural patterns. In the second half of the 2nd millennium, however, western Iran—at first perhaps gradually, and then with striking suddenness—came under the influence of the gray and gray-black ware cultures that had developed earlier in the northeast. There the impact of these influences was such as definitely to suggest a major cultural dislocation and the introduction of a whole new culture and probably a new people into the Zagros. It was this development that marked the end of the Bronze Age in western Iran and ushered in the early protohistoric period.

The Elamites. Whereas the Iranian Plateau did not experience the rise of urban, literate civilization in the late 4th and early 3rd millennia on the Mesopotamian pattern, lowland Khuzistan did. There was centred Elamite civilization.

Geographically, Elam included more than Khuzistan; it was a combination of the lowlands and the immediate highland areas to the north and east. Elamite strength was based on an ability to hold these various areas together under a coordinated government that permitted the maximum interchange of the natural resources unique to each region. Traditionally this was done through a federated governmental structure.

Law of succession

Closely related to that form of government was the Elamite system of inheritance and power distribution. The normal pattern of government was that of an overlord ruling over vassal princes. In earliest times the overlord lived in Susa, which functioned as a federal capital. With him ruled his brother closest in age, the viceroy, who usually had his seat of government in the native city of the currently ruling dynasty. This viceroy was heir presumptive to the overlord. Yet a third official, the regent or prince of Susa (the district), shared power with the overlord and the viceroy. He was usually the overlord's son or, if no son was available, his nephew. On the death of the overlord, the viceroy became overlord. The prince of Susa remained in office, and the brother of the old viceroy nearest to him in age became the new viceroy. Only if all brothers were dead was the prince of Susa promoted to viceroy, thus enabling the overlord to name his own son (or nephew) as the new prince of Susa. Such a complicated system of governmental checks, balances, and power inheritance often broke down despite bilateral descent and levirate marriage (*i.e.*, the compulsory marriage of a widow to her deceased husband's brother). What is remarkable is how often the system did work; it was only in the Middle and Neo-Elamite periods that sons more often succeeded fathers to power.

Elamite history can be divided into three main phases: the Old, Middle, and Late, or Neo-Elamite, periods. In all periods Elam was closely involved with Sumer, Babylonia, and Assyria, sometimes through peaceful trade, more often through war. In like manner, Elam was often a participant in events on the Iranian Plateau. Both involvements were related to the combined need of all the lowland civilizations to control the warlike peoples to the east and to exploit the economic resources of the plateau.

The Old Elamite period. The earliest kings in the Old Elamite period may date to approximately 2700 BC. Already conflict with Mesopotamia, in this case apparently with the city of Ur, was characteristic of Elamite history. These early rulers were succeeded by the Awan (Shustar) dynasty. The 11th king of this line entered into treaty relations with the great Naram-Sin of Akkad (*c.* 2254–*c.* 2218 BC). Yet there soon appeared a new ruling house, the Simash dynasty (Simash may have been in the mountains of southern Luristan). The outstanding event of this period was the virtual conquest of Elam by Shulgi of the 3rd dynasty of Ur (2094–2047 BC). Eventually the Elamites rose in rebellion and overthrew the 3rd Ur dynasty, an event long remembered in Mesopotamian dirges and omen texts. About the middle of the 19th century BC, power in Elam passed to a new dynasty, that of Eparti. The third king of this line, Shirukdukh, was active in various military coalitions against the rising power of Babylon, but Hammurabi (more correctly, Hammurapi, *c.* 1792–1750 BC) was not to be denied, and Elam was crushed in 1764 BC. The Old Babylon kingdom, however, fell into rapid decline following the death of Hammurabi, and it was not long before the Elamites were able to gain revenge. Kutir-Nahhunte I attacked Samsuiluna (1749–1712 BC), Hammurabi's son, and dealt so serious a defeat to the Babylonians that the event was remembered more than 1,000 years later in an inscription of the Assyrian king Ashurbanipal. It may be assumed that with this stroke Elam once again gained independence. The end of the Eparti dynasty, which may have come in the late 16th century BC, is buried in silence.

Conquest by Ur

The Middle Elamite period. After two centuries for which sources reveal nothing, the Middle Elamite period opened with the rise to power of the Anzanite dynasty, whose homeland probably lay in the mountains northeast of Khuzistan. Political expansion under Khumban-numena (*c.* 1285–*c.* 1266 BC), the fourth king of this line, proceeded apace, and his successes were commemorated by his assumption of the title "Expander of the Empire." He was succeeded by his son, Untash-Gal (Untash (d) Gal, or Untash-Huban), a contemporary of Shalmaneser I of Assyria (*c.* 1274–*c.* 1245 BC) and the founder of the city of Dur Untashi (modern Choga Zanbil). In the years immediately following Untash-Gal, Elam increasingly found itself in real or potential conflict with the rising power of Assyria. Tukulti-Ninurta I of Assyria (*c.* 1244–*c.* 1208 BC) campaigned in the mountains north of Elam. The Elamites under Kidin-Khutran, second king after Untash-Gal, countered with a successful and devastating raid on Babylonia. In the end, however, Assyrian power seems to have been too great. Tukulti-Ninurta managed to expand, for a brief time, Assyrian control well to the south in Mesopotamia, Kidin-Khutran faded into obscurity, and the Anzanite dynasty came to an end.

After a short period of dynastic troubles, the second half of the Middle Elamite period opened with the reign of Shutruk-Nahhunte (*c.* 1160 BC). Two equally powerful and two rather less impressive kings followed this founder of a new dynasty, whose home was probably Susa, and in this period Elam became one of the great military powers of the Near East. Tukulti-Ninurta died in *c.* 1208 BC, and Assyria fell into a period of internal weakness and dynastic conflict. Elam was quick to take advantage of this situation by campaigning extensively in the Diyālā River area and into the very heart of Mesopotamia. Shutruk-Nahhunte captured Babylon and carried off to Susa the stele on which was inscribed the famous law code of Hammurabi. Shilkhak-In-Shushinak, brother and successor of Shutruk-Nahhunte's eldest son, Kutir-Nahhunte, still anxious to take advantage of Assyrian weakness, campaigned as far north as the area of modern Kirkūk. In Babylonia, however, the 2nd dynasty of Isin led a native revolt against such control as the Elamites had been able to exercise there, and Elamite power in central Mesopotamia was eventually broken. The Elamite military empire began to shrink rapidly. Nebuchadrezzar I of Babylon (*c.* 1124–*c.* 1103 BC) attacked Elam and was just barely beaten off. A second Babylonian attack succeeded, however, and the whole of Elam was apparently overrun, ending the Middle Elamite period.

Period of military greatness

It is noteworthy that during the Middle Elamite period the old system of succession to, and distribution of, power appears to have broken down. Increasingly, son succeeded father, and less is heard of divided authority within a federated system. This probably reflects an effort to increase the central authority at Susa in order to conduct effective military campaigns abroad and to hold Elamite foreign conquests. The old system of regionalism balanced with federalism must have suffered, and the fraternal, sectional strife that so weakened Elam in the Neo-Elamite period may have had its roots in the centrifugal developments of the 13th and 12th centuries.

The Neo-Elamite period. A long period of darkness separates the Middle and Neo-Elamite periods. In 742 BC, a certain Huban-nugash is mentioned as king in Elam. The land appears to have been divided into separate principalities, with the central power fairly weak. The next 100 years witnessed the constant attempts of the Elamites to interfere in Mesopotamian affairs, usually in alliance with Babylon, against the constant pressure of Neo-Assyrian expansion. At times they were successful with this policy, both militarily and diplomatically. But on the whole they were forced to give way to increasing Assyrian power. Local Elamite dynastic troubles were from time to time compounded by both Assyrian and Babylonian interference. Meanwhile, the Assyrian army whittled away at Elamite power and influence in Luristan. In time these internal and external pressures resulted in the near total collapse of any meaningful central authority in Elam. In a series of campaigns between 692 and 639 BC, in an effort to clean up a political and diplomatic mess that had become a chronic headache for the Assyrians, Ashurbanipal's armies utterly destroyed Susa, pulling down buildings, looting, and sowing the land of Elam with salt. (T.C.Y., Jr.)

Defeat by the Assyrians

Art and architecture. The most artistically important, though not the most ancient, prehistoric painted pottery



Metalwork of Urartu.
(Top left) Fragment of a bronze model of a city, from Toprakkale, near Van, Turkey, late 8th century BC. Height 30.5 cm. (Top centre) Cast of a gold medallion representing a goddess and a worshipper, original found at Toprakkale. (Right) Bronze figurine of a god, from Van, early 7th century BC. Height 19.7 cm. (Bottom) Bronze relief showing scenes from the first campaign of Shalmaneser III against Urartu in *c.* 859 BC, from gates found at Balawat in Assyria (now Iraq). Height 24 cm. All in the British Museum.
By courtesy of the trustees of the British Museum

of Iran is derived from Susiana. The stylized shapes of animals and birds are used, their curves and contrasting angularities skillfully adapted to the sensitive shapes of the clay vessels in a manner that implies a long history of evolutionary experiment. The simplified silhouettes, which once must have appeared as individual figures, are linked together in a repetitive frieze to form a band of enrichment or to articulate some plastic feature. Nothing of this sort is to be seen in the contemporary Ubaid pottery of southern Mesopotamia, although the two cultures show a parallel progression.

In the succeeding Protoliterate period each produced an independent form of pictographic writing. In Iran, this development took place in Elam, a region bordering on southern and central Mesopotamia. The Elamite use of pictographs was short-lived, however, and for a long time no further attempt was made to develop a written language. Modern knowledge of Elamite history during the centuries that followed thus depends exclusively on references in Mesopotamian literature, occasioned by the perpetual contacts between the two countries. Trade was on a large scale, and even the wars that so frequently interrupted it supplied the Elamites with Mesopotamian goods, including works of art that could serve as a pattern for Iranian artists. The stelae of Naram-Sin and Hammurabi were among major works brought as loot to the Elamite capital at Susa, where they have since been found. It is thus understandable that the native products of Elamite artists in the 3rd millennium BC—

Mesopotamian influence

statues, reliefs, and smaller objects, such as cylinder seals—followed the conventions current in Mesopotamia. One interesting exception is a type of carved steatite bowl, at first thought to be an Elamite product but since found at sites as far apart as Mari on the Euphrates and Mohenjo-daro in the Indus Valley, where it originated, thus emphasizing the part played by Susa in transcontinental trade.

During the late 2nd millennium BC, Elam experienced a period of great prosperity and political stability under a line of powerful rulers. The character of Elamite architecture at this time was revealed by the discovery and excavation of a vast temple complex at Choga-Zambil (Dur Untash) in the Susa area. It was built by King Untash-Gal (*c.* 1265–45 BC) to fulfill the functions, first, of a holy shrine and centre of pilgrimage and, second, of a mausoleum for his family. Its central feature is a remarkably well-preserved ziggurat, 345 feet (105 metres) square and originally 144 feet (44 metres) high, built for the most part of baked bricks each weighing 18 kilograms. Apart from the fact that it has been reconstructed with four receding stages and a summit temple, the structure of the ziggurat has little in common with those in Mesopotamia. As described by one scholar,

> gates in the centre of each face led to diverging staircases up to the first stage, except on the southwest face, where a triple diverging staircase also led up to the second and higher stages.

Gates to the main stairways were flanked by glazed terracotta guardian animals—bulls or griffins. The paved area adjoining the ziggurat had a surrounding wall with im-

posing gateways. A similar outer wall enclosed a number of subsidiary temples, their dedication indicated by cuneiform inscriptions in the Elamite language. The ornamental sculptures from this complex were predominantly of glazed terra-cotta. A more striking example of contemporary Elamite sculpture is the life-size bronze statue of Queen Napirasu from Susa (now in the Louvre).

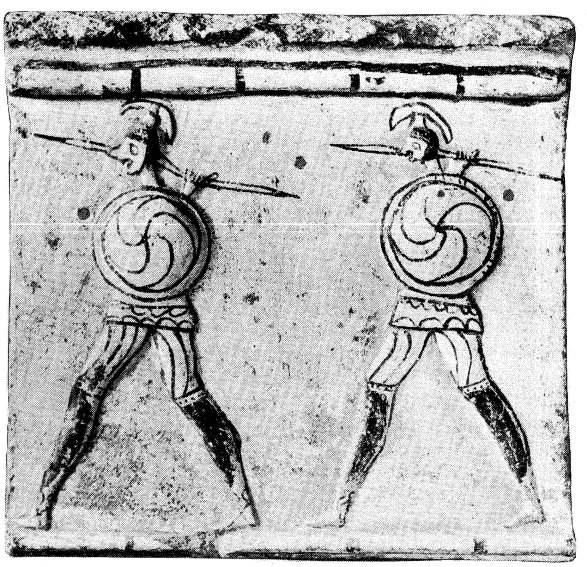
(S.H.F.L.)

The protohistoric period and the kingdom of the Medes. The beginning of the Iron Age is marked by major dislocations of cultural and historical patterns in western Iran (almost nothing is known of the eastern half of the plateau in the Iron Age). The Iron Age itself is divided into three periods: Iron Age I (*c.* 1300–*c.* 1000 BC), Iron Age II (*c.* 1000–*c.* 800/750 BC), and Iron Age III (*c.* 750–*c.* 550 BC). The latter is the archaeological equivalent of what historically can be called the Median period.

The coming of the Iranians. Though isolated groups of speakers of Indo-European languages had appeared and disappeared in western Iran in the 2nd millennium BC, it was during the Iron Age that the Indo-European Iranians rose to be the dominant force on the plateau. By the mid-9th century BC two major groups of Iranians appear in cuneiform sources: the Medes and the Persians. Of the two, the Medes were the more widespread and, from an Assyrian point of view, the more important group. When Assyrian armies raided as far east as modern Hamadān they found only Medes. In the more western Zagros they encountered Medes mixed with indigenous, non-Iranian peoples. Early in the 1st millennium, Iranian Medes already controlled almost all of the eastern Zagros and were infiltrating, if not actually pushing steadily, into the western Zagros, in some areas right up to the edge of the plateau and to the borders of lowland Mesopotamia. Persians (Parsua, Parsuash, Parsumash) also appear in roughly the same areas, though their exact location remains controversial. At times they seem to have settled in the north near Lake Urmia, at times in the central western Zagros near Bakhtaran (formerly Kermānshāh), later certainly in the southwestern Zagros somewhere near the borders of Elam, and eventually, of course, in the province of Fārs. It has been argued that these various locations represent a nomadic tribe on the move; more likely they represent more than one group of Persians. What is reasonably clear from the cuneiform sources is that these Medes and Persians (and no doubt other Iranian peoples not identified by name) were moving into western Iran from the east. They probably followed routes along the southern face of the Elburz Mountains and, as they entered the Zagros, spread out to the northwest and southeast following the natural topography of the mountains. Where they could, as, for example, along the major pass across the mountains from Hamadān to Bakhtaran, they infiltrated farther west. In doing so they met resistance from the local settled populations, who often appealed to Urartu, Assyria, and Elam for assistance in holding back the newcomers. Such appeals were, of course, most welcome to the great powers, who were willing to take advantage of the situation both to advance their interests vis-à-vis each other and to control the Iranian threat to themselves.

Penetration into the Zagros

It has been suggested that the introduction of gray and gray-black pottery into western Iran from the northeast, which signals the start of the Iron Age, is the archaeological manifestation of this pattern of a gradual movement of Iranians from east to west. The case is by no means proved but is a reasonable reading of the combined evidence. If it is so, then the earliest Iranians in the Zagros can be dated to Iron Age I times, *c.* 1300 BC. Archaeologically, the culture of Iron Age II times can be seen as having evolved out of that of the Iron Age I period, and, though the development is less clear, the same can be said of the relationship between the cultures of Iron Age II and III. The spread of the Iron Age I and II cultures in the Zagros is restricted and would appear to correspond fairly well with the distribution of Iranians known from the written documents. The distribution of the Iron Age III culture on the other hand is, at least by the 7th century, much more widespread and covers almost the whole of the Zagros. Thus, the argument that links these archaeological patterns with the Iranian migration into the area associates the Iron Age I and II cultures with the early penetration of the Iranians into the more eastern Zagros and with their infiltration westward along the major routes crosscutting the main mountain alignments. Those areas where traces of the Iron Age I and II cultures do not appear were the regions still under the control of non-Iranian indigenous groups supported by Urartu, Assyria, and Elam. The widespread Iron Age III culture is then associated with the rise to power of the Median kingdom in the 7th and early 6th





Phrygian ornament and architecture.
(Left) Terra-cotta architectural plaque from Pazarlı, Anatolia (now Turkey), 8th century BC. (Right) Rock-cut temple facade at Aslankaya, near Afyonkarahisar, Anatolia.

By courtesy of (left) the Hitit Muzesi Mudurlugu, Ankara; photograph, Arts Council of Great Britain; (right) from E. Akurgal, *Die Kunst Anatoliens*

centuries BC and the Iranianization of the whole of the Zagros. (T.C.Y., Jr.)

Architecture and art. Much less is known about the northwestern part of modern Iran during the 2nd millennium BC, but, from about 1000 BC onward, it acquired a new importance. The southward migration of Indo-European peoples, many centuries earlier, that had brought the Hittites to Anatolia and the Kassites to Babylon was now followed by a new wave of peoples. Referred to as "Iranian," these peoples included both the Persians and the Medes. At first they were not more than a loose confederation of tribes, occupying a wide area of Iran south of the Caspian sea. In settling there they had bypassed the eastern provinces of Urartu and the prosperous country called Mannai, south of Lake Urmia, also avoiding contact with mountain tribes of Luristan. In the 8th and 7th centuries BC, however, the weakening of Urartu and Mannai by Assyrian conquests, followed by a Scythian invasion and the final collapse of Elam, made all of western Iran accessible, and the foundations of Median Empire were laid by Cyaxares.

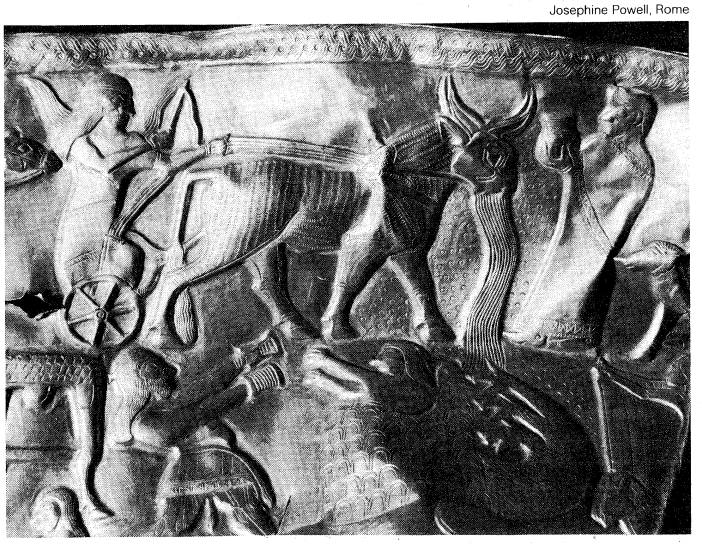
With this complicated infusion of new elements and influences, it is hardly surprising that Iranian art of this period shows a rich synthesis of novel characteristics. Its most conspicuous surviving products are derived from four principal sources: an excavation at Hasanlu, probably the capital of the Mannaeans; a chance-found treasure from Ziwiye (Ziwiyeh) in Iranian Kurdistan; tomb finds at Marlik, near Kazvin; and excavated graves in Luristan.

The Mannaean capital had a powerfully walled citadel, built between 1000 and 800 BC, and within it an enormous palace or temple, consisting of two-story buildings around a partially roofed courtyard. It had been destroyed by fire, and interest in its architecture, which should be considerable, has been eclipsed by the beauty of some objects recovered from the burnt debris. One was a silver cup decorated with electrum figures in two registers; there are chariots and soldiers above and, below, a lion and a horse opposed heraldically between archers. Even more remarkable was a bowl of solid gold (Archaeological Museum, Tehrān), elaborately figured with repoussé ornament, the religious or heroic imagery of which is largely incomprehensible. Though it shows no Assyrian influence, the associations of its linear style are with the West, and some links with Hurrian mythology have been detected in the design.

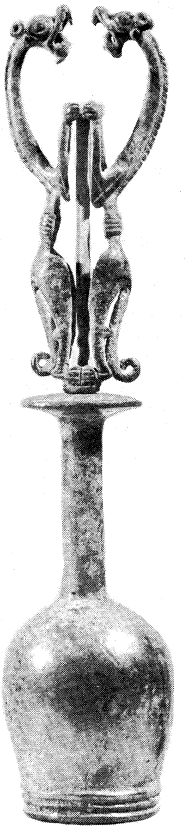
Ziwiye treasure

The hilltop fortress at Ziwiye falls within the territory of the Mannaean state, but the famous treasure found there dates from its later history (7th century BC). The marvellously designed objects making up the treasure include an elaborate pectoral (breast) ornament, a bracelet with lions' heads, a similarly adorned torque (rigid necklace), a ram's

Josephine Powell, Rome



Detail of a Mannaean gold bowl from Hasanlu, Azerbaijan, *c.* 900 BC. In the Archaeological Museum, Tehrān. Diameter of entire bowl 27.9 cm.



Bronze standard and base from Luristan, Iran, *c.* 7th–5th century BC. In the Oriental Institute, the University of Chicago.

By courtesy of the Oriental Institute, the University of Chicago

head rhyton, daggers, and enrichment for furniture—all of gold. As would be expected at this period, they show a variety of styles, sometimes combined in a single design, and carry suggestions of influence from Assyria, Syria, and even Egypt. The collection is thought by some to have been the property of a Scythian chief who temporarily ruled Mannai.

Dated to a period contemporary with the foundation of the Hasanlu citadel, though removed from it geographically, are the royal tombs found at Marlik. The tombs contained gold and silver vessels, comparable in design to those found at Hasanlu, and ingeniously grotesque animal figures in terra-cotta.

The Luristan Bronzes include objects basically homogeneous in style but varying considerably in date and excavated from burial grounds in the eastern Zagros Mountains. There appear to have been more than 400 of these burial grounds, each comprising about 200 graves, so that the number of ornamental bronze objects reaching museums and private collections must have been very great. The burials appear to cover a long period of time; but two main groups have been distinguished, belonging, respectively, to the latter half of the 2nd millennium BC and to the Iron Age from 1000 BC onward. In the second period, particularly, there was a continual movement of Indo-European horsemen migrating from the north—a fact that helps to explain the character of the bronzes themselves, for they are nomads' gear. These portable goods and chattels of wandering tribes must have been produced by the settled craftsmen in the towns and villages of Iran for the nomads who passed through. Chariot or harness fittings are predominant: they include rein rings, bits, and ornaments, as well as weapons and various types of talisman. All suggest an indigenous school of craftsmanship catering to a changing clientele whose tastes and requirements varied with their origin. In the earlier group some very ancient Mesopotamian motifs are seen—rein rings, for instance, resembling those from the Ur tombs—but these have been explained as the belated

use, in the 2nd millennium, of devices long superseded within Sumer itself.

There can be little doubt that, during the first half of the 8th century, when the Medes ruled northwestern Iran from their capital at Ecbatana (Hamadan), they developed some characteristic forms of architecture; this has been confirmed, for instance, by the discovery of a magnificent brick fortress at Nush-e Jan in that area. But the evolution of a style capable of expressing the full genius of Iranian invention fell to the lot of their Persian successors and, fortunately, is better documented by material remains.

(S.H.F.L.)

The kingdom of the Medes. Traditionally, the creator of the Median kingdom was one Deioces who, according to Herodotus, reigned from 728 to 675 BC and founded the Median capital Ecbatana (modern Hamadan). Attempts have been made to associate Daiaukku, a local Zagros king mentioned in a cuneiform text as one of the captives deported to Assyria by Sargon II in 714 BC, with the Deioces of Herodotus, but such an association is highly unlikely. To judge from the Assyrian sources, no Median kingdom such as Herodotus describes for the reign of Deioces existed in the early 7th century BC; at best, he is reporting a Median legend of the founding of their kingdom.

Foundation of the capital

According to Herodotus, Deioces was succeeded by his son Phraortes (675–653 BC), who subjugated the Persians and lost his life in a premature attack against the Assyrians. Some of this tale may be true. Assyrian texts speak of a Kashtariti as the leader of a conglomerate group of Medes, Scythians, Mannaeans, and miscellaneous other local Zagros peoples that seriously threatened the peace of Assyria's eastern borderlands during the reign of Esarhaddon (680–669 BC). It is possible that Phraortes is this Kashtariti, though the suggestion cannot be proven either historically or linguistically. That a Median king in this period exerted political and military control over the Persians is entirely reasonable, though it cannot be proven.

Beginning as early as the 9th century, and with increasing impact in the late 8th and early 7th centuries, groups of nomadic warriors entered western Iran, probably from across the Caucasus. Dominant among these groups were the Scythians, and their entrance into the affairs of the western plateau during the 7th century may perhaps mark one of the important turning points in Iron Age history. Herodotus speaks in some detail of a period of Scythian domination, the so-called Scythian interregnum in Median dynasty history. His dating of this event remains uncertain, but traditionally it is seen as falling between the reigns of Phraortes and Cyaxares and as covering the years 653 to 625 BC. Whether such an interregnum ever actually occurred, and, if it did, whether it should not be dated later than this, are open questions. What is clear is that, by the mid-7th century BC, there were a great many Scythians in western Iran; that they, along with the Medes and other groups, posed a serious threat to Assyria; and that their appearance threw previous power alignments quite out of balance.

Herodotus reports how, under Cyaxares of Media (625–585 BC), the Scythians were overthrown when their kings were induced at a supper party to get so drunk that they were then easily slain. It is more likely that about this time either the Scythians withdrew voluntarily from western Iran, and went off to plunder elsewhere, or that they were simply absorbed into a rapidly developing confederation under Median hegemony. Cyaxares is a fully historical figure, who appears in the cuneiform sources as Uvakhshatra. Herodotus speaks of how Cyaxares reorganized the Median army into units built around specialized armaments: spearmen, bowmen, and cavalry. The unified and reorganized Medes were a match for the Assyrians. They attacked one of the important Assyrian border cities, Arrapkha, in 615 BC, surrounded Ninevah in 614 BC but were unable to capture it, and instead successfully stormed the Assyrian religious capital, Ashur. An alliance between Babylon and the Medes was sealed by the betrothal of Cyaxares' granddaughter to Nabopollassar's son, Nebuchadrezzar II (605–562 BC). In 612 BC the attack on Ninevah was renewed and the city fell in late August (the Babylonians arrived rather too late to participate fully in the battle). The Babylonians and the Medes together pursued the fleeing Assyrians westward into Syria. Assyrian appeals to Egypt for help came to nought, and the last Assyrian ruler, Ashur-uballit II, disappeared from history in 609 BC.

Defeat of the Assyrians

The problem, of course, was how to divide the spoils among the victors. The cuneiform sources are comparatively silent, but it would seem that the Babylonians fell heir to all of the Assyrian holdings within the fertile crescent, while their allies took over all of the highland areas. The Medes gained control over the lands in eastern Anatolia that had once been part of Urartu and eventually became embroiled in war with the Lydians, the dominant political power in western Asia Minor. In 585 BC, probably through the mediation of the Babylonians, peace was established between Media and Lydia, and the Halys (Kızıl) River was fixed as the boundary between the two kingdoms. Thus a new balance of power was established in the Near East among Medes, Lydians, Babylonians, and, far to the south, Egyptians. At his death, Cyaxares controlled vast territories: all of Anatolia to the Halys, the whole of western Iran eastward, perhaps as far as the area of modern Tehrān (Teheran), and all of southwestern Iran, including Fārs. Whether it is appropriate to call these holdings a kingdom is debatable; one suspects that authority over the various peoples, Iranian and non-Iranian, who occupied these territories was exerted in the form of a confederation such as is implied by the ancient Iranian royal title, king of kings.

Astyages followed his father, Cyaxares, on the Median throne (585–550 BC). Comparatively little is known of his reign. All was not well with the alliance with Babylon, and there is some evidence to suggest that Babylonia may have feared Median power. The latter, however, was soon in no position to threaten others, for Astyages was himself under attack. Indeed, Astyages and the Medians were soon overthrown by the rise to power in the Iranian world of Cyrus II the Great of Persia.

The rise of the Persians under Cyrus II. The ruling dynasty of the Persians settled in Fārs in southwestern Iran (possibly the Parsumash of the later Assyrian records) traced its ancestry back to an eponymous ancestor, Hāxamanish, or Achaemenes. There is no historical evidence of such a king's existence. Traditionally, three rulers fall between Achaemenes and Cyrus II: Teispes, Cyrus I, and Cambyses I. Teispes, freed of Median domination during the so-called Scythian interregnum, is thought to have expanded his kingdom and to have divided it on his death between his two sons, Cyrus I and Ariaramnes. Cyrus I may have been the king of Persia who appears in the records of Ashurbanipal swearing allegiance to Assyria after the devastation of Elam in the campaigns of 642–639 BC, though there are chronological problems involved with this equation. When Median control over the Persians was supposedly reasserted under Cyaxares, Cambyses I is thought to have been given a reunited Persia to administer as a Median vassal. His son, Cyrus II, married the daughter of Astyages and in 559 BC inherited his father's position within the Median confederation. Cyrus II certainly warranted his later title, Cyrus the Great. He must have been a remarkable personality, and certainly he was a remarkable king. He united under his authority several Persian and Iranian groups who apparently had not been under his father's control. He then initiated diplomatic exchanges with Nabonidus of Babylon (556–539 BC), which justifiably worried Astyages. Eventually, he openly rebelled against the Medes, who were beaten in battle when considerable numbers of Median troops deserted to the Persian standard. Thus, in 550 BC, the Median Empire became the first Persian Empire, and the Achaemenid kings appeared on the international scene with a suddenness that must have frightened many.

Rise of Cyrus the Great

Cyrus immediately set out to expand his conquests. After apparently convincing the Babylonians that they had nothing to fear from Persia, he turned against the Lydians under the rule of the fabulously wealthy Croesus. Lydian appeals to Babylon were to no avail. He then took Cilicia, thus cutting the routes over which any help might have reached the Lydians.

Croesus attacked and an indecisive battle was fought in 547 BC on the Halys River. Since it was late in the campaigning season, the Lydians thought the war was over for that year, returned to their capital at Sardis, and dispersed the national levy. Cyrus, however, kept coming. He caught and besieged the Lydians in the citadel at Sardis and captured Croesus in 546 BC. Of the Greek city-states along the western coast of Asia Minor, heretofore under Lydian control, only Miletus surrendered without a fight. The others were systematically reduced by the Persian armies led by subordinate generals. Cyrus himself was apparently busy elsewhere, possibly in the east, for little is known of his activities between the capture of Sardis and the beginning of the Babylonian campaign in 540 BC.

Nowhere did Cyrus display his political and military genius better than in the conquest of Babylon. The campaign actually began when he lulled the Babylonians into inactivity during his war with Lydia, which, since it was carried to a successful conclusion, deprived the Babylonians of a potential ally when their turn came. Then he took maximum advantage of internal disaffection and discontent within Babylon. Nabonidus was not a popular king. He had paid too little attention to home affairs and had alienated the native Babylonian priesthood. Second Isaiah, speaking for many of the captive Jews in Babylon, was undoubtedly not the only one of Nabonidus' subjects who looked to Cyrus as a potential deliverer. With the stage thus set, the military campaign against Babylon came almost as an anticlimax. The fall of the greatest city in the Near East was swift; Cyrus marched into town in the late summer of 539 BC, seized the hands of the statue of the city god Marduk as a signal of his willingness to rule as a Babylonian not as a foreign conquerer, and was hailed by many as the legitimate successor to the throne. In one stride Cyrus carried Persian power to the borders of Egypt, for with Babylon came all that it had seized from the Assyrians and had gained in the sequel.

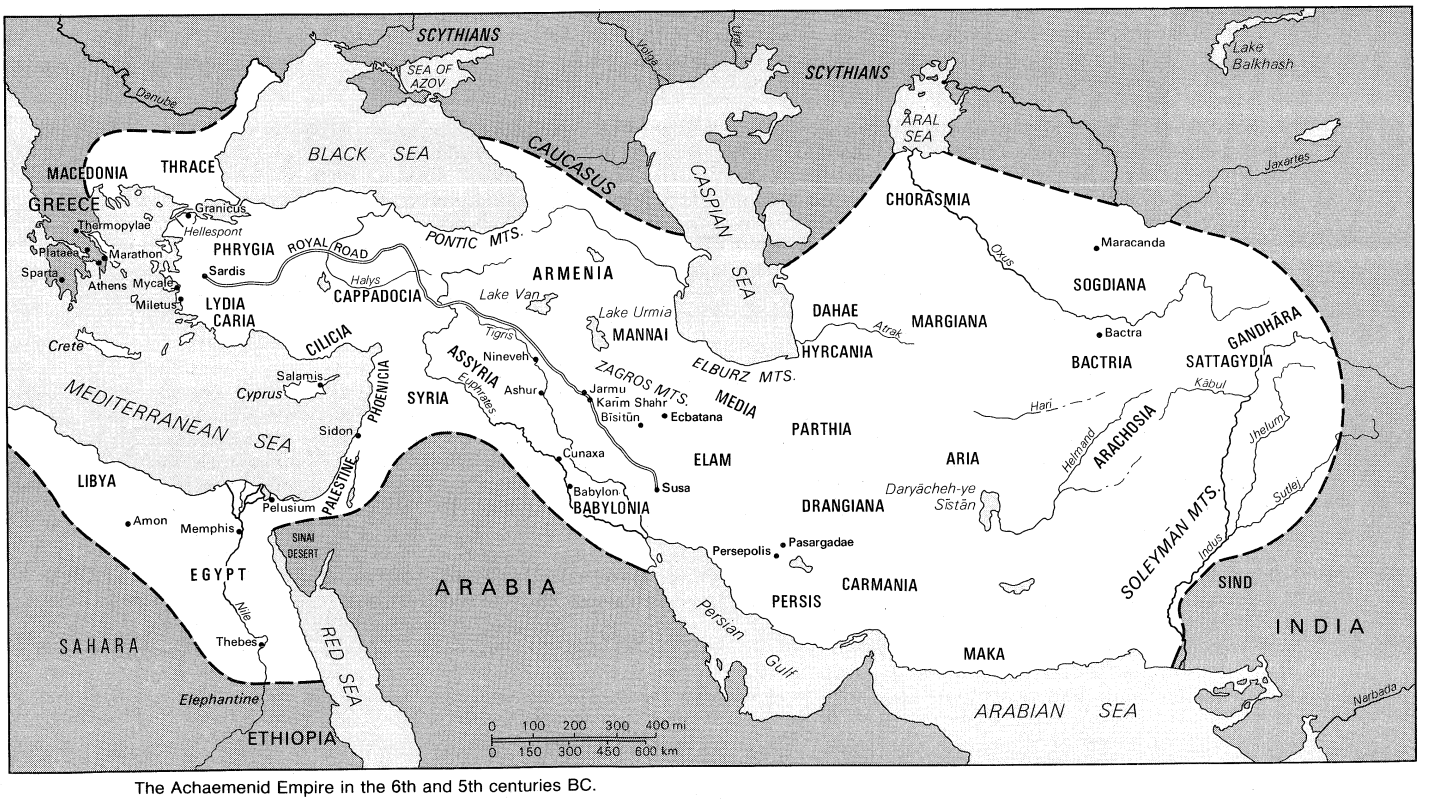
The fall of Babylon

Little is known of the remainder of Cyrus' reign. The rapidity with which his son and successor, Cambyses II, initiated a successful campaign against Egypt suggests that preparations for such an attack were well advanced under Cyrus. But the founder of Persian power was forced to turn east late in his reign to protect that frontier against warlike tribes who were themselves in part Iranians and who threatened the plateau in the same manner as had the Medes and the Persians more than a millennium earlier. One of the recurrent themes of Iranian history is the threat of peoples from the east. How much Cyrus conquered in the east is uncertain. What is clear is that he lost his life in 529 BC, fighting somewhere in the region of the Oxus and Jaxartes rivers.

The Achaemenid dynasty. *Cambyses.* On the death of Cyrus the Great the empire passed to his son, Cambyses II (529–522 BC). There may have been some degree of unrest throughout the empire at the time of Cyrus' death, for Cambyses apparently felt it necessary secretly to kill his brother, Bardiya (Smerdis), in order to protect his rear while leading the campaign against Egypt in 525 BC. The pharaoh Ahmose II of the 26th dynasty sought to shore up his defenses by hiring Greek mercenaries, but was betrayed by the Greeks. Cambyses successfully managed the crossing of the hostile Sinai Desert, traditionally Egypt's first and strongest line of defense, and brought the Egyptians under Psamtik III, son and successor of Ahmose, to battle at Pelusium. The Egyptians lost and retired to Memphis; the city fell to the Persians and the Pharaoh was carried off in captivity to Susa. Three subsidiary campaigns were then mounted, all of which are reported as failures: one against Carthage, but the Phoenician sailors, who were the backbone of the Persian navy, declined to sail against their own colony; one against the oasis of Amon (in the Egyptian desert west of the Nile), which, according to Herodotus, was defeated by a massive sandstorm; and one led by Cambyses himself to Nubia. This latter effort was partly successful, but the army suffered badly from a lack of proper provisions on the return march. Egypt was then garrisoned at three major points: Daphnae in the east delta, Memphis, and Elephantine, where Jewish mercenaries formed the main body of troops.

Cambyses' campaign against Egypt

In 522 BC news reached Cambyses of a revolt in Iran led by an impostor claiming to be Bardiya, Cambyses' brother. Several provinces of the empire accepted the new ruler, who bribed his subjects with a remission of taxes for three years. Hastening home to regain control, Cambyses died—possibly by his own hand, more probably from infection following an accidental sword wound. Darius, a leading general in Cambyses' army and one of the princes of the Achaemenid family, raced homeward with



The Achaemenid Empire in the 6th and 5th centuries BC.

the troops in order to crush the rebellion in a manner profitable to himself.

Cambyses has been rather mistreated in the sources, thanks partly to the prejudices of Herodotus' Egyptian informers and partly to the propaganda motives of Darius I. He is reported to have ruled the Egyptians harshly and to have desecrated their religious ceremonies and shrines. His campaigns out of Egypt were all reported as failures. He was accused of suicide in the face of revolt at home. It was even suggested that he was mad. There is, however, little solid contemporary evidence to support these charges.

Darius I. Darius I, later called the Great, tells the story of the overthrow of Bardiya and of the first year of his own rule in detail in his famous royal inscription cut on a rock face at the base of Bīsitūn mountain, a few miles east of modern Bakhtaran. Six leading Achaemenid nobles assisted in slaying the false Bardiya and together proclaimed Darius the rightful heir of Cambyses. Darius was a member of the Achaemenid royal house. His great-grandfather had been Ariaramnes, son of Teispes, who had shared power in Persia with his brother Cyrus I. Ariaramnes' son, Arsames, and his grandson, Hystaspes (Darius' father), had not been kings in Persia, as unified royal power had been placed in the hands of Cambyses I by Cyaxares. Neither is named a king in Darius' own inscriptions. Hystaspes was, however, an important prince of the blood, who, at the time of revolt of the false Bardiya had apparently been the governor of Parthia. Darius himself was in the mold of Cyrus the Great—a powerful personality and a dynamic ruler.

It took more than a year (522–521 BC) of hard fighting to put down revolts associated with Bardiya's claim to the throne and Darius' succession to power. Almost every province of the empire was involved, including Persia and, most particularly, Media. A balanced policy of clemency backed by the swift and thorough punishment of any captured rebel leader, combined with a well-coordinated and carefully timed distribution of loyal forces, eventually brought peace to the empire and undisputed power to Darius. He then turned his attention to the organization and consolidation of his inheritance, and it was for this role—that of lawgiver and organizer—that he himself, to judge from his inscriptions, most wished to be remembered.

Darius' campaign of expansion

Such activities, however, did not prevent Darius from following an active expansionist policy. Campaigns to the east confirmed gains probably made by Cyrus the Great and added large sections of the northern Indian subcontinent to the list of Persian-controlled provinces. Expansion in the west began about 516 BC when Darius moved against the Hellespont as a first step toward an attack on the Scythians along the western and northern shores of the Black Sea. The real strategic purpose behind this move probably was to disrupt and if possible to interrupt Greek trade with the Black Sea area, which supplied much grain to Greece. Crossing into Europe for the first time, Darius campaigned with comparatively little success to the north of the Danube. He retreated in good order, however, with only limited losses, and a bridgehead across the Hellespont was established.

Perhaps in part in response to these developments, perhaps for more purely internal reasons, the Ionian Greek cities on the west coast of Asia Minor revolted against Persian rule in 500 BC. The Persians were apparently taken by surprise, and at first the rebellion prospered. The Ionians received some limited assistance from the Athenians and in 498 BC felt strong enough to take the offensive. With one hand Darius negotiated; with the other he assembled a counterattack. The first Persian military efforts proved only partially successful, however and the Ionians enjoyed another respite in the years 496–495 BC. A renewed Persian offensive in 494 BC was successful. The Greek fleet was badly beaten off Miletus, and the Persian land army began a systematic reduction of the rebel cities. In *c.* 492 BC Mardonius, a son-in-law of Darius, was made special commissioner to Ionia. He suppressed local tyrants and returned democratic government to many cities. In time the wounds caused by the revolt and its suppression healed, and by 481 BC Xerxes was able to levy troops in this region with little trouble.

By 492 BC Mardonius had also recovered Persian Thrace and Macedonia, first gained in the campaign against the Scythians and lost during the Ionian Revolt. There followed the Persian invasion of Greece that led to Darius' defeat at the Battle of Marathon on August 12, 490 BC. The "Great King" was forced to retreat and to face the fact that the Greek problem, which had probably seemed to the Persians a minor issue on the western extremity of the empire, would require a more concerted and massive effort. Thus began preparations for an invasion of Greece on a grand, coordinated scale. These plans were interrupted in 486 BC by two events: a serious revolt in Egypt, and the death of Darius.

Xerxes I. Xerxes (486–465 BC), Darius' eldest son by Queen Atossa, was born after his father had come to the throne; he had been designated official heir perhaps as early as 498 BC, and while crown prince he had ruled as the King's governor in Babylon. The new King quickly suppressed the revolt in Egypt in a single campaign in 485 BC. Xerxes then broke with the policy followed by Cyrus and Darius of ruling foreign lands with a fairly light hand and, in a manner compatible with local traditions, ruthlessly ignored Egyptian forms of rule and imposed his will on the rebellious province in a thoroughly Persian style. Plans for the invasion of Greece begun under Darius were then still further delayed by a major revolt in Babylonia about 482 BC, which also was suppressed with a heavy hand.

Invasion of Greece

Xerxes then turned his attention westward to Greece. He wintered in Sardis in 481–480 BC and thence led a combined land and sea invasion of Greece. Northern Greece fell to the invaders in the summer of 480, the Greek stand at Thermopylae in August of 480 came to nought, and the Persian land forces marched on Athens, taking and burning the Acropolis. But the Persian fleet lost the Battle of Salamis and the impetus of the invasion was blunted. Xerxes, who had by then been away from Asia rather long for a king with such widespread responsibilities, returned home and left Mardonius in charge of further operations. The real end of the invasion came with the Battle of Plataea, the fall of Thebes (a stronghold of pro-Persian forces), and the Persian naval loss at Mycale in 479 BC. Of the three, the Persian loss at Plataea was perhaps the most decisive. Up until Mardonius was killed, the issue of the battle was probably still in doubt, but once leaderless, the less organized and less disciplined Persian forces collapsed. Time and again in later years this was to be the pattern in such encounters, for the Persians never solved the military problem posed by the disciplined Greek hoplites.

The formation of the Delian League, the rise of Athenian imperialism, troubles on the west coast of Asia Minor, and the end of Persian military ambitions in the Aegean followed rapidly in the decade after Plataea. Xerxes probably lost interest in the proceedings and sank deeper and deeper into the comforts of life in his capital cities of Susa, Ecbatana, and Persepolis. Harem intrigues, which were steadily to sap the strength and vitality of the Achaemenid Empire, led to the assassination of the Great King in 465 BC.

Artaxerxes I to Darius III. The death of Xerxes was a major turning point in Achaemenid history. Occasional flashes of vigour and intelligence by some of Xerxes' successors were too infrequent to prevent eventual collapse but did allow the empire to die gradually. It is a tribute to Cyrus, Cambyses, and Darius that the empire they constructed was as resilient as it proved to be after Xerxes.

The three kings that followed Xerxes on the throne—Artaxerxes I (464–425 BC), Xerxes II (425–424 BC), and Darius II (423–404 BC)—were all comparatively weak individuals and kings, and such successes as the empire enjoyed during their reigns were mainly the result of the efforts of subordinates or of the troubles faced by their adversaries. Artaxerxes I faced several rebellions, the most important of which was that of Egypt in 459 BC, not fully suppressed until 454 BC. An advantageous peace (the Peace of Callias) with Athens was signed in 448 BC, whereby the Persians agreed to stay out of the Aegean and the Athenians agreed to leave Asia Minor to the Achaemenids. Athens broke the peace in 439

BC in an attack on Samos, and in its aftermath the Persians made some military gains in the west. Xerxes II ruled only about 45 days and was killed in a drunken stupor by the son of one of his father's concubines. The assassin was himself killed by Darius II, who rose to the throne through palace intrigue. Several revolts marred his reign, including one in Media, which was rather close to home.

The Peloponnesian War

The major event of these three reigns was the Peloponnesian War between Sparta and Athens that lasted, with occasional pauses, from 460 to 404 BC. The situation was ripe for exploitation by the famous "Persian archers," the gold coins of the Achaemenids that depicted an archer on their obverse and that were used with considerable skill by the Persians in bribing first one Greek state and then another. Initially, the Persians encouraged Athens against Sparta, and from this gained the treaty of Callias. Then, after the disastrous Athenian campaign against Sicily in 413 BC, the Persians intervened on Sparta's side. By the treaty of Miletus in 412 BC, Iran recovered complete freedom in western Asia Minor in return for agreeing to pay for seamen to man the Peloponnesian fleet. Persian gold and Spartan soldiers brought about Athens' fall in 404 BC. Despite the fact that the Persians played the two sides against each other to much advantage, they should have done better. One observes a certain lack of control from Susa by the king in these proceedings, and the two principal governors in Asia Minor who were involved, Tissaphernes of Sardis and Pharnabazus of Hellespontine Phrygia, seemed to have permitted a personal power rivalry to stand in the way of a really coordinated Persian intervention in the Greek war.

Artaxerxes II came to the throne in 404 BC and reigned until 359 BC. The main events of his long rule were the war with Sparta that ended with a peace favourable to the Persians; the revolt and loss to the empire of Egypt; the rebellion of Cyrus the Younger, brother of the king; and the uprising known as the revolt of the satraps.

Sparta, triumphant over Athens, built a small empire of its own and was soon involved in a war against the Persians, the principal issue again being the Greek cities of Asia Minor. While Sparta played one Persian governor in Anatolia against the other, the Persians spent gold in Greece to raise rebellion on Sparta's home ground. The Persians rebuilt their fleet and placed a competent Athenian admiral, Conon, in command. The contest continued from 400 to 387, with Sparta forced to act on an ever-shrinking front. A revitalized Athens, supported by Persia, created a balance of power in Greece, and eventually Artaxerxes was able to step in, at Greek request, and dictate the so-called King's Peace of 387–6 BC. Once again the Greeks gave up any claim to Asia Minor and further agreed to maintain the status quo in Greece itself. When Egypt revolted in 405 BC, Persia was unable to do much about it, and from this point forward Egypt remained essentially an independent state.

Cyrus the Younger, though caught in an assassination attempt at the time of Artaxerxes' coronation, was, nevertheless, forgiven, thanks to the pleadings of the Queen Mother, and was returned to the command of a province in Asia Minor. But he revolted again in 401 BC and, supported by 10,000 Greek mercenaries, marched eastward to contest the throne. He was defeated and killed at the Battle of Cunaxa in Mesopotamia in the summer of 401. The Greek mercenaries, however, were not broken and, though harried, left the field in good order and began their famous march, recorded in the *Anabasis* of Xenophon, north to the Black Sea and home. Probably no other event in late Achaemenid history revealed more clearly to the Greeks the essential internal weakness of the Achaemenid Empire than the escape of so large a body of men from the very heart of the Great King's domain.

Revolt of the satraps

Since 379 BC Greek mercenaries had been gathered together in order to mount a campaign against Egypt. An attack in 373 failed against the native 30th dynasty. On the heels of this failure came the revolt of the satraps. Several satraps, or provincial governors, rose against the central power, and one, Aroandas, a late satrap of Armenia, went so far as to stamp his own gold coinage as a direct challenge to Artaxerxes. The general plan of the rebels appears to have been for a combined attack. The rebel satraps were to coordinate their march eastward through Syria with an Egyptian attack, under the pharaoh Tachos (Zedhor), supported by Greek mercenaries. The Egyptian attack was called off because of a revolt in Egypt by Tachos' brother, and Artaxerxes managed to defeat the satraps who were left alone to face the Great King's wrath. How different would have been the wrath of Darius! Several of the satraps, including Aroandas, were actually forgiven and returned to their governorships. In general the impression is that, in the end, rather than fight the central authority, the satraps were willing to return to their own provinces and plunder there in the name of the Great King. Perhaps they saw that they actually had more authority and more control over real events in their own provincial territories than Artaxerxes had in his empire.

Reign of Artaxerxes III

Plot and counterplot, harem intrigue, and murder brought Artaxerxes III to the throne in 359 BC. He promptly exterminated many of his relatives who might have challenged his rule—all to no avail, for revolts continued to rock the empire. A fresh attempt to win back Egypt was thrown back in 351–350. This setback encouraged revolt in Sidon and eventually in all of Palestine and Phoenicia. Parts of Cilicia joined the rebellion but the revolt was crushed the same year it had begun, 345 BC. Peace was achieved only temporarily; mercenaries from Thebes and the Argives, as well as from the Greek cities of Asia Minor, gathered for a new attempt on Egypt, which, led by Artaxerxes III himself, succeeded in 343 BC. But the local dynasty fled south to Nubia, where it maintained an independent kingdom that kept alive the hopes of a national revival. Persia then misplayed its hand in Greece by refusing aid to Athens against the rising power of Philip II of Macedon. In 339 BC Persian troops were fighting alone in Thrace against the Macedonians, and in the following year, at the Battle of Chaeronea, Philip extended his hegemony over all of Greece—a united Greece that was to prove impervious to Persian gold.

Artaxerxes was poisoned by his physician at the order of the eunuch Bagoas. The latter made Arses king (338–336 BC) in hopes of being the power behind the throne, but Arses did not bend easily to Bagoas' will. He attempted to poison the kingmaker but was himself killed in retaliation. Bagoas then engineered the accession of Darius III, a 45-year-old former satrap of Armenia. So many members of the royal house had been murdered in the court intrigue that Darius probably held the closest blood claim to the throne by virtue of being the grandnephew of Artaxerxes II. Darius was able to put down yet another rebellion in Egypt under Khababash in 337–336 BC, but the beginning of the end came soon afterward, in May 334, with the loss of the Battle of Granicus to Alexander the Great. Persepolis fell to the invader in April 330, and Darius, the last Achaemenid, was murdered in the summer of the same year while fleeing the conqueror. His unfinished tomb at Persepolis bears witness to his lack of preparation.

The invasion of Alexander the Great

Alexander did not win his victories easily, however, and the catalog of troubles that marked the latter part of the Achaemenid Empire—rebellions, murders, weak kings trapped in the harem, missed chances, and foolish policies—cannot be the whole story. The sources, mostly Greek, are often prejudiced against the Persians and tend to view events from but a single point of view. No government could have lasted so long, found its way somehow through so many difficulties, and in the end actually have fought so hard against the conqueror without having much virtue with which to balance its vices.

Achaemenid society and culture. Achaemenid society and culture was in reality the collective societies and cultures of the many subject peoples of the empire. From this mosaic it is sometimes difficult to sort out that which is distinctively Persian or distinctively a development of the Achaemenid period, and therefore perhaps an early Iranian contribution to general Near Eastern society and culture.

Language. The languages of the empire were as varied as its peoples. The Persians, at least originally, spoke Old Persian, a southwestern dialect of Iranian (Median was

a northwestern Iranian dialect), but they were illiterate. Their language was first written when Darius commanded that a script suitable for this purpose be invented so that he might inscribe the record of his rise to power at Bīsitūn (the inscriptions in Old Persian are attributed to earlier kings as either late historical forgeries or as probably written during the reign of Darius). That few could read Old Persian might be the reason why Darius at Bīsitūn established the tradition that royal inscription should be trilingual in Old Persian, Babylonian, and Elamite. Old Persian was never a working written language of the empire. Elamite, written on clay tablets, appears to have been the language of many of the administrators in Fārs and, it may be assumed, in Elam. Archives of administrative documents in Elamite have been found at Persepolis. Aramaic, however, was the language of much of the empire and was probably the language most used in the imperial bureaucracy. The beginnings of the strong influence of Aramaic on Persian, which is so evident in the Middle Persian of Sāsānian times, can already be seen in the Old Persian royal inscriptions of late Achaemenid times.

Social organization. Little is known of Iranian social organization in the period. In general, it was based on feudal lines that were in part drawn by economic and social function. Traditional Indo-Iranian society consisted of three classes, the warriors or aristocracy, the priests, and the farmers or herdsmen. Crosscutting these divisions was a tribal structure based on patrilineal descent. The

Kingship

title king of kings, used even in the 20th century by the shahs of Iran, implies that the central authority exercised power through a pyramidal structure that was controlled at levels below the supreme authority by individuals who were themselves, in a certain sense, kings. Traditionally the king was elected from a particular family by the warrior class; he was sacred, and a certain royal charisma attached to his person.

Such a method of organizing and controlling society undoubtedly changed under the influences and demands of imperial power and underwent much modification as Iranians increasingly borrowed social and political ideas from the peoples they ruled. Even in later times, nevertheless, there is evidence that the original Iranian concepts of kingship and social organization were still honoured and remained the ideals of Persian culture.

Religion. Iranian religion in the pre-Achaemenid and Achaemenid periods is a subject on which there is little scholarly agreement. When the Iranians first entered the semilight of the protohistoric period, they were certainly polytheists whose religious beliefs and practices closely paralleled other Indo-Iranian and Indo-European groups at the same stage in history. Their gods were associated with natural phenomena; with social, military, and economic functions; and with abstract concepts such as justice and truth. Their religious practices included, among others, animal sacrifice, a reverence for fire, and the drinking of the juice of the haoma plant, a natural intoxicant.

Probably around 600 BC there arose in the northeast of the plateau the great Iranian religious prophet and teacher,

Zoroastrianism

Zoroaster. The history of the religion that he founded is even more complicated and controversial than the history of pre-Zoroastrian Iranian religion. Yet certain features of his religious reform stand out. He was an ethical prophet of the highest rank, stressing constantly the need for man to act righteously and to speak the truth and abhor the lie. In his teaching the lie was almost personified as the Druj, chief in the kingdom of the demons, to which he relegated many of the earlier Indo-Iranian deities. His god was Ahura Mazdā, who, it seems likely, was a creation in name and attributes of Zoroaster. Though in a certain sense technically monotheism, early Zoroastrianism viewed the world in strongly dualistic terms, for Ahura Mazdā and the "Lie" were deeply involved in a struggle for the soul of man. Zoroaster, as might be expected, attempted to reform earlier Iranian religious practices as well as beliefs. He first rejected and then perhaps allowed the practice of the haoma cult in a modified form; he clearly condemned the practice of animal sacrifice; and he elevated to central importance in the ritual a reverence for fire. Fire worship, however, is a misnomer since the Zoroastrians have never worshipped fire but rather have revered it as the symbol *par excellence* of truth.

The crucial question is: were the Achaemenids Zoroastrians or at least followers of the prophet in the terms in which they understood his message? Possibly Cyrus the Great was, probably Darius I was, and almost certainly Xerxes and his successors were. Such a simple answer to the question is possible, however, only if we understand that Zoroastrianism as a religion had already undergone considerable development and modification since Zoroaster's lifetime, influenced by beliefs and practices and by the religions of those people of the Near East with whom the expanding Iranians had intimate contact.

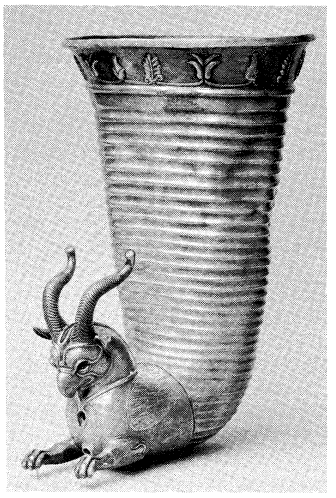
The god of the Achaemenid kings was the great Ahura Mazdā, from whom they understood they had received their empire and with whose aid they accomplished all deeds. Xerxes and his successors mention other deities by name, but Ahura Mazdā remains supreme. Darius I names only Ahura Mazdā in his inscriptions. More significant, however, is Darius' tone, which is entirely compatible with the moral tone of Zoroaster and, in some instances, even compatible with details of Zoroaster's theology. During the reigns of Darius and Xerxes the archaeological record reveals that religious rituals were in force that were also compatible with an evolved and evolving Zoroastrianism. The haoma cult was practiced at Persepolis, but animal sacrifice is not attested. More important, fire clearly played a central role in Achaemenid religion.

There may have been religious overtones in the quarrel between Cambyses and Darius on the one hand and the false Bardiya, a Magian or Median priest, on the other. Certainly there were religious as well as political motivations behind Xerxes' suppression of the Daeva worshippers and the destruction of their temple. It is possible that there was some conflict among the royal Achaemenids, who were followers of one form of Zoroastrianism, the supporters of a different version of Zoroastrianism as practiced by other Iranians, believers in older forms of Iranian religion, and foreign religions, which in the light of the Prophet's teachings were reprehensible. Compromises and syncretism, however, probably could not be prevented. Though the Zoroastrian calendar was adopted as the official calendar of the empire in the reign of Artaxerxes I, by the time of Artaxerxes II, the ancient Iranian god Mithra and the goddess Anahita had been accepted in the royal religion alongside Ahura Mazdā.

Thus, in a sense, the Achaemenid kings were Zoroastrians, but Zoroastrianism itself was probably no longer exactly the religion Zoroaster had attempted to establish. What the religion of the people beyond court circles may have been is almost impossible to say. One suspects that a variety of ancient Iranian cults and beliefs were prevalent. The Magi, the traditional priests of the Medes, may have wielded more influence in the countryside than they did at court, and popular beliefs and practices may have been more deeply influenced by contact with other peoples and other religions. Later classical Zoroastrianism, as known in the Sāsānian period, was an amalgam of such popular cults, of the religion of the Achaemenid court, and of the teachings of the Prophet in their purer form (see also ZOROASTRIANISM AND PARSIISM). (T.C.Y., Jr.)

Art. Achaemenid art, like Achaemenid religion, was a blend of many elements. For Cyrus and his successors the access of temporal power and authority called for expression in the embellishment of their cities and the creation of an appropriately magnificent setting for their everyday lives. For an unsophisticated and hitherto nomadic people, this was an unprecedented requirement. Inspiration was at first sought in the existing formulas of Mesopotamia, Elam, Urartu, and Media, but from the beginning the matrix of the new style was the deep-seated artistic aptitude of the Iranians themselves. The marvels of Achaemenian art were conceived with native—and extraordinary—ingenuity. This continued to be the case even after the large-scale importation of Greek craftsmen in the time of Darius the Great (reigned 522–486 BC).

Architecture. The earliest stages in the evolution of Achaemenian architecture are to be seen in the rather scanty remains of Cyrus' capital city at Pasargadae, north of



Griffon rhyton (drinking vessel), silver, from the Oxus Treasure, Achaemenid period (559–330 BC). In the British Museum.

By courtesy of the trustees of the British Museum

Persepolis. The layout retained the character of an nomadic encampment: widely separated buildings—including gatehouse, residential palace, and audience hall—standing in a vast park, surrounded by a wall 13 feet (four metres) thick. The audience hall provides the earliest example of a formula in design that was to become a criterion of the Achaemenian architecture: a columned hall with corner towers and external colonnades, called by the Persians an *apadana.* Other features are the Tomb of Cyrus, a gabled stone building on a stepped plinth, and a Zoroastrian fire temple (Zendan), a tower-like structure the plan of which recalls the standard Urartian temple. Replicas of the Zendan were built later at Naqsh-e Rostam and elsewhere. Also at Pasargadae, the workmanship of Greek stonemasons was already recognizable, but their full contribution to the new Achaemenian style of architecture is better seen at Persepolis, to which Darius transferred the state capital in 518 BC.

Achaemenian palaces

Achaemenian palaces are built on rock terraces levelled to receive them. The terrace at Persepolis measures approximately 1,600 × 1,000 feet (500 × 300 metres) and is over 43 feet (13 metres) high, with traces of a high enclosure wall of mud brick and around its outer edge and a magnificent stairway approach at its northeast corner. The palace buildings themselves, started by Darius in 518 BC and completed during the following half century by Xerxes and Artaxerxes I, are rather closely grouped together, with little consideration for the overall composition. They are decorated, for the most part externally, with portal sculptures and long ranges of relief carving. Apart from the ornamental facades of the stairways by which the buildings are approached, sculpture is confined to the decoration of doors or windows. These stone-built features together with a few internal columns are all that have survived.

In the absence of the mud-brick walls themselves, it is difficult to gain a true impression of the terrace's original appearance. Most characteristic of these Persian palaces is the proliferation of columns and the tendency to plan them around a square central chamber. In the main gatehouse, with its guardian bulls and bull-men, the square appears as an independent unit. Facing it at a higher level is the largest building of all, the great Apadana (hall) of Darius. It is 272 feet (83 metres) square and said to have accommodated 10,000 people. The four corner towers presumably contained guardrooms and stairs. The sculptured stairway by which it was reached bears the famous relief of the tribute bearers. Next comes the Throne Hall or Hall of a Hundred Columns. It has a portico on the north side with 16 pillars and guardian bulls built into the tower walls at either end. Seven sculptured windows in the north wall are balanced by corresponding niches elsewhere, and the reveals, or jambs, of the doors are also decorated with reliefs. Again approached by an ornamental stairway, a "tripylon" unit between these main buildings leads to others only tentatively identified. The plan of the building, called the Harlem by archaeologists, is to some extent self-explanatory. The character of the Treasury is indicated by security precautions in its planning. In this building the columns were of wood, heavily plastered and painted in bright colours. Elsewhere, columns are fluted in the Greek manner, while the more elaborate capitals and bases have a floral treatment that, like much other Achaemenian ornament, is half Greek and half Egyptian. The most notable form of capital and one peculiar to Achaemenian design is the "double-figure" impost (an upper addition to the capital), taking the form of paired bulls, bull-men, or dragons. Some of these features reappear in the contemporary palace at Susa. Also from this source are figured panels of molded and glazed brick (now in the Louvre), reminiscent of Nebuchadrezzar's Babylon.

(S.H.F.L.)

The palace at Susa

In describing, with justifiable pride, the construction of his palace at Susa, Darius says,

> The cedar timber—a mountain by name Lebanon—from there it was brought . . . the *yakā*-timber was brought from Gandara and from Carmania. The gold was brought from Sardis and from Bactria . . . the precious stone lapis-lazuli and carnelian . . . was brought from Sogdiana. The . . . turquois from Chorasmia . . . The silver and ebony . . . from Egypt . . . the ornamentation from Ionia . . . the ivory . . . from Ethiopia and from Sind and from Arachosia . . . The stone-cutters who wrought the stone, those were Ionians and Sardians. The goldsmiths . . . were Medes and Egyptians. The men who wrought the wood, those were Sardians and Egyptians. The men who wrought the baked brick, those were Babylonians. The men who adorned the wall, those were Medes and Egyptians.

This was an imperial art on a scale the world had not seen before. Materials and artists were drawn from all the lands ruled by the Great King, and thus tastes, styles, and motifs became mixed together in an eclectic art and architecture that in itself mirrored the empire and the Persians' understanding, of how that empire ought to function. Yet the whole was entirely Persian. Just as the Achaemenids were tolerant in matters of local government and custom, as long as Persians controlled the general policy and administration of the empire, so also were they tolerant in art so long as the finished and total effect was Persian. At Pasargadae, the capital of Cyrus the Great and Cambyses in Fārs, the Persian homeland, and at Persepolis, the neighbouring city founded by Darius the Great and used by all of his successors, one can trace to a foreign origin almost all of the several details in the construction and embellishment of the architecture and the sculptured reliefs, but the conception, planning, and overall finished product are distinctly Persian and could not have been created by any of the foreign groups who supplied the king of kings with artistic talent. So also with the small arts, at which the Persians excelled: fine metal tableware, jewelry, seal cutting, weaponry and its decoration, and pottery. It has been suggested that the Persians called on the subject peoples for artists because they were themselves crude barbarians with little taste and needed quickly to create an imperial art to match their sudden rise to political power. Yet excavations at sites from the protohistoric period show this not to have been the case. Cyrus may have been the leader of Persian tribes not yet so sophisticated nor so civilized as the Babylonians or Egyptians, but when he chose to build Pasargadae, he had a long artistic tradition behind him that was probably already distinctly Iranian and that was in many ways the equal of any. Two examples suffice: the tradition of the columned hall in architecture and fine gold work. The former can now be seen as belonging to an architectural tradition on the Iranian Plateau that extended back through the Median period to at least the beginning of the 1st millennium BC. The rich Achaemenid gold work, which inscriptions suggest may have been a specialty of the Medes, was in the tradition of the delicate metalwork found in Iron Age II times at Hasanlu and still earlier at Marlik. In its carefully proportioned and well-organized ground plan, rich architectural ornament, and magnifi-

cent decorative reliefs, Persepolis, primarily the creation of Darius and Xerxes, is one of the great artistic legacies of the ancient world.

The organization and achievement of the empire. At the centre of the empire sat the king of kings. Around him was gathered a court composed of powerful hereditary landholders, the upper echelons of the army, the harem, religious functionaries, and the bureaucracy that administered the whole. This court lived mainly in Susa but in the hot summer months went to Ecbatana (Hamadān), probably in the spring to Persepolis in Fārs, and perhaps sometimes to Babylon. In a smaller version it travelled with the king when he was away in the provinces.

The satraps

The provinces, or satrapies, were ruled by governors (satraps), technically appointed by the central authority but who often became hereditary subkings, particularly in the later years of the empire. They were surrounded and assisted in their functions by a court modelled on that of the central government and were powerful officials. The great king was nevertheless theoretically able to maintain considerable control in local affairs. He was the last court of appeal in judicial matters. He controlled directly the standing military forces stationed in the provinces, though as time went on, the military and civil authority in the provinces tended to become combined under the satrap. The king was also aided in keeping control in the provinces by the so-called king's eyes, or better, the king's ears, officials from the central government who traveled throughout the empire and who reported directly back to the king on what they learned. The number of satrapies and their boundaries varied greatly from time to time; at the beginning of Darius' reign there were 20 provinces. In general, as time went on, the number of satrapies increased, partly because of the need to reassert control over the satraps by decreasing their power base, partly because the feudal structure that underlay Persian society required rewarding more and more people with a role in government, and partly because the original 20 satrapies were undoubtedly simply too large to permit efficient administration.

The army

The army was a particularly important element within the empire. It, too, developed and changed with time. After Cyrus the Persian tribal levy, based on the responsibility of all male Persians to fight for the king, was replaced by a professional army supplemented by a troop levy from the subject peoples in time of intensive military activity. The elite of the standing army were the 10,000 "immortals" composed of Persians and Medes, 1,000 of whom were the personal guard of the king. The person who controlled this elite, as did Darius on the death of Cambyses, usually controlled all. The troops of the imperial levy fought with the regular army in national units, were armed according to their individual customs, but were usually officered by Persians. Permanent bodies of troops were stationed at strategic points throughout the empire, and, to judge from the garrison at Elephantine in Egypt, these were actually military colonies, firmly settled into the local countryside. Greek mercenaries were used with increasing frequency in later years and many Greeks fought faithfully for Persian silver.

Both the civil and the military administration, as well as public and private trade, were greatly facilitated by the famous royal Achaemenid road system. Communications throughout the empire were better than any previous Near Eastern power had maintained. The famous road from Susa to Sardis in western Asia Minor is the best known of these imperial highways. It was an all-weather road, maintained by the state. Over it ran a governmental postal system based on relay stations with remounts and fresh riders located a day's ride apart. The speed with which a message could travel from the provinces to the king at Susa was remarkable.

On the whole, Persian rule sat lightly on the subject peoples, at least under the early Achaemenids. It was a conscious policy of Cyrus and Darius to permit conquered nations to retain their own religion, customs, their methods of doing business, and even to some extent their forms of government. Cyrus' attitude toward the Babylonians, which led to his being accepted as the rightful successor of Nabonidus, his willingness to permit the Jews to return to Palestine and to their own way of life, and his successors' concern that this promise be honoured; Cambyses' behaviour in Egypt and his acceptance by the Egyptians as founder of a legitimate new Egyptian dynasty; and the policy adopted under Mardonius toward the Ionian cities following their rebellion are all examples of such a policy. Perhaps even too often in the later empire, rebellious peoples, governments, and leaders were forgiven and not suppressed with the thoroughness sometimes characteristic of other regimes. Lapses from this policy, such as Xerxes' violent reaction to rebellion in Babylon, stand out in the record.

Law

Law played an important role in the administration of the empire, and stories of Persian justice abound in the Greek sources. Darius particularly wished to be remembered as the great lawgiver, and law reform was one of the cornerstones in his program for reorganizing the empire. To judge from the Babylonian evidence, two sets of law, possibly administered by two sets of courts, were in force in the provinces. One was the local law undoubtedly based on custom and previous local codifications; the other was the Persian, or imperial, law, based ultimately on the authority of the great king. A new word for law appeared in the Near East in Achaemenid times, the Iranian *dāta,* and was borrowed by the Semitic languages used in the empire. In Babylonian and Aramaic, sources give evidence for Persian judges called by the Iranian word *dāta-bar.* These were probably the judges of the imperial courts. With legal reform came reform and unification of tax structures. The tax structure of the empire was apparently based on the principle that all the conquered lands were the actual property of the king. Thus taxes were rather rents, and the Persians and their land, Fārs, by virtue of not being a conquered people, were always tax-free. Each satrapy was required to pay a fixed yearly amount in gold or silver and each vassal state paid a fixed tribute in kind. Again going on the Babylonian evidence, where in previous times agricultural taxes were levied in fixed amounts regardless of the fluctuating quality of the harvest, under Darius all land was surveyed, an estimate of its yield based on an average of the harvests over several years was from time to time established, and taxes were levied in fixed amounts based on a percentage of that average yield. This was not quite an income tax, since it was not based on a percentage of each year's production, but it was at least a reasonable figure based on a reasonable average production.

Breakdowns often occurred in the Achaemenids' effort to maintain a productive balance between local social structures, customs, laws, and government and the demand of the empire. The failure of the Persians to find such a balance when dealing with what was, for them, that extremely strange system of social and political organization, the Greek polis, or city-state, probably lay at the heart of their never-ending troubles in Ionia as much as did the power and ambitions of mainland Greeks. Yet even the Ionians, at the best of times, often realized the mutual advantages and benefits of the king's peace and a unified western Asia under a tolerant central administration.

The economy

The economy of the empire was very much founded on that king's peace; it was when the peace broke down with ever-increasing frequency during the last century of Achaemenid rule that the economy of the empire went into a decline that undoubtedly contributed significantly to eventual political and military collapse. Wealth in the Achaemenid world was very much founded on land and on agriculture. Land was the principal reward that the king had available for those who gave service or who were in positions of great political or military power in the empire. Under Darius there was a measure of land called a "bow" that was originally a unit considered sufficient to support one bowman, who then paid his duty for the land in military service. At the other end of the scale were enormous family estates, which often increased in size over the years and which were or became hereditary holdings. They were often administered by absentee landlords. Such major landholdings were, as one would expect, usually in the hands of Iranians, but non-Iranians were also able to

amass similar wealth and power, thereby testifying once again to the inherent tolerance with which the empire was administered. The Achaemenids themselves took a positive role in the encouragement of agriculture by investing state funds and effort in irrigation and the improvement of horticulture.

They also invested in and endeavoured to encourage trade, a major source of imperial wealth. The effect of the state-maintained road system on the encouragement of trade has already been mentioned. Equal attention was paid to the development of sea borne trade. State-sponsored voyages of exploration were undertaken in order to search for new markets and new resources. Darius completed a project, begun by the Egyptians, of linking the Nile with the Red Sea by a canal, so that routes across the Arabian Sea and into the Persian Gulf could be used to link the eastern and western ends of his empire. As part of the same program, port development on the Persian Gulf coast was encouraged. An imperial standardization of weights and measures, efforts to encourage the development and use of coinage, and the standardization in the king's name of that coinage were all policies intended to encourage commerce and economic activity within the realm. Banking played a role in the economy. Documents have survived from a family banking business in Babylonia—the house of Murashu and Sons of Nippur—covering the years *c.* 455–403 BC; the firm evidently prospered greatly by lending money and by acting as a middleman in the system of tax collection. Interest rates were high, but borrowers were numerous.

As time went on, there were more and more such borrowers, for the later empire is marked by a general economic decline. The principal cause for this decline was the unsettled political conditions, but other, more indirect causes were unwise government interference in the economy, overtaxation, and the removal of too much hard money from the economy. Gold and silver tended to drain into the treasury of the central government from the provinces, and too little found its way back into the economy. Disastrous inflation was the result. The large sums of money paid to foreign mercenaries and as bribes to foreign governments must have also contributed to an unfavourable balance of payments that in turn stimulated inflation. Such conditions hardly strengthened the empire and must have contributed, in ways that cannot be documented with certainty, to the political unrest that was their own main cause.

The significance of the Achaemenids

Ultimately, the achievement of the Achaemenid Persians was that they ruled with much creative tolerance over an area and a time that, for both the Near East and for Europe, saw the end of the ancient and the beginning of the modern world. In one sense, the ancient Near East died when Cyrus marched into Babylon. Others would argue that its death came when Alexander burned Persepolis. The question remains open. What is clear is that the Achaemenid Empire, the largest anyone had ever yet tried to hold together and one that was not to be surpassed until Rome reached its height, was a profound force in western Asia and in Europe during an important period of ferment and transition in human history. That period witnessed major developments in art, philosophy, literature, historiography, religion, exploration, economics, and science, and those developments provided the direct background for the further changes, along similar lines, that made the Hellenistic period so important in history. Hellenism probably would not have been possible, at least not in the form we know it, if it had had to build directly on the rather more narrow and less ambitious base of the individual civilizations of Bablyon, Egypt, or Greece. In a sense the Achaemenid Persians passed on a concept of empire that, much modified by others, has remained something of a model throughout history of how it is possible for diverse peoples with variant customs, languages, religions, laws, and economic systems to flourish with mutual profit under a central government. In narrower terms, but for the Iranians themselves no less important, the Achaemenid Empire was the beginning of the Iranian nation, one of the pivotal peoples in the modern Near East. (T.C.Y., Jr.)

THE HELLENISTIC AND PARTHIAN PERIODS

Alexander and his successors. Between 336 and 330 BC Alexander completed the conquest of the whole Achaemenid Empire. (For the story of the conquest see ALEXANDER.) Alexander's burning of the royal palace at Persepolis in 330 symbolized the passing of the old order and the introduction of Greek civilization into western Asia. Greek and Macedonian soldiers settled in large numbers in Mesopotamia and Iran. Alexander encouraged intermarriage and fostered Greek culture, but he also retained a large part of the Achaemenid administrative structure and introduced Oriental elements and Greek political institutions.

Alexander left no heir. His death in 323 BC signalled the beginning of a period of prolonged internecine warfare among the Macedonian generals for control of his enormous empire. By the end of the 4th century BC Seleucus I Nicator had consolidated his control over that part of Alexander's territory that had corresponded to the Achaemenid Empire.

Seleucus, who, with his son Antiochus I, assumed supreme power, established a government with two capitals: Antioch on the Orontes River in Syria and Seleucia on the Tigris River in Babylonia. The greatest part of Asia—from the Aegean to the Punjab—belonged to this vast kingdom, and to its diverse and varied populace must be added several allied Greek cities, both in Greece and in Asia Minor.

The nobles and the nomads. As he was finishing the conquest of eastern Iran—and at a moment when his attention was being drawn toward the conquest of India—Alexander was confronted by two human factors that were of the greatest importance for the future of his empire. The first of these was the powerful local aristocracy of this part of the Achaemenid Empire, an aristocracy that held enormous properties and dominated the indigenous population; the second was the nomad population that for centuries had wandered along the northern and northeastern frontiers of Iran.

Alexander and the Iranian nobility

Alexander seems to have admired greatly the barons of eastern Iran; he had taken note of their ardour during the two years of hard and constant fighting in his conquest of northeastern Iran. Realizing how such a force could benefit the future of his empire, Alexander convoked an assembly of Bactrian nobles. He ordered 30,000 young men to be chosen for training in the Macedonian military disciplines. He understood the importance and effectiveness of the Iranian light cavalry armed with the bow, and his army would make use of this training in its march toward the plains of India. And finally, Alexander married Roxana of Sogdiana, daughter of a chief of one of the conquered countries, thereby symbolizing the union of the two races.

But Alexander was not unaware that other measures were needed to ensure his control of these vast territories. He founded many new cities, or refounded some that were already in existence. Many of these were placed strategically along the northern frontiers as protection. Almost half of these new cities were located in the "high (eastern) satrapies." This policy of Alexander was soon abandoned by the Seleucids, whose efforts at city planning were mostly confined to their western possessions.

In contrast with Alexander, the Seleucids were unable to maintain the good rapport with the eastern Iranian nobility that Alexander had believed essential. And this deficiency, a result of the Seleucids' "pro-Macedonian" policies, was one of the principal causes for the progressive decline of the Seleucid Empire.

The second of the human factors was the nomads who inhabited the immense territories beyond the northern frontiers. They fought constantly with the settled populations, but could nevertheless occasionally ally with them in the face of necessity. When Alexander arrived on the banks of the Jaxartes River, it marked the limit of the "civilized" world; beyond stretched the Eurasian wilderness. The Roman historian Quintus Curtius recounts Alexander's meeting with a delegation of Scythians who gave him a warning. They told him,

> Just cross the Tanais [properly the Jaxartes] and you will see how far Scythia stretches. You will never conquer the Scythians. Our poverty makes us quicker than your army, which bears plunder from so many nations. Just when you think we are far away, then will you see us in your camp. We know how to pursue and how to flee with the same swiftness. [One recalls here the famous "Parthian shot."] We seek out those deserts totally devoid of human culture rather than the cities and the rich countryside.

These words sum up what the nomad world represented to an empire that stretched several thousand miles from east to west. The non-nomad population knew the threat only too well. Alexander was not the first to cross swords with the nomads. Cyrus, founder of the Achaemenid Empire, had paid with his life while fighting them; and Darius, believing he could take them from behind through southern Russia, suffered a crushing defeat in his campaign against the Scythians along the shores of the Black Sea.

If the nomads and the eastern Iranian nobility were the two dominant factors in the decline of the Seleucid kingdom, and if the events they provoked were some of the principal causes for the exhaustion and eventual fall of that state, these same causes played a not inconsiderable role in the collapse of Parthian power. This power was undermined by an aristocracy that retained its military power and refused to bend before the royal will or to give up its meddling in the country's politics.

In the meantime, the kingdom's unruly neighbours to the north and the northeast, at the cost of the lives of several Parthian sovereigns, weakened the kingdom and sometimes added a complementary element to the often numerous intrigues of the pretenders to supreme power during the course of the almost half a millennium of the existence of the Parthian kingdom.

The Seleucids. In the struggle for power after Alexander's death, Seleucus I brought under his control the whole eastern part of Alexander's empire. But even before he had consolidated his control over this territory, the eastern provinces on the Indian frontier had begun to revolt.

Loss of eastern provinces

By about 304 BC Seleucus was forced to abandon these to Candragupta Maurya, the founder of the great Indian Maurya Empire. This was a serious loss to the Seleucids, for they lost not only the Indian territory conquered by Alexander but also frontier districts west of the Indus. As recompense for his losses, Seleucus received 500 elephants, which he took back with him to Syria. From this time on, the west was dominant in the Seleucids' politics to the detriment of their eastern possessions. This near disinterest of the Seleucids in these far-off eastern regions must have alienated the Greeks of these communities who had settled far from their homeland; and the thought of taking back their full independence could not have been far from their minds.

Soon afterward, toward 290–280 BC, the two eastern provinces of Margiana and Aria suffered an invasion by nomads. But the invasion was repelled and the nomads pushed back beyond the Jaxartes. Demodamas, general to the first two Seleucid kings, crossed the river and even put up altars to Apollo, ancestor of the dynasty. Alexandria of Margiana and Heraclea of Aria, founded by Alexander, were rebuilt by Antiochus I under the names of Antioch and Achaea, and a wall nearly 100 miles (160 kilometres) long was put up to protect the oasis of Merv against future invasions, the menace of which was never far. Patrocles received a commission to explore the Caspian Sea.

Seleucus I and his successors hoped to Hellenize Asia and held the conviction that the Greeks and Macedonians were a superior race and the bearers of a superior civilization. A network of cities and military colonies was built to assure the stability of a state whose inhabitants would be Asians. The Greek language made deep inroads, especially among the families of those numerous Greeks who married the local women and among those engaged in commerce. But after the 2nd century BC and the slowing of the Greco-Macedonian immigration, the Greek language lost ground and the local element became dominant.

Superficial Hellenization

The people of Iran, particularly those in the upper stratum of society, borrowed nothing from Hellenism but its exterior forms. Even the Iranians who lived in such cities as Seleucia or Susa do not seem to have been deeply affected by Greek ideas.

The movement of Iranian peoples. The victories of Alexander had brought the Greeks to the limits of the known world. But less than a century after Alexander's death there began a great movement back, propelled by stirring of peoples in the Iranian world. In a movement westward, and from the 3rd century BC, the Sarmatians occupied the northern shore of the Black Sea. While driving back their close relatives, the Scythians, they succeeded in "sarmatizing" the Greek cities along its shores. At the end of the 3rd century, there began in Chinese Turkistan a long migration of the Yüeh-chih, an Iranian people who about 130 BC invaded Bactria, putting an end to the Greco-Bactrian kingdom there. (In the 1st century AD they created the Kushan (Kuṣāṇa) Empire, which extended from Afghanistan to the Ganges and from Russian Turkistan to the estuary of the Indus.) Finally, in the mid-3rd century BC, taking a median direction between the other two movements described, the Parni, a nomadic or seminomadic people from Iran, took over the Seleucid satrapy of Parthia and created the Parthian kingdom. Restoring Achaemenid power for nearly half a millennium, the arrival of the Parthian state coincided with the expansion of Rome and played a significant role in the destinies of the world during the three centuries before and two centuries after Christ.

Revolt of the high satrapies. The empire of the Seleucids, like that of the Achaemenids before them, was shaken by revolts of the satraps. The difficult situation in the west and the grave reverses suffered by the royal house accelerated the weakening of the Macedonian kingdom. The loss of its eastern possessions in the 3rd century BC proved fatal to the Seleucid cause. Diodotus, a Greek who found himself at the head of the satrapy of Bactria, led a revolt that brought independence in about 250 BC; at about the same time Arsaces led the Scythian Parni into Parthia and defeated Andragoras, establishing an independent native dynasty.

Parthian independence

Parthia was the first province to detach itself from the Seleucid Empire, just as it had been the first to rise up on the occasion of the accession of Darius the Great. Andragoras, although he did not declare himself king, showed his independence by minting his own coins. At this time Parthia was one of the poorer of the high satrapies. Caught between the mountains and the great central desert, without large agricultural resources, this satrapal independence might seem surprising if it were not for the fact that the main route for the silk trade crossed right through Parthia over a distance more than 100 miles. The tolls the caravans paid must have produced a sizable income.

The defection of Diodotus is still easier to understand. Bactria, a vast country of a "thousand cities," was located at the junction of the routes to China and India, and it was rich in cultivable land. The Greco-Bactrian kingdom founded by Diodotus expanded rapidly, embracing Sogdiana and Aria, and extending southward and southeastward.

Being at some distance from the west, Diodotus and his successors adopted little by little the customs and lifestyles of their subjects. The closer these ties were drawn, the stronger became the loyalty of the Bactrians. It is believed that the separation of Diodotus from the Seleucids might, over the long term, have seemed to the Bactrians and Sogdians as the realization of their national destiny, and they might have looked on these satraps as men acting in their interest. For more than a century (230–130 BC) this kingdom held the frontiers and barred the route to the nomads.

The rise of the Parthians. *Invasion of the Parni.* Arsaces, who was chief of the Parni, a member tribe of the Dahae confederation, must have begun his struggle against the Seleucids from 247 BC, the year from which the Parthians dated their history. This does not necessarily mean that Arsaces was crowned king in 247. Other Iranian dynasties (*e.g.*, the Sāsānids; see below) dated the beginning of their eras from the time when they began to establish their power rather than from the time of coronation of the first monarch of their line.

Daho-Parno-Parthian tribes "chose chiefs for war and princes for peace" from among the closest circle of the princely family. They were famous for their breeding of horses, for their combat cavalry, and for their fine archers. Alexander encountered them during his Bactrian campaign, and the Greek writers who recorded his reign remarked on their agility and effectiveness as horsemen. They were a people who kept the traditions of patriarchal tribal organization. The Parni, with Arsaces at their head, took the province of Parthia after having beaten Andragoras; soon, neighbouring Hyrcania was annexed and the Caspian reached. Arsaces had himself crowned in the city of Asaak, and the tribe took the name of the Parthians, their close relatives, a name that meant "exiled." Their language was closely related to Scythian and Median. The dynasty these people produced never broke its links with the people, and rare was the Arsacid dynastic sovereign who did not turn to his people in time of danger.

Formation of the Parthian state. The two new kingdoms, that of Arsaces I's Parthians and the Greco-Bactrian kingdom of Diodotus, sprang up almost simultaneously and very near each other; there were, however, notable differences between them. The motivating force behind the rebellion in Bactria was an association—or perhaps even a collaboration—between the local nobility (large landholders who dominated the whole indigenous population) and the local Greek community. Both groups were opposed to the Macedonian domination represented by the Seleucid dynasty.

Relations with the nomads

The makeup of the Parthian kingdom seems to have been different. It was essentially built on the relationship of the inhabitants of Parthia to the neighbouring tribes outside the static frontiers, an ethnic mass, half nomad, half settled, that inhabited the north of Iran. The success of Arsaces and his men was based on their strength, their spirit, and the weakness of their enemies. The Greek element present in Parthia does not seem to have played a role similar to that played by their counterparts in Bactria. In fact the Parthians, at least initially, may have been hostile to the local Greek populations. During their war with the Seleucid king Antiochus III (see below), they massacred all the Greek inhabitants of the city of Syrinx in Hyrcania.

Arsaces. Arsaces seems to have enjoyed great fame among the tribes. His name remained linked with the names of the sovereigns of this dynasty, who succeeded each other for the four and a half centuries of the Parthian state. His image regularly appeared on the obverse of Parthian coins right down to the end of the period.

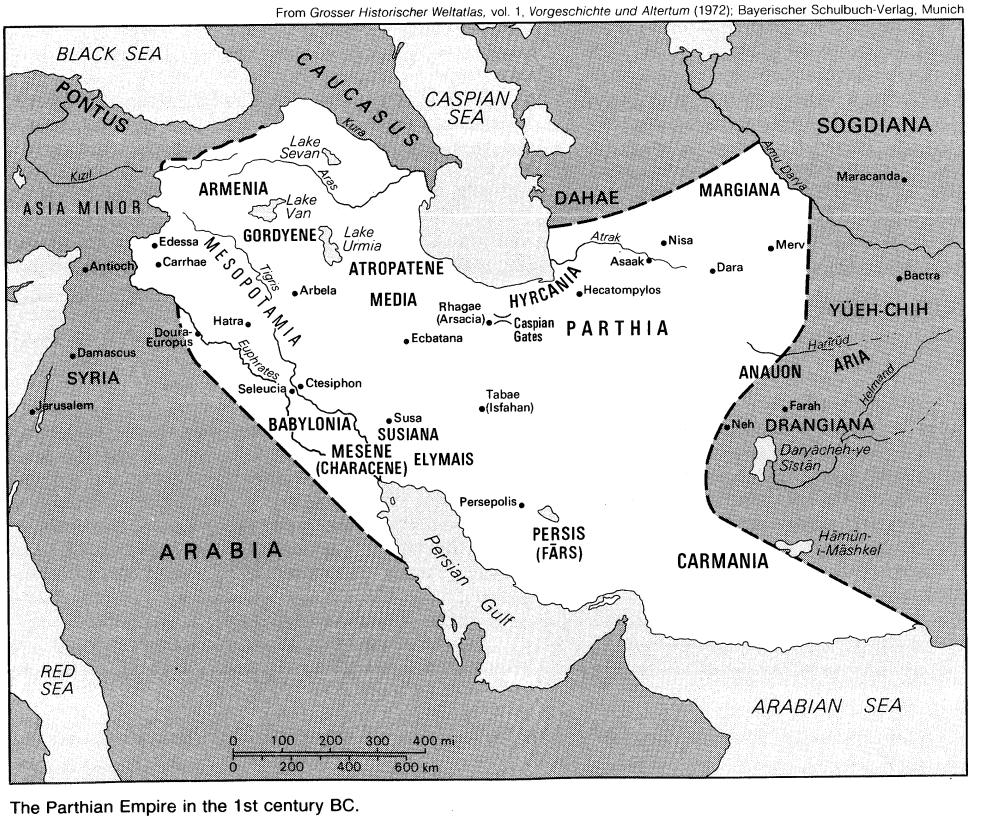
The rupture of the communications link between the Seleucid capitals and the east caused by Arsaces' success placed Diodotus in a difficult situation. He seems to have wanted to collaborate with Seleucus II in a campaign he was preparing against the Parthians. The death of Diodotus (*c.* 234 BC) and the succession of his son, Diodotus II, reversed matters, for the young successor changed his father's policy and joined with Arsaces. It was not until 232 or 231 BC that Seleucus arrived in the east to put down the rebellion. Arsaces, who had remained closely allied with the nomads to the north, sensing his own weakness in the face of Seleucus' army, fled to the home of the Apasiacae, or "Scythians of the Waters." Seleucus II tried to cross the Jaxartes but, having suffered losses at the hands of the nomads, decided to return to Syria after receiving alarming news from the west. He made peace with Arsaces, who recognized his suzerainty.

Founding of new cities

From this time on, Arsaces changed his policy: he no longer acted as a nomad but rather as a chief of state, a worthy successor to the Seleucids, whose example he followed. He had himself crowned; besides Asaak and Dara (an impregnable fortress), he founded such cities as Nisa, where he would be buried. These new cities were usually named after the king or the dynasty. Arsaces seems not to have infringed upon the rights of the Greeks and Macedonians living in these cities, perhaps hoping to win their support. From the beginning, while maintaining the autonomy of the cities, he made use of propaganda to ensure their continuing obedience. He installed his capital at Hecatompylos, on the silk route. His death is dated between 217 and 211 BC.

Artabanus. His successor, Artabanus I (sometimes known as Arsaces II, reigned *c.* 211–191), continued the work of consolidation. Being already solidly established in Parthia and Hyrcania, Artabanus tried to extend his possessions toward Media. But events in the neighbouring Greco-Bactrian kingdom worked against him: Diodotus II lost his throne, accused, it is thought, of treason to Hel-

From *Grosser Historischer Weltatlas*, vol. 1, *Vorgeschichte und Altertum* (1972); Bayerischer Schulbuch-Verlag, Munich



The Parthian Empire in the 1st century BC.

lenism through his alliance with the nomads. His throne had passed to Euthydemus when the Syrian army of Antiochus III arrived in Hyrcania.

The wave of revolts by the eastern satraps, which began a movement away from unity in the state, also affected western Iran. The beginning of the reign of the young Antiochus III the Great (223–187 BC) was marked by the dissidence of Molon and his brother Alexander, satraps of Media and Persis. Antiochus III did not undertake his campaign for recovery of the high satrapies—a project his father had planned and never carried out—until 212 BC.

It is admitted that at this time his kingdom in the east did not stretch farther than Media, Persis, Susiana, and Carmania. His operations against Artabanus I were successful; he took Hecatompylos and crossed the mountains separating that province from Parthia, which he occupied. Artabanus fled and took refuge with the friendly Apasiacae, as had his father, Arsaces I. The conflict between the Seleucids and Parthia, however, was ended by a compromise, just as it had been at the time of the invasion of Seleucus II. A much more important struggle against the Bactrian kingdom of Euthydemus awaited Antiochus III. He preferred to make peace with Artabanus, to whom he accorded the title of king in exchange for recognition of his fealty, and obliged the Parthian to send troops to reinforce the Syrian army. The rear of the Seleucid king was safeguarded, but the two provinces held by Artabanus were definitively lost by the Macedonians.

The period following Antiochus' campaign was marked by a strong resistance by the Bactrian cavalry at the frontier and by the siege of Bactra, for two years the capital (208–207 BC). There, too, the Seleucid king made peace with Euthydemus, who, like Artabanus, kept his title of king. Demetrius, son of Euthydemus, married a daughter of Antiochus the Great, thus preserving his political prestige.

Having acquired war elephants and provisions for his army in Bactria, Antiochus crossed the Hindu Kush into the Kabul Valley, where he concluded a pact with the Indian king Sophagasenos, secured still more elephants, and returned by way of southern Iran. The results of this long campaign were meagre. Antiochus recognized the independence of two kingdoms, that of the Parthians and that of Euthydemus, which had been no more than satrapies. The struggle must have weakened these two, but after receiving legalization of their status, the two states proceeded to the reestablishment of their material and military resources.

Phraates I. Precise information is not available concerning the reign of Priapatius (*c.* 191–176 BC), who succeeded Artabanus and whose name appears in documents found in excavations at Nisa. Under his son Phraates I (*c.* 176–171 BC), the young Parthian kingdom seems to have recuperated sufficiently to have taken up once again its expansionist activities. It attacked Media and was successful in the conquest of the Mardi tribe near the Caspian and set up a defense of the "Caspian Gates," an important strategic point of penetration in Phraates' possessions. Overturning tribal tradition, which reserved the succession to the throne to the eldest son, he designated as a successor—even though he had several sons—his brother Mithradates. His choice was no mistake.

Parthian expansion

The "phil-Hellenistic" period (c. 171 BC–AD 10). The accession of Mithradates I opened a new period in the destinies of the kingdom, which historians call "phil-Hellenistic" and which lasted from the accession of this prince (*c.* 171 BC) until *c.* AD 10. This period was characterized by a strong Hellenistic cultural influence, manifested in the use of the Greek language and in particular in the arts, where, however, national traditions were not completely abandoned.

Mithradates I. From the accession of Mithradates I an expansion of Parthian power in the military, political, and economic domains was manifest. The king began with an attack on the Greco-Bactrian kingdom, which was going through a period of weakness; then he turned against the west and declared himself independent of the Seleucids. To show his complete independence—the first of the Parthian sovereigns to do so—he began issuing coins bearing his likeness, wearing a royal diadem like the Seleucid kings. On the reverse side was a representation of Arsaces, ancestor of the dynasty, seated on an omphalos (hemispheric altar) and holding a bow, in imitation of Seleucid coins that showed Apollo in the same way, as the ancestor of the Seleucids.

The action Antiochus IV Epiphanes undertook against Mithradates ended with the death of Antiochus at Tabae (or Gabae, probably present Isfahan). His death brought a widespread dislocation of the Macedonian kingdom, which crumbled into several smaller states. Toward 160 BC most of the "high satrapies" and the eastern satrapies were apparently lost to the Seleucids. The power that could unite them into one could only be that of the Parthians, who, under Mithradates I, began the assault. In 155 BC Media was occupied, which opened the route to Mesopotamia. In 148–147 BC Mithradates reached Ecbatana, where he moved his capital. Rhagae was "refounded" and given the dynastic name of Arsacia. In 141 BC Mithradates took Seleucia on the Tigris and was recognized king of Babylonia. His forces conquered Susiana and Elymais, either at this time or after 139. In the same year (141), he was obliged to leave Hyrcania for his eastern possessions, which were evidently being menaced by hostile movements of the nomads. There he spent the remainder of his reign.

Demetrius II, probably acquainted with Mithradates' difficulties in the east, undertook an effort to retrieve the situation and recover Mesopotamia, but after a few successes he suffered defeat and was taken prisoner (139 BC). Sent to Hyrcania, he was married there to a daughter of Mithradates, who, by this union, became related to the house of Seleucus. The army of Demetrius included Greco-Bactrian and Elymaian troops—which is understandable—as well as men from Persis, or Persians, who by their cooperation with the Macedonians seem to indicate their opposition to the expansionism of the Parthians, whom they considered foreigners and conquerors. Persia under the Parthians was an empire, but not yet a nation.

Persian resistance to Parthia

Phraates II. Like his father, Mithradates I, Phraates was to remain for some time in the eastern provinces. He was also to endure a last Macedonian attempt to break the Parthian advance. Antiochus VII Sidetes, brother of Demetrius II, who had been taken prisoner, assembled a powerful army, which once more included men of Persis and Elymais. The strength in numbers and the wealth of this army made an impression on contemporaries, who reported that even the simple soldiers wore shoes cobbled with gold. Phraates was beaten in several battles, but time worked on his side. With the arrival of winter, Antiochus quartered his troops in several localities in Media. The local population, exasperated by the undisciplined Syrian soldiery, rose up in revolt. Antiochus was killed and his son taken prisoner (129 BC). Thanks to the loyalty of the Medians, whose sentiments contrasted with those of the Persians, Phraates was victorious. The year 129 BC was a turning point in the history of the eastern Mediterranean: Greco-Macedonian domination received a decisive blow, which it would survive for only 46 years.

The route to great acquisitions in the west seemed to open before Phraates, if the nomads did not stop him. Weakened in his struggle against Antiochus VII, he called upon the Śaka nomads to the north of his frontiers for aid, promising them payment. The reinforcements arrived too late to be of use; he sent them back, which provoked them to revolt and pillage the countryside. The Greek prisoners drafted by Phraates into his army participated in the pillage, and Phraates lost his life fighting them. The same fate was reserved for his successor and uncle, Artabanus II (*c.* 128–124/123). The Śakas were pushed back with some difficulty toward Drangiana, to which they gave their name Seistan (Sakastan). Another branch of the vast nomad movement crossed the Oxus and put an end to the Greco-Bactrian kingdom, on the ruins of which the powerful kingdom of the Kushāns was to be built.

The second stage of the phil-Hellenistic period extends from the first quarter of the 2nd century until about 30 BC and embraces a period when Parthia reached the apogee of its power and worldwide territorial expansion.

Mithradates II. The reign of this prince constitutes the most glorious chapter of Parthian history. It put an end to the ambitions of Artabanus' son Himerus, left by his father as governor of Mesopotamia, and brought back Hyspaosines, king of Characene, who had extended his possessions too far toward the north, into submission. In the east, the Śakas were on the move—soon an independent state would be formed there that would push toward eastern Iran and India; in the 1st century BC, two dynasties, the Indo-Scythian and the Indo-Parthian, whose members would remain closely linked to the Arsacid dynasty, were to reign in this region. They would disappear after having been absorbed by the kingdom of the Kushāns.

The eastern frontiers of Mithradates II incorporated Margiana and Aria. Once order was restored in the east, the King turned toward the west: he placed Tigranes I the Great on the throne of Armenia, and, extending his hegemony over this kingdom and over eastern Asia Minor, he organized pressure on the last Seleucids; among them, a pretender, Demetrius III, was taken prisoner. A meeting with Rome, which had already formed a "Province of Asia" in Asia Minor, became inevitable and took place in 92 BC on the Euphrates between the Roman general Sulla and the Parthian ambassador Orobaze. Mithradates II wisely refused to agree to follow in the Roman path and preferred to retain his neutrality in the struggle between Rome and Mithradates VI of Pontus. Rome in the west and Parthia in the east met as Alexander's successors and, with a common accord, settled the inheritance. The two parties recognized the Euphrates as a common frontier. It seems there was no longer a question either of an alliance or of a signed convention. Upon his return, Orobaze paid with his head for the *lèse majesté* he had committed by accepting a seat lower than Sulla's at their meeting.

Relations with China

For the first time, Parthian power entered into direct contact with the Chinese empire and received an embassy from the Han emperor Wu Ti (140–87 BC), who dispatched to meet them an escort of 20,000 men. The Chinese were particularly interested in the horses raised in Fergana, which they needed to create a cavalry to fight the nomadic Hsiung-nu, or Huns, on their northern border.

At the zenith of his power, Mithradates II took the title of "king of kings"; in the East, as well as in the West, his empire achieved a position of a power and stability previously unknown. He maintained diplomatic relations with the two greatest world powers, Rome and China. Mithradates I, Phraates II, and Mithradates II were the true creators of the Parthian state, winning for it military and economic victories and raising it to a level comparable to that of the Achaemenid Empire. After the death of Mithradates II, a short period of intrigue and rivalry saw the succession, in turn, of Gotarzes, Orodes I, and Sanatruces. The latter came to power late in life and was replaced in 70 BC by his son, Phraates III (70–58/57 BC), under whom took place sustained contacts with Rome.

Wars with Rome. The Roman general Lucullus, in charge of looking after Roman interests in the East (69 BC), hoped to lure Phraates III into an alliance that would help Rome in its struggle against Pontus and Armenia, but the Parthian king, while still maintaining "friendly" relations with Rome, retained his neutrality. An agreement with the Romans renewed the Euphrates line as a frontier. Three years later, in 66 BC, the Roman general Pompey replaced Lucullus and succeeded in concluding a real alliance with Phraates III. This proved, however, to be of short duration, for affairs in Armenia, aggravated by Roman operations on Parthian territory, had brought the two empires to a parting of the ways. To Phraates' protestations Pompey replied with the occupation of Gordyene, a vassal state of the Parthians, and treated with Phraates using the simple title "king." Pompey did not trouble himself over entering into direct relations with the sovereigns of Media and Elymais, vassals of Phraates. The position taken by the Romans toward the king of kings was rather more like that of conquerors than of allies. Pompey's policy became clear: from the Caucasus to the Persian Gulf he hoped to create a wall of states friendly to Rome that would encircle Parthia, in preparation for Roman conquest.

That action fell within the jurisdiction of the Roman triumvir Crassus. As early as 57 BC a conflict with Rome broke out over the case of Mithradates III (58/57–55 BC), who, opposing Orodes II (*c.* 57–37/36 BC), his brother, a parricide like himself of their father, Phraates III, fled to Syria and asked the legate Gabinius for aid and asylum. The Roman Senate forbade Gabinius to involve himself in the dispute over the succession to the Parthian throne. Three years later the tension between the two powers was settled in bloody fashion, and the rupture consummated in 53 BC. Without provocation, the army of Crassus—the only one of the triumvirs with Caesar without military glory (Caeser was conqueror of Gaul, and Pompey conqueror of the Near East)—crossed the Euphrates. Orodes II protested and invoked the treaty of friendship in vain. Crassus refused to reply until he arrived at Seleucia on the Tigris. It was a brutal breaking of all the agreements concluded in 69 and 66 BC.

The Battle of Carrhae

The battle near Carrhae (53 BC), led by Surenas with his light and heavy cavalry, cost Rome seven legions and the lives of Crassus and his son. Through Surenas' brilliant victory the routes to Iran and India were closed to Rome, and its ambitions in the Orient were so weakened that the Euphrates became not only a political but also a spiritual frontier; no effort at latinization was possible any longer. A united Greco-Iranian Asia front protected against the Romanization of Iranianized Hellenism and destroyed the myth of Roman invincibility.

The insignia of the Roman legions fell into Parthian hands, and 10,000 Roman prisoners were sent into captivity in Margiana. The victory over Crassus had great repercussions among the peoples of the East. It shook the Roman position in Asia Minor, Syria, and Palestine, while it restored the Parthians' confidence in their power and in their ability to resist Rome, promising them a dominant position among the peoples of the East. According to the Greek writer Plutarch, the severed head of Crassus was brought to Orodes like a hunting trophy while he was attending a presentation of Euripides' play *The Bacchae.*

The Parthian counterthrust in 52–50 BC under the command of Prince Pacorus (Pakores) was not crowned with success. The Arsacid army did not know how to organize long campaigns or how to lay siege to fortified cities. But soon, civil war in Rome reinforced the position of the Parthians, and Pompey, after being defeated by Caesar, thought of taking refuge among them. It is thought that Orodes, taking advantage of this lull, succeeded in resolving difficulties in the east with the Yüeh-chih, or even perhaps the Kushāns. In 48 BC, with Pompey dead, Caesar was absolute master of the Roman world and was preparing to revenge Crassus' defeat when he was assassinated in 44 BC. The duty of following through on Caesar's project fell to Mark Antony. Pacorus, anticipating Antony, crossed into Syria after having concluded an agreement with Quintus Labienus, a Roman commander on the side of Caesar's assassins who had gone over to the Parthians. The successes of the two armies were startling: Labienus took all of Asia Minor, Pacorus all of Syria and Palestine. For nearly two years all the western provinces of the Achaemenids remained in Parthian hands. In Rome it was rumoured that the Parthians were planning to invade Italy itself. But the successes of the Arsacid armies were as ephemeral as they were remarkable. Disagreement between the two generals weakened their effect. In 39 BC Labienus was conquered and slain. Asia Minor was recovered by the Romans, and the following year the same fate struck Pacorus and his conquests.

The defeat of Pacorus and Labienus

Under Orodes II the Parthians had reached the zenith of their power: in the west, the Arsacids had for a short time reestablished the empire of the Achaemenids almost in its entirety. Their successes in the east seem to have been equally important. Their capital was moved to Ctesiphon, where the military camp was transformed into a great metropolis, facing Seleucia across the Tigris. At Nisa the city was expanded, the royal palaces made larger, and the royal hypogea were enriched with precious pieces of fine Greco-Iranian art.

In 37 BC Orodes was assassinated by his son Phraates IV, who also did away with his brothers and his eldest son. In

36 BC Mark Antony began to carry out the revenge Caesar had planned. He brought his army to Armenia, through which he planned to enter Media and attack Parthia from the north. But cold weather and Phraates' cavalry combined to force Antony to abandon the fight and return to Syria. In 34 BC he launched another campaign and again suffered heavy losses, and his power struggle with Octavian forced him to abandon his plans for war against the Parthians.

Toward 30 BC Tiridates II, a pretender to the throne of Parthia supported by Rome, obliged Phraates IV to leave Mesopotamia and take refuge with his eastern neighbours, the Scythians, who restored him to power. Driven out, Tiridates took refuge at Rome. He returned again in 28–27 BC, after which Phraates was able to definitively reestablish his power at the same time Octavian was inaugurating the imperial period of Roman history.

Settlement with Rome. The new stage in the phil-Hellenistic period began toward 31 BC, when, after his victory over Antony, Octavian (now Augustus) was sole master in Rome. Before that, however, he had already proposed to Phraates an alliance and a treaty ending the war. The Battle of Carrhae and Antony's defeat raised Parthia to a major power in the eyes of Rome. A renewed "friendship" would permit the return of Roman prisoners and insignia. Augustus put pressure on Phraates through the pretender Tiridates and even tried military intervention. In the end a pact was signed in 20 BC that allowed the return of Roman prisoners and the insignia of the conquered legions.

Recognition of the Euphrates frontier

A new stage began in relations between the two states, marked by the conclusion of a real peace that recognized the Euphrates as a frontier between them. Phraates IV was dealt with as the sovereign of a great nation. Rome renounced its ambitions in the east, and Augustus inaugurated a policy of respect. The two states could do nothing but profit from the agreement, for a defeat would have been fatal to either power, and a victory hazardous. The caravan route to India and China was opened. Augustus received ambassadors from the many eastern peoples, including the Indo-Scythians and the Sarmatians. The only country in the east where Rome remained active was Armenia.

All obstacles, however, were not necessarily eliminated. There remained the question of Armenia, which, if controlled by Rome, would be a channel for penetration into Parthia from the north, but if controlled by Parthia would offer an outlet on the Black Sea, over which Rome asserted its authority. The rivalry of the two powers over this country would remain for centuries a stumbling block to peace.

Toward 10 or 9 BC Phraates sent his four sons and grandsons to Rome, a gesture that must be understood as one of confidence in a "friendly" power but also as a guarantee that his throne would pass to his son by Musa, an Italian slave girl given him by Augustus. This son, Phraates V, would assassinate his father with his mother's help and occupy the throne from 2 BC to AD 4 after having married his mother.

The end of this "phil-Hellenistic" period is marked by the clash of the ruling class with foreign influences that had penetrated life in Parthian society. These influences came from Rome and were often introduced by princes of the Arsacid house returning from stays abroad. The short reign of Orodes III (AD 4–6/7) was followed by that of Vonones (7/8–11), son of Phraates IV, who because of his Roman habits was driven out by the Parthian nobility, whose role at this time became dominant in internal politics and dynastic questions. Vonones' fall brought a change in the destinies of the country.

The "anti-Hellenistic" period (AD 12–162). The new and very important period in Parthian history, often called "anti-Hellenistic," embraces a century and a half, from AD 12 to 162. It is characterized by an expansion of the Parthian national culture and an opposition to all things foreign. The weakness of the reigning dynasty opened wide avenues to the nobility to involve themselves in the official existence of the state. They chose the sovereign whose reign opened the first stage in this new period.

Artabanus III. The king chosen by the barons to replace Vonones was Artabanus III (12–38). They were certainly mistaken in believing they would find in him an easy instrument to manipulate. Artabanus was the son of a viceroy of Hyrcania and was only Arsacid on his mother's side. Under his rule Parthia entered a brilliant but troubled era, one completely dominated by the personality of this violently anti-Roman sovereign, who was eager to drive Rome out of Asia. After an abortive attempt to place his son on the throne of Armenia, Artabanus avoided precipitating matters with Rome and dedicated himself to internal reforms, among which centralization occupied the place of first importance.

Rise of the nobles

The humbling of the great nobles, an enterprise in which he was sustained by the lesser nobles, became necessary. He had to reduce the hereditary privileges the barons had carved out for themselves. It was also necessary to reorganize the states that made up the kingdom. He put princes of his family on the various thrones of these states: Mesene, Persis, Elymais, Atropatene, all little states that were governed by men loyal to the throne. But it proved impossible for him to put down a revolt in the eastern possessions, where the Indo-Parthian king Gondophares declared himself independent (*c.* 19) and took the title "king of kings."

It is thought that the position taken toward the city-states, about which precise information is lacking, was the reason for the seven-year-long revolt of Seleucia on the Tigris. The fighting took place there between the Greek and Hellenized elements and the Semites, who demanded their right to participate in the autonomy of the city and who supported pretenders against Artabanus III.

A new attempt to place a son on the throne in Armenia angered Rome, which, with the aid of the nobility, sent for Tiridates III, a pretender the barons had crowned at Ctesiphon, obliging Artabanus III to take refuge with the Dahae, who helped him win back his throne. In 37 a meeting with a representative of Rome on a bridge in the middle of the Euphrates allowed an agreement to be reached that maintained the status quo in Armenia and recognized the Parthian sovereignty with the river as the frontier.

The strong personality of Artabanus III did not seek to impose his kingdom as a world power, but he did not hesitate to make plans to regain the western province, the former Achaemenid possessions.

Dissolution of the Parthian state. The period from 51 to 122 shows a slow dissolution of the Parthian state and its decomposition into several small countries. This was an inevitable result of the weakness of the central power.

Pliny's description of Parthia

In the 1st century AD the Parthian Empire, according to the Roman historian Pliny, was composed of 18 kingdoms, 11 in the north and seven in the south, some governed by Arsacid princes and others by local dynasties. In 58 Hyrcania became independent. In the realm of external affairs an effort was made to maintain good relations with Rome, especially because of the new kingdom of the Kushāns, which was causing concern on the eastern frontiers. It might be for this reason that in 87 Parthia sent an embassy to neighbouring China to the east of the Kushāns. Internally, the nationalistic upsurge became more accentuated.

After the short reign of Vonones II (51), the throne passed to Vologases I (51–77/78), an ardent anti-Roman. One of his brothers, Vonones, was made king of Media. Vologases I wanted his second brother, Tiridates, to be king of Armenia, which put him in the position of a break with Rome, which opposed him militarily. Upon orders from Nero, Corbulo undertook operations broken off by the exchange of ambassadors. An agreement was finally reached: in 66 Tiridates left for Rome with his whole family surrounded by a retinue of princes and 3,000 Parthian nobles. He received from Nero the crown of Armenia, and an end to hostilities was announced by the closing of the doors to the Temple of Janus.

Nationalist sentiment, which under Artabanus III had found expression in the invention of a genealogical table proving the Achaemenid descent of the Arsacid house, manifested itself under Vologases I by the compilation of the Avesta, the holy book of the Iranians, and by the

issuance of coins on which, for the first time, Pahlavi characters were added to the Greek legend.

In 78 Pacorus II came to the throne, to be replaced in 79 by the ephemeral Artabanus IV (80/81), who was then replaced permanently by Pacorus II. During his reign the country showed signs of a profound decomposition. The barons refused to obey the crown. In the provinces the army and finances were in the hands of the nobility. Aristocrats occupied the highest positions, and these positions became hereditary. Plots with Rome were hatched, and the nobility felt itself the equal of the dynasty, ready to revolt in defense of their privileges. Externally, the dynasty was unable to count on Rome, which constantly plotted in support of new pretenders. In 109/110 Pacorus II was replaced by Osroes, his brother or brother-in-law.

Invasion of Trajan

In 114 the emperor Trajan invaded Armenia. In vain did the King put his crown at Trajan's feet—he was defeated by the Roman soldiery. With Armenia occupied, the Emperor descended with his army into Mesopotamia. All Babylonia was taken and Ctesiphon, the capital, fell into the hands of the Romans, who carried off a daughter of Osroes and the golden throne of the Parthian kings. Victorious, Trajan went as far as the Persian Gulf. Iranian reaction was not late in coming. Faced with the gravity of the Roman offensive, all the princes of the royal house, formerly divided by internal strife, united against the invader. At Ctesiphon Trajan crowned a new vassal king, but revolt was in the wind and attempts to disunite the Parthian chiefs failed. The Romans suffered losses, and, after a reverse on the walls of Hatra, Trajan abandoned the campaign and died on his way home. Trajan's successor, Hadrian (117–138), abandoned all pretensions to Armenia, Mesopotamia, and Assyria.

Hadrian's desire for peace seems to have been sincere. He sent back Osroes' daughter, promised to return the golden throne, and did not try to profit from the long power struggle between Osroes and Vologases II. He even invited Osroes to come to Rome.

Peace with Rome. This was a time of 40 years of peace with Rome. The status quo it maintained with its western neighbour seems even to have been a necessity for the Parthian kingdom, the expansion of the kingdom of the Kushāns on the eastern frontiers having reached the peak of its power under King Kaniṣka (Kanishka). Accurate information about the relations between the Kushāns and the Parthians is not available, but this long peace sought with Rome suggests that certain precautions were necessary for the kingdom of Iran.

The end of the Parthian Empire (AD 162–226). The 40 years' peace was succeeded by almost uninterrupted hostilities with Rome, with varied success, Iran remaining more vulnerable because of the exposed position of its capital.

Apparent stability in the 2nd century

The reign of Vologases II (105/106–?147) and especially that of Vologases III (148–192), the latter not having to dispute the throne with a pretender, could by their length be a sign of a certain stability the country might have experienced. But underneath the apparent calm the intrigues continued, with Rome receiving embassies from the Hyrcanians, the Bactrians, and doubtless from the Kushāns.

A new clash with Rome came in 161, this time upon the initiative of Vologases III, who considered himself strong enough to attack. He occupied Armenia, crossed the Euphrates, and invaded Syria, which for two centuries had not seen Parthian cavalry. And although the country had been Roman since the time of Pompey, the Syrian population, which included Jews driven from Palestine by the Romans, received the Iranians as liberators. The situation became so serious that Lucius Verus, co-emperor with Marcus Aurelius, was dispatched to the east with strong reinforcements taken from the fronts on the Danube and Rhine. The Romans retook Armenia (163) and succeeded in a campaign similar to Trajan's: Doura-Europus was taken and remained Roman until its destruction by the Sāsānids; Seleucia on the Tigris, despite the welcome it reserved for the Romans, was sacked; and in 164 or 165 for the second time Ctesiphon fell into the hands of Romans, who razed the royal palace.

But once more success was not continuous. The Roman army had come from Armenia and had crossed through Azerbaijan, known even today as a country where plague is endemic. Contaminated, the Roman army was sorely tried by disease and obliged to retreat, but not definitively. Lucius Verus, repeating his campaigns in Armenia and northern Mesopotamia, inflicted heavy losses on the Parthians.

The tensions between the two states did not diminish when Vologases IV (191–207) supported a pretender (Niger) against the emperor Septimius Severus. The latter became emperor in 193 and began operations that permitted him to occupy first northern and then southern Mesopotamia and, for the third time in a century, Ctesiphon.

The Parthians in their retreat adopted a scorched-earth policy. As under Trajan, the starving Roman army went back up the Tigris, failed in its attempt to take Hatra, and left the country.

Vologases V, son of the previous king, succeeded him (207–222), and his throne was contested from 213 by another prince, Artabanus V (213–224), who was able to maintain himself thanks to the support of the kingdom of Media. A new invasion of Mesopotamia took place under Caracalla, the *casus belli* being the refusal of Artabanus V to give Caracalla his daughter in marriage. The young Roman emperor dreamed of rebuilding Alexander's empire but succeeded only in the pillage of Media and the destruction at Arbela of the hypogea of the Arsacid kings, whose bones he scattered.

Invasion of Caracalla

The Parthian reply was harsh. Artabanus V avenged himself by invading the Roman provinces and destroying several cities. Rome sued for peace. Artabanus' conditions were too hard and were refused. Hostilities were taken up again and turned in favour of the Parthians, who obtained such a success that the emperor Macrinus paid 200,000,000 sesterces to make peace.

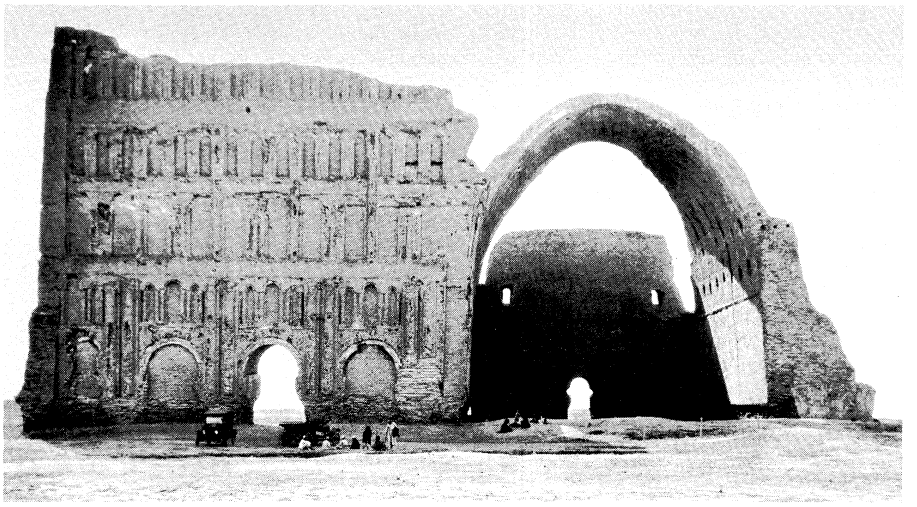
Since 208 Pāpak (Bābak), a lesser prince of Persis, had been preparing a revolt, which his son Ardashir finally declared openly. A battle took place between him and Artabanus V in 224; the Parthian was killed, and the throne of Iran passed into the hands of the Sāsānids, a new national dynasty, originally from Fārs, cradle of the Achaemenids.

The Iran of the Parthians, a region strategically crucial for international commerce, maintained open roads, created cities, and encouraged exchanges that were the lifeblood of this great nation stretching from the portals of China and India to the Roman Empire. Tolerant in religion, it was Parthia that contributed to the dissemination of Buddhism to China, where a Parthian prince spread the word of Buddha near the middle of the 2nd century AD. For nearly half a millennium Parthia pursued its great ambition to recover the western provinces of the Achaemenids. An empire of the middle, between Rome on the west and the Kushāns on the east, undermined by internal weaknesses, Parthia finally succumbed, leaving its great dreams to its successors, the Sāsānids. (R.Gh.)

Architecture and art. In Iran and Mesopotamia, this long era of Parthian occupation is poorly represented by newly built towns, but there are a few notable examples. One was Ctesiphon, originally a Parthian military camp facing Seleucia, the older capital city on the other side of the Tigris River. Another was Hatra, a fortress city in the Jazīrah Desert between the Tigris and Euphrates rivers; and a third was Gūr-Fīrūzābād, south of Shiraz. All of these show the approximately circular plan of military tradition. The palaces, sometimes built of good ashlar masonry, and even the private houses are distinguished by a feature later characteristic of Islāmic architecture: the *īwān,* or three-sided hall, the fourth side of which is replaced by an open archway. At Hatra and in a Parthian palace at Ashur, the *īwān*s multiplied in number, and the adjoining facades are decorated with engaged columns, singly or in tiers, corresponding approximately to the orders of classical architecture.

Parthian palaces

Hatra is by far the best preserved and the most informative example of a Parthian city. It has an inner and an outer city wall, the former built of ashlar masonry, and a temenos (temple enclosure) enclosing the principal sacred buildings. The great central group, built during the



The facade of the Palace of Shāpūr I at Ctesiphon, Iraq, Sāsānian period, 3rd century AD.
By courtesy of the Oriental Institute, the University of Chicago

2nd and 1st centuries BC and still standing with many of its vaulted chambers intact, includes the Shamash temple, a characteristic Iranian sanctuary with a square central chamber, ambulatory, and outer wall incorporating a staircase to the roof. Comparable buildings have been found as far afield as Taxila in the Indus Valley. Excavations by Iraqi archaeologists have revealed temples dedicated to Al-Lāt and Shahiro, who, with Shamash, comprise a Hatran trinity.

In the 1st and 2nd centuries AD, Hatra was ruled by a dynasty of Arab princes whose written language was Aramaic. Excavations have yielded a great wealth of sculptures dating from this period, including portrait statues of the princes and their families. All of this material, when published, will throw new light on the whole subject of Parthian art, which depended hitherto on such isolated survivals as the magnificent bronze statue from Shami, in the old Elamite homeland. Rock reliefs at Tang-e Sarwak, Bīsitūn, and elsewhere reveal one salient characteristic of Parthian design—that is, the preference for "frontality" or, as one scholar describes it, "the tendency for the figures to ignore one another and to the face the spectator with an unblinking stare." Frontality is also characteristic of the Romano-Syrian sculptures of Palmyra and later passed into the fabric of Byzantine art.

Petra and Palmyra. Two cities, strategically placed in Jordan and eastern Syria, respectively, were at times associated with Parthian history and have left monuments suggesting a compromise between Roman and Near Eastern art. Both were caravan cities and each in turn acquired wealth and importance from its position at the junction of arterial trade routes. Petra, in biblical Edom, was ruled by Nabataean kings from late in the 2nd century BC until 106 AD, when it became a Roman colony. Its subsequent decline was due in part to the temporary ascendancy of Palmyra, on the western fringe of the Syrian desert, whose remarkable queen Zenobia ruled a miniature empire until defeated by the Romans in AD 272.

The significance of Petra in the history of ancient Near Eastern art is difficult to assess, since its art is in large part restriced to the design of rock-cut tombs or temples. Imitation building facades carved on cliff faces surround the tomb and temple entrances. Their most distinctively oriental characteristic is the baroque effect obtained by ingenious contortions of classical Roman formulas. The dramatic beauty of their natural setting and the chromatic peculiarity of the stone from which they were hewn are assets that have enhanced their reputation. Stripped of these, their interest is primarily academic.

Architecture of Palmyra

The architecture of Palmyra is more conventionally Roman, but it is freely associated with sculpture; and many bas-relief carvings decorate the tombs of rich merchants and other notables. Parthian influence is to be seen in their style, particularly in the frontality of the human figures.

(S.H.F.L.)

THE SĀSĀNIAN PERIOD

Foundation of the empire. *Rise of Ardashīr I.* At the beginning of the 3rd century AD, the Arsacid Empire had been in existence for some 400 years. Its strength had been undermined, however, by repeated Roman invasions, and the empire became once more divided, this time between Vologases V (207–222), who seems to have ruled at Ctesiphon, on the left bank of the middle Tigris in what is now Iraq, and Artabanus V (*c.* AD 213–224), who was in control of Iran and whose authority at Susa, in southwestern Iran, is attested by an inscription of AD 215.

It was against Artabanus V that a challenger arose in Persis. Ardashīr I, son of Pāpak and a descendant of Sāsān, was the ruler of one of the several small states into which Persia had gradually been divided. His father had taken possession of the city and district of Istakhr (Estakhr), which had replaced the old residence city of Persepolis, a mass of ruins after its destruction by Alexander the Great in 330 BC. Pāpak was succeeded by his eldest son, who was soon killed in an accident, and in AD 208 Ardashīr replaced his brother. He first built for himself a stronghold at Gūr, called after its founder Ardashīr-Khwarrah (Ardashīr's Glory), now Fīrūzābād, southeast of Shīrāz in Fārs. He subdued the neighbouring rulers and disposed, in the process, of his own remaining brothers. His seizure of such areas as Kermān, Isfahan, Elymais, and Characene (Mesene), to the east, north, and west of Fārs, respectively, led to war with Artabanus, his suzerain. The conflict between the two rivals lasted several years, during which time the Parthian forces were defeated in three battles. In the last of these, the battle in the plain of Hormizdagān (AD 224), Artabanus was killed.

Numismatic evidence concerning Ardashīr

There is evidence to support the assumption that Ardashīr's rise to power suffered several setbacks. Thus, Vologases V struck coins at Seleucia, on the Tigris, as late as AD 228/229 (the Seleucid year 539). Another Parthian prince, Artavasdes, a son of Artabanus V, known from coins on which he is portrayed with the distinguishing feature of a forked beard, seems to have exercised practical independence even after AD 228. Numismatic evidence further reflects the stages of Ardashīr's struggle for undisputed leadership. He appears on his coins with four different types of crown: as king of Fārs, as claimant to the throne before the battle at Hormizdagān (AD 224), and as emperor with two distinctly different crowns. It has been suggested that this evidence points to two separate coronation ceremonies of Ardashīr as sovereign ruler, the second, perhaps, indicating that he may have lost the throne temporarily.

According to aṭ-Ṭabarī, the Arabic historian (9th–10th centuries), Ardashīr, after having secured his position as a ruler in western Iran, embarked on an extensive military campaign in the east (AD 227) and conquered Seistan (Sakastan), Gorgān (Hyrcania), Merv

Table 1: Sāsānian Dates Established on Direct Ancient Evidence

event	reign years	Seleucid Era	Christian Era
Accession of Artabanus (Ardawān) V	1		212/213
Inscription of Khwasak at Susa names Artabanus "king of kings"			215
Birth of Mani	5	527	216/217
Artabanus V overthrown and killed by Ardashīr			224
Official first year of Ardashīr		538*	226/227
Last coin of Vologases V, minted at Seleucia		539	228/229
Mani, in 13th year, receives divine revelation		539	228/229
Official first year of Shāpūr I		553	241/242

*Syrian reckoning.

(Margiana), Balkh (Bactria), and Khwārezm (Chorasmia). The inference that this campaign resulted in the defeat of the powerful Kushān Empire is supported by the further statement of aṭ-Ṭabarī that the king of the Kushāns was among the eastern sovereigns, such as the rulers of Tūrān (Quzdar, south of modern Quetta) and of Mokrān (Makran), whose surrender was received by Ardashīr. These military and political successes were further extended by Ardashīr by his taking possession of the palace at Ctesiphon, by his assuming the title "king of kings of the Iranians," and by his refounding and rebuilding of the city of Seleucia, located on the Tigris River, under the new name of Weh-Ardashīr, the Good Deed of Ardashīr.

Sāsānian chronology

The chronology of events in the early Sāsānian period was calculated by the German orientalist T. Nöldeke in 1879, and his system of dating is still generally accepted. The discovery of fresh evidence in manuscript materials dealing with the life of Mani, a religious leader whose activities fall in the early Sāsānian period, led to a reassessment of Nöldeke's calculations by W.B. Henning, by which the principal events are dated about two years earlier. Another alternative was proposed by S.H. Taqizadeh, who preferred a sequence by which the same events are placed about six months later than the dates established by Nöldeke. Since the dating systems employed by the Sāsānians themselves were based on the regnal years of the individual kings, whose exact coronation dates are often subject to dispute, several details remain uncertain, and their definite solution has not been possible. A firmer basis of calculation is obtained when the ancient sources quote dates in terms of the Seleucid era, either according to the computation that prevailed in Babylonia, which started from 311 BC, or after the Syrian reckoning, beginning in 312 BC. Tables 2 and 3 give the dates as they can be established on direct numismatic or literary evidence in the differing chronological systems of Nöldeke, Henning, and Taqizadeh.

Wars of Shāpūr I. Shortly before his death, probably because of failing health, Ardashīr abdicated the throne in favour of his chosen heir, his son Shāpūr I. The latter assumed the responsibilities of government but delayed his coronation until after his father's death. Coins thus exist showing Ardashīr together with his son as heir apparent and Shāpūr alone wearing the eagle cap, indicating the exercise of royal rule before his coronation—besides the normal series of Shāpūr crowned as king.

Table 2: Chronological Systems of Nöldeke, Henning, and Taqizadeh

event	Nöldeke	Henning	Taqizadeh
Ardashīr's first year begins		Sept. 27, 223	Sept. 26, 226
Ardashīr's actual accession	Sept. 26, 266	April 28, 244	April 6, 227
Shāpūr's first year begins		Sept. 23, 239	Sept. 22, 241
Shāpūr's actual accession	Sept. 22, 241		
Shāpūr's coronation		April 12, 240	April 9, 243
Shāpūr's death		May 270	April 273
Accession of Hormizd I	Sept. 14, 272		
Hormizd I's death		June 271	April 274
Accession of Bahrām I	Sept. 14, 273		
Death of Mani (about age 60)		March 2, 274	Feb. 26, 277
Death of Bahrām I		Sept. 274	July 277
Accession of Bahrām II	Sept. 13, 276		

Shāpūr's defeat of the Romans

Shortly after his accession, Shāpūr was faced with an invasion of Persia by the emperor Gordian III (238–244):

> . . . the emperor Gordian levied in all of the Roman empire an army of Goths and Germans and marched against Asūristān [Iraq], the empire of Iran and us. On the border of Asūristān, at Massice [Misikhe on the Euphrates], a great battle took place. The emperor Gordian was killed and we destroyed the Roman army. The Romans proclaimed Philip [the Arab; 244–249] emperor. The emperor Philip came to terms, and as ransom for their lives he gave us 500,000 dinars and became our tributary. For that reason, we renamed Massice Fīrūz-Shāpūr ["victorious (is) Shāpūr"].

Several years later, in AD 256 (or AD 252), another confrontation between the Persians and Romans occurred:

> We attacked the Roman empire and we destroyed an army of 60,000 men at Barbalissus [in Syria]. Syria and its surrounding areas we burned, devastated and plundered. In this one campaign we captured of the Roman empire 37 cities. . . ,

including Antioch, the capital of Syria, itself. A third encounter took place when the emperor Valerian (253–260) came to the rescue of the city of Edessa, in Syria, which was besieged by the Persian army:

> He (Valerian) had with him (troops from) Germania, Rhaetia . . . [follow the names of some 29 Roman provinces], a force of 70,000 men. Beyond Carrhae and Edessa there was a great battle between the emperor Valerian and us. We made the emperor Valerian prisoner with our own hands; and the commanders of that army, the praefectus praetorii, senators and officers, we made them all prisoner, and we transported them to Persia. We burned, devastated and plundered Cilicia and Cappadocia . . . [follow the names of 36 cities].

The source for these quotations is Shāpūr's own account of the events. It was unknown until 1938, when expeditions of the Oriental Institute in Chicago discovered a long inscription on the walls of an Achaemenid building known as the Kaʿbe-ye Zardusht (Kaʿba of Zarathushtra). The text is in three languages, Sāsānian Pahlavi (Middle Persian), Parthian, and Greek. Besides the narrative of the military operations, the inscription provides a description of the Persian Empire of the time and an inventory of Zoroastrian religious foundations established by Shāpūr I to commemorate his victorious wars. These foundations were fire temples dedicated to the "soul" (memory) of the founder himself, of members of the royal family, and of prominent officials who had served under Shāpūr and his predecessor. The list of the officials who are specified by the positions they held throw light on the administrative organization of the empire.

Organization of the empire. In contrast to his father, who claimed to be "king of kings of Iran" (*shāhanshāh ērān*), Shāpūr I assumed the title "king of kings of Iran and non-Iran" (*shāhanshāh ērān ud anērān*). This formula was retained by his successors as the regular designation of the Sāsānian emperors. The hereditary local dynasties, which under the Arsacids had ruled many of the most important provinces, were to a large extent abolished. Instead, such areas as Maishān (Mesene), in western Iran, and Sakastan (Sīstān), in eastern Iran, were now ruled by members of the Sāsānian family, who were appointed by the sovereign with the title of *shāh* (king). Among such provincial governors, precedence was often given to the heir to the throne, who was placed in control of large territories, such as the former Kushān Empire (Kūshānshahr) and Armenia, with the title "great king" (*wuzurg shāh*). This arrangement lasted until the early 4th century AD, and such emperors as Shāpūr I and Hormizd II are known to have first held the title *kūshānshāh* as governors of the areas of Bactria, Sogdiana, and Gandhāra. Next in the hierarchy came a few remaining hereditary vassals, such as the kings of Iberia (now Georgia) in the Caucasus, and the chief nobles of the empire, among whom the Warāz, Sūrēn, and Karēn families retained their prominent position from Parthian times. Next in line were the satraps, whose importance had diminished and who were no more than the administrators of larger cities or court officials.

The list of provinces given in the inscription of Kaʿbe-yi Zardusht defines the extent of the empire under Shāp-



The Sāsānian Empire at the time of Shāpūr I.

ūr I, in clockwise geographical enumeration: (1) Persis (Fārs); (2) Parthia; (3) Susiana (Khūzestān); (4) Maishān (Mesene); (5) Asūristān (Iraq); (6) Adiabene; (7) Arabistān (northern Mesopotamia); (8) Atropatene (Āzārbāījān); (9) Armenia; (10) Iberia (Georgia); (11) Machelonia; (12) Albania (eastern Caucasus); (13) Balāsagān up to the Caucasus Mountains and the Gate of Albania (also known as Gate of the Alans, now the Darreh Āhū Pass in the central Caucasus); (14) Patishkhwagar (all of the Elburz Mountains); (15) Media; (16) Hyrcania (Gorgān); (17) Margiana (Merv); (18) Aria; (19) Abarshahr; (20) Carmania (Kermān); (21) Sakastan (Sīstān); (22) Tūrān; (23) Mokrān (Makran); (24) Paratān (Paradene); (25) India (probably restricted to the Indus Delta area); (26) Kūshānshahr, until as far as Peshāwar and until Kashgar and (borders of) Sogdiana and Tashkent; and (27), on the further side of the sea, Mazun (Oman). This empire, considerably more extensive than that controlled by the Arsacid dynasty, was governed by members of the royal family and by appointed officials directly responsible to the throne. The greater degree of centralization thus attained by the Sāsānian government partly explains its increased military effectiveness in comparison with the Arsacid administration. Tight organization of the numerous central and provincial officials, whose ranks in the bureaucratic structure on different levels were strictly defined, also contributed toward general administrative efficiency.

The extent of Shāpūr I's empire

Another trend that developed in the Sāsānian period, although it had already made itself felt under the Arsacids, was a strict principle of dynastic legitimacy. For a usurper not of the royal blood to come to the throne was an extremely rare occurrence, though it was in fact accomplished by Bahrām VI Chūbīn. Loyalty was given, however, to the whole royal house, rather as it was in the later Ottoman Empire. The person of the individual ruler was a matter of comparatively lesser importance, and one member of the dynasty could readily be removed and replaced by another. In accordance with this principle of legitimacy, Persian tradition carried the Sāsānian line back to the Achaemenids and, ultimately, to the kings of the legendary period.

Religious developments. *Zoroastrianism.* The ancestors of Ardashīr had played a leading role in the rites of the fire temple at Istakhr, known as Ādur-Anāhīd, the Anāhīd Fire. With the new dynasty having these priestly antecedents, it seems only natural that there would have been important developments in the Zoroastrian religion during the Sāsānian period. In fact, the evolution of Zoroastrianism as an organized religion into something resembling its modern form can be regarded as having begun in this period. Under the Parthians, local Magi had no doubt continued to perform the traditional ceremonies associated with the old Iranian deities, the fire cult, the creed preached by Zoroaster, with its emphasis on the worship of Ahura Mazdā, and even the cults of cosmopolitan deities that were introduced in the Hellenistic period and later.

Under the Sāsānians, stress came to be placed on the fire cult and the worship of Ahura Mazdā. Strong mutual relationships, furthermore, were developed between religion and the state, and an ecclesiastical organization was set up in which every local district of any importance had its own *mobed* ("priest"; originally *magupat,* "chief priest"). At their head stood the *mobedān mobed* ("priest of priests"), who, in addition to his purely religious jurisdiction, appears, especially in later times, to have had a more or less decisive voice in the choice of a successor to the throne and in other matters of state. There is also some evidence that the *mobeds*, by virtue of their proficiency in reading and writing in general and in the interpretation of the sacred scriptures in particular, performed the offices of registrars and scribes in semireligious or nonreligious matters, after the fashion of the Christian clergy in medieval Europe. This situation in turn makes it likely that the priestly library buildings not only contained the sacred texts, charters, and other church records but also served as repositories of local archives, title deeds, and other documents of a legal nature. The

Religion and the state

Table 3: Sāsānian Kings*

name	reign years	name	reign years
Defeat of Artabanus V (Ardavan)	226	Kavadh (Qobād) I (first reign)	488–496
Ardashīr I	224–241	Jāmāsb	496–499
Shāpūr I	241–272	Kavadh (second reign)	499–531
Hormizd I	272–273	Khosrow I	531–579
Bahrām I	273–276	Hormizd IV	579–590
Bahrām II	276–293	Khosrow II Parvīz (first reign)	590
Bahrām III	293	Bahrām VI	590–591
Narses	293–302	Khosrow II Parvīz (second reign)	591–628
Hormizd II	302–309	Bestām (rebel in Media)	591–596
Shāpūr II	309–379	Kavadh (Qobād) II Shīrūye (Siroes)	627–628
Ardashīr II	379–383	Ardashīr III	628–630
Shāpūr III	383–388	Shāhrbarāz	630
Bahrām IV	388–399	Pūrāndokht	629–631
Yazdegerd I	399–420	Hormizd V	631–632
Bahrām V Gūr	420–438	Khosrow III	632–633
Yazdegerd II	438–457	Yazdegerd III	633–651
Hormizd III	457–459		
Fīrūz	457–484		
Balāsh	484–488		

*Based mainly on T. Nöldeke's chronology.

building known as Kaʿbe-ye Zardusht and referred to as a *bun-khānag* ("foundation house") may well have served this very purpose.

In the matter of religious practice, the theology of the Sāsānians appears to have developed from that previously current in their home province of Persis. There, extraneous religious influences were limited. The opposition between the good spirit of light and the demons—between Ormizd (Ahura Mazdā) and Ahriman (Angra Mainyu)—remained the essential dogma. All the other gods and angels were restricted to the role of subordinate servants of Ormizd, whose highest manifestation on earth was not so much the sun or the sun god Mihr (Mithra) but rather the holy fire guarded and attended by his priests. At the same time the names of such deities as Wahrām (Verethraghna), Mihr, and Anāhīd (Anahita) were still associated with the names of fire temples or classes of fires. Divine names were also used to designate the 30 days of each month and of the 12 30-day months of the year, plus five epact days called *gahānīg,* to align the lunar with the solar year.

Ritual practice

All the prescriptions of purity were scrupulously observed. The elaborate ritual still maintained in modern times by the Parsi for the purification and custody of the sacred fire was no doubt observed under the Sāsānians. The officiating priest was girt with a sword and carried in his hand the *barsman* (*barsom*), or bundle of sacred grass. His mouth was covered to prevent the sacred fire from being polluted by his breath. The practice of animal sacrifice, abhorred by the modern followers of Zoroaster, is attested for the Sāsānian period at least as late as the reign of Yazdegerd I (399–420). On the days of the important festivals, such as *Nōgrūz* (*Nowrūz*), the first day of the vernal equinox, and on the day of *Mihragan* (the 16th day of the seventh month), the sacred fire was displayed to the faithful (*wehden*) at nightfall from some vantage point. Under the Sāsānians, the injunction not to pollute the earth by contact with corpses but to expose the dead on mountain tops to vultures and dogs was strictly observed. Ahura Mazdā preserved his character as a national god, who bestowed victory and world dominion on his worshippers. In rock-relief sculptures he appears on horseback as a god of war.

Theology was further developed, and an attempt was made to modify the old dualistic concept by considering both Ormizd and Ahriman as emanations of an original principle of infinite time (Zurvān). This doctrine enjoyed a certain degree of official recognition in early Sāsānian times. In the reign of Khosrow I (531–579), however, the "sect of the Zurvanites" was declared to be heretical. The chief trend of Sāsānian religion, apart from the process of institutionalization, was toward the elaboration of ritual and of the doctrine of purity. A complete and detailed system of casuistry was developed, which dealt with all things allowed and forbidden and with the forms of pollution and the expiation of each. One of the consequences of this development was the increasing emphasis placed on orthodoxy and rigorous obedience to priestly injunctions. Nonorthodox and heretical cults and forbidden manners and customs came to be regarded as a pollution of the land and a serious offense to the true God. It was the duty of the believer to combat and destroy the unbelievers and the heretics. In short, the tolerance of the Achaemenids and the indifference of the Arsacids were gradually replaced by religious intolerance and persecution.

Despite his priestly family origin, Ardashīr himself seems not to have been the person responsible for initiating these new directions in religious affairs. It was once believed that the institutionalization of the Zoroastrian church and the codification of its scriptures and beliefs was the work of a high priest named Tansar, a contemporary of Ardashīr I, of whose activities an account is preserved in the *Letter of Tansar,* contained in the history of Ṭabaristān by the Persian writer Ibn Isfandiyār (12th–13th centuries). New inscriptional evidence, however, rather suggests that, if Tansar was, in fact, a historical personage, his role in religious matters was overshadowed by Kartēr (Karder).

Role of the priest Kartēr

The latter, a *herbed* ("teacher priest") and *mobed* ("priest") already prominent under Shāpūr I, appears during the reigns of Bahrām I (273–276) and Bahrām II (276–293) as the dominant figure in the Zoroastrian church. As stated in the Kaʿbe-ye Zardusht inscription of Kartēr, he claims credit for the suppression of non-Zoroastrian religious communities in Iran ("and Jews, Buddhists, Brahmins, 'Nazoreans,' Christians . . . were struck upon"), for the imposition of orthodoxy and discipline on the priesthood ("the heretics [*ahlomog*] . . . who in the Magus estate did not attend to the Mazdean religion and the services to the gods with discrimination, I struck them with punishment and I castigated them"), and for the establishment of royal foundations for the maintenance of priests and of sacred fires.

Christianity. The reference in the Kartēr inscription to two sects of Christians continues the indications from Syriac sources that Christianity had by this time, the second half of the 3rd century, gained a firm footing in the lands of the Tigris and the Euphrates, where it was strongest among the Aramaic-speaking communities. Ultimately, Christian missionary effort came to expand over the whole of Iran and even beyond. As long as the Roman Empire remained pagan, the Christian communities of Iran lived undisturbed by persecution, while the Christians themselves showed outspoken hostility toward such heterodox sects as the Manichaeans and the Gnostic followers of Marcion and Bardesanes, who existed side by side with them. Once the emperor Constantine the Great (306–337) made Christianity the official religion of the Roman world, the Iranian Christians were, on the one hand, drawn to feel a certain sympathy for their foreign coreligionists, while, on the other, political significance came to be attached by the Sāsānian rulers to these religious connections with an often hostile foreign power. After 339 the Christians of Iran were subjected to severe persecutions at the hands of Shāpūr II and his successors. Substantial Christian communities survived, nonetheless, in parts of Iran long after the end of the Sāsānian dynasty.

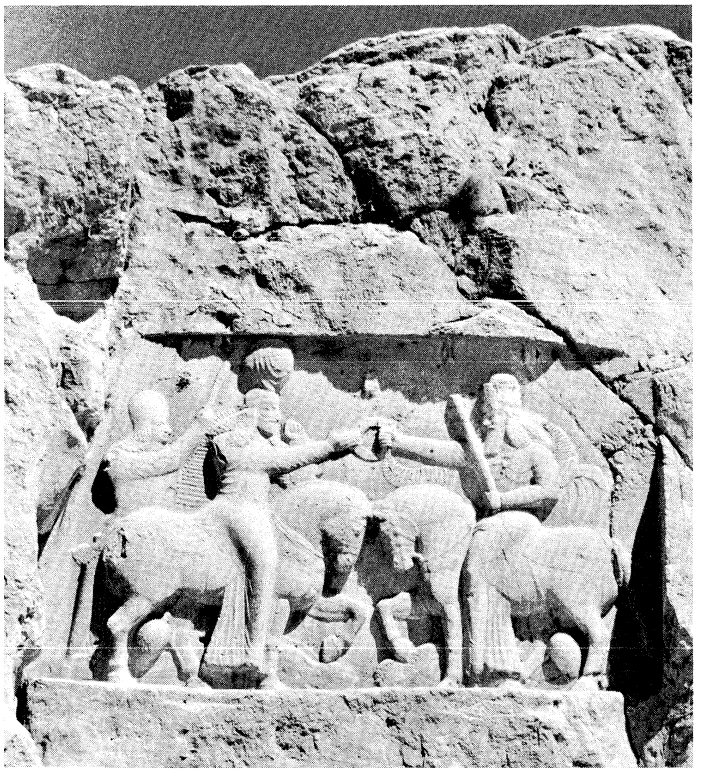
Manichaeism. During the reign of Shāpūr I a new religious leader and movement made their appearance. Mani (between 216? and 274?) was the offspring of a Parthian family resident in Babylonia ("a thankful disciple I am, risen from Babel's land") but himself a speaker of Aramaic. In the early 20th century, knowledge of his teachings was greatly increased by the discovery of many fragments of Manichaean literature in eastern Turkistan. Subsequently, a large part of the *Kephalaia,* a collection of the religious injunctions of Mani, was recovered in a Coptic version, found in Egypt. These texts can now be collated with the versions of Manichaean doctrines as reported by the Church Fathers, including St. Augustine. From this cumulative documentation, to which other sources can be added, it appears, among other things, that Mani's teachings were formulated under the strong influence of Gnostic ideas and philosophy. Mani proclaimed himself to be the last and greatest Apostle of Jesus as well as the paraclete announced in the Gospel of St. John. With the Gnostic interpretation of the Gospel, Mani tried to combine the doctrines of Zoroaster and Jesus in order to create a new religion of a universal character. There is a tradition that he made his first appearance as a teacher on the coronation day of Shāpūr I (April 12, 240, or April 9, 243), but other evidence suggests that Mani was not necessarily in Iran at the time and may have been on a sea journey to India when he started preaching. He later returned and found many followers, among whom were Fīrūz (Pērōz) and Mihrshāh, governor of Maishān (Mesene), both brothers of Shāpūr I. Even the King himself is said to have been impressed and to have granted the prophet several personal interviews. On the last such occasion, Mani presented the King with his first book, the *Shāpuragān* (*Shabuhragan*), a summary of his teachings ("dedicated to Shāpūr") written in the Middle Persian language, further evidence of a degree of royal favour. During Shāpūr's reign the religion of Mani was thus propagated in and beyond Iran. The heir to the throne, Hormizd I, was also favourably disposed toward him. Shāpūr's younger son, Bahrām I, however, yielded to pressure from the priestly establishment, and Mani was executed. After that, Manichaeism was persecuted and destroyed in Iran. Yet it

Suppression of Manichaeism

maintained itself not only in the West, penetrating far into the Roman Empire, but also in the East, in Khorāsān and beyond the boundaries of the Sāsānian Empire. There the seat of its pontiff was at Samarkand, whence it penetrated into Central Asia.

Art and literature. Perhaps the most characteristic and certainly among the most impressive relics of Sāsānian art are the great rock sculptures carved on the limestone cliffs that are found in many parts of the country. The best known groups are at Naqsh-e Rostam and Naqsh-e Rajab, both near Persepolis, and at Bishāpūr, an ancient city a few miles north of Kāzerūn in Fārs. At Fīrūzābād, the ancient Gūr, also in Fārs, are two reliefs of Ardashīr I, one depicting the overthrow of Artabanus V, the other an investiture scene. Not far away, in the valley at Sar Mashhad, a representation of Bahrām II shows that king in the process of slaying two lions. At Dārābgerd, about 180 miles southwest of Shīrāz, Shāpūr I is shown triumphing over three Roman emperors, Gordian III, Philip the Arabian, and Valerian. At Naqsh-e Bahrām, north of Kāzerūn, Bahrām III is depicted enthroned. The same ruler appears at Qaṣr-e Abū Nasr, near Shīrāz, and at Gūyom, not far from there. Sāsānian sculptured reliefs are less numerous outside Fārs, but a Sāsānian equestrian that once existed at Rayy (ancient Rhagae), southeast of Tehrān, was replaced in the 19th century by a representation of Fatḥ

Paul Popper



Sāsānian rock relief showing the investiture in AD 226 of Ardashīr I at Naqsh-e Rostam, Persia (Iran).

ʿAlī Shāh, a member of the then-ruling Qajar dynasty. At Salmās, near Lake Urmia, Ardashīr I is shown on horseback while receiving the surrender of a Parthian personage. There are also later Sāsānian sculptures at Ṭāq-e Bostān, near Bakhtaran, showing Ardashīr II, Shāpūr III, and Khosrow II. In many of these representations the Sāsānian kings can often be identified by their individual crowns.

Architecture

The most ambitious and celebrated architectural achievement of the dynasty is the vast palace at Ctesiphon, built by Khosrow II (590; 591–628), of which a part is still standing. It is known as the Ṭāq Kisrā and is notable for its great barrel vault in baked brick, a typically Sāsānian architectonic device. Many Sāsānian buildings can also be seen in Fārs, where the characteristic construction is of limestone blocks embedded in strong mortar. The most important of these are the palace of Ardashīr I at Fīrūzābād, south of Shīrāz, and a small, well-preserved palace at Sarvestān, southeast of Shīrāz, in which the rooms are roofed with domes and squinches, features often found in Sāsānian architecture. Excavations at Bishāpūr, or Shāhpūr, near Kāzerūn, have revealed some mosaic floors and other features of this important Sāsānian town. Numerous fire temples of the period survive, especially in Fārs; these are square buildings roofed by a dome over four arches. Sāsānian remains of considerable extent also exist at Qaṣr-e Shīrīn, on the road from Baghdad to Tehrān, and at Gondēshāpūr, modern Shāhābād, south of Dezfūl.

Metalwork and coinage

Generally speaking, the Sāsānian era was one of a renascence in Iranian art, which, if not quite on the same level as the Achaemenid achievement, was of no small importance. Metalwork reached a high level of artistry and craftmanship; its most characteristic decorative themes are hunting scenes portraying the Sāsānian kings in action. A gold and enamel drinking vessel (now in the Bibliothèque Nationale in Paris) from the time of Khosrow I (531–579)—known as the "Cup of Solomon" and, according to one tradition, a gift of the caliph Hārūn ar-Rashīd to Charlemagne—is perhaps the most sumptuous specimen of Sāsānian metalworking. The art of gem engraving produced many fine intaglio stamp seals and cameos. The coins invariably bear a Pahlavi (Middle Persian) inscription; on the obverse is the head of the king, wearing his characteristic crown, accompanied by his name and title, on the reverse the fire altar with its guardians and the legend "Fire of Ardashīr, Shāpūr, etc." or, in the later period, an abbreviated mint name and the regnal date.

The acquaintance with Greek language and literature maintained by the Arsacid court had begun to decline during the last century of that dynasty. Greek versions nonetheless accompany the Parthian and Middle Persian texts of the inscriptions of Ardashīr I and Shāpūr I, as in the case of the Kaʿbe-yi Zardusht inscription. Later inscriptions, however, are only in Parthian and Middle Persian, as in the case of the inscription of Narses at Paikuli.

Literature

Most of the comparatively few remains of literature in (Book) Pahlavi—a form of Middle Persian somewhat different from that used in the Sāsānian inscriptions—is of late or post-Sāsānian date in its actual form, if not in content. This is partly due to the fact that the transition from an oral to a written literary tradition took place in the later part of the Sāsānian era. This is true of both religious and secular compositions. A passage in a religious text states that "it is proper to consider the living spoken word more weighty than the written." It should be added that most Sāsānian literary remains are primarily of religious and historical rather than of literary interest. Just as foreign learning appears in religious works, likewise foreign prose works of entertainment came to Persia, where they were translated; among them, in the time of Khosrow I, were Hellenistic romance literature and Indian books of tales, such as *Kalīlag and Dimnag,* based on the Indian *Pañcatantra* or the legends of *Barlaam and Josaphat* (*Balauhar and Budasaf*).

Foreign policy. In foreign policy the problems under the Sāsānian kings remained, as of old, the defense and, when possible, the expansion of the eastern and western frontiers. The successful military campaigns in the eastern areas by Ardashīr I and Shāpūr I, which resulted in the annexation of the western part of the Kushān Empire, have already been mentioned.

Conflicts with Rome. In the west, the old contest for northern Mesopotamia with the fortified cities of Carrhae, Nisibis, and Edessa continued. The Sāsānians were all the more eager to regain and retain control of Armenia because there the Arsacid dynasty still survived and turned for protection to Rome, with which, in consequence, new wars continually broke out. In the reign of Bahrām II (276–293), the emperor Carus (282–283) invaded Mesopotamia without meeting opposition and reached Ctesiphon. His sudden death, however, caused the Roman army to withdraw. Bahrām II had been prevented from meeting the Roman challenge by the rebellion of his brother, the *kūshānshāh* Hormizd, who tried to establish an independent eastern empire. This attempt ended in

failure, however, and Bahrām II appointed his younger son, the future Bahrām III, as viceroy of Sakastan (Sīstān). After Bahrām II had died (293), Narses, the youngest son of Shāpūr II, contested the succession of Bahrām III and won the crown. In memory of his victory, Narses erected a tower at Paikuli, in the mountains west of the upper Diyālā River, which was discovered in 1843 by the English Orientalist Sir Henry Rawlinson. Decorated with busts of Narses, the monument has a long inscription in Parthian and Middle Persian that tells the story of the events. In 296, Narses was forced to conclude a peace treaty with the Romans by which Armenia remained under Roman suzerainty and certain areas in northern Mesopotamia were ceded to Rome. By this treaty, which lasted for 40 years, the Sāsānians withdrew completely from the disputed districts. The Roman Empire had meanwhile become Christian, and the Syro-Christian populations of Mesopotamia and Babylonia began to feel sympathy with Roman policies for religious reasons. Christianity also became predominant in Armenia after King Tiridates adopted the Christian faith in 294. The Sāsānian emperors consequently felt the need to consolidate their Zoroastrianism, and efforts were made to perfect and enforce state orthodoxy. All heresy was proscribed by the state, defection from the official faith was made a capital crime, and persecution of the heterodox, the Christians in particular, began. Competition between Iran and Rome–Byzantium thus took on a religious dimension.

Religious factors in foreign policy

A new war was inevitable. It was begun by Shāpūr II in 337, the year of the death of Constantine the Great. Shāpūr besieged the fortress city of Nisibis three times without success. The emperor Constantius (337–361) conducted the war weakly, but Shāpūr was distracted by the appearance of a new enemy, the nomadic Chionites (Huns), on his eastern frontier. After a long campaign against them (353–358), he returned to Mesopotamia and, with the help of Chionite auxiliaries, captured the city of Amida (modern Diyarbakır) on the upper Tigris, an episode vividly narrated by the Roman historian Ammianus Marcellinus (*c.* 330–400). The emperor Julian the Apostate (361–363) reopened hostilities after the death of Constantius (361) but died after having reached the vicinity of Ctesiphon. His successor, Jovian (363–364), was forced to give up the Roman possessions on the Tigris, including Nisibis, and to abandon Armenia and his Arsacid protégé, Arsaces III, to the Persians. The greater part of Armenia then became a Persian province.

Intermittent conflicts from Yazdegerd I to Khosrow I. After about two decades of disturbed reigns (Ardashīr II, Shāpūr III, Bahrām IV), Yazdegerd I came to the throne. His reign is viewed differently by Christian and Zoroastrian sources. The former praise his clemency; the latter refer to him as "Yazdegerd the Sinful." His initial inclination toward tolerance of the Christian and Jewish religions was met by resistance on the part of the nobility. Because of their attitude and because of the growing fanaticism of the Christians, Yazdegerd was forced to turn to repression. After his death (420), the nobles refused to admit any of Yazdegerd's sons to the throne. But one of them, Bahrām V, had the support of al-Mundhir, Arab prince of al-Ḥīrah (east of the lower Euphrates) and a Sāsānian vassal, and also, apparently, of Mihr-Narseh, chief minister of Yazdegerd's last years, who was retained in office, and Bahrām eventually won the throne. As King Bahrām V, surnamed Gūr (Wild Ass), he became the favourite of Persian popular tradition, which exuberantly celebrates his prowess in hunting and in love. Unsuccessful in war with Byzantium (421–422), Bahrām V made a 100-year peace and granted freedom of worship to the Christians. In the east he did succeed in repelling an invasion by a new wave of Hephthalites. In the following decades, however (second half of the 5th century), Hephthalite attacks continued to harass and weaken the Sāsānians. Fīrūz (457–484) fell in battle against them; his treasures and family were captured, and the country was devastated. His brother Balāsh (484–488), unable to cope with continuing incursions, was deposed and blinded. The crown fell to Kavadh (Qobād) I, son of Fīrūz. While the empire continued to suffer distress, he was dethroned and imprisoned (496), but he escaped to the Hephthalites and was restored (499) with their assistance. The Nestorian doctrine (claiming that divine and human persons remained separate in the incarnate Christ) had by then become dominant among the Christians in Iran and was definitely established as the accepted form of Christianity in the Sāsānian Empire.

Reign of Kavadh I

Kavadh I proved himself a vigorous ruler. He restored peace and order in the land. A campaign against the Romans (502) resulted in the destruction of Amida, but another inroad by the Hephthalites in the east compelled him to ratify a peace treaty with the Byzantines. Toward the close of his reign, in 527, he resumed the war and defeated the Byzantine general Belisarius at Callinicum (531) with the support of al-Mundhir II of al-Ḥīrah. Earlier in his reign he had moved away from the Zoroastrian church and favoured Mazdak, the founder of a new socio-religious movement that had found support among the people. The crown prince, Khosrow, however, was an orthodox Zoroastrian; toward the end of his father's reign, in collaboration with the chief *mobed,* he contrived the condemnation of the Mazdakites, who were destroyed in a great massacre (528). On his father's death, after acceding as Khosrow I (531–579), he concluded peace with the Byzantine emperor Justinian (532). He reestablished Zoroastrian orthodoxy, and, although some persecution of Christian communities occurred during periods of tension with Byzantium, the restoration of peace brought about a considerable amount of religious tolerance.

Khosrow I was one of the most illustrious Sāsānian monarchs. From his time dates a new and more equitable adjustment of the imperial tax system. The levying of land revenue in kind was replaced by a fixed assessment in cash, and these assessments continued in force later under the Arab administration. His reputation as an enlightened and just ruler was high during his lifetime and later became legendary. When Justinian, in 529, closed the philosophy school in Athens, the last Neoplatonists turned to Khosrow in hopes of finding in him the true philosopher-king. Though they were disillusioned by conditions at his court, their gratitude was great when Khosrow obtained for them the right to return. From 540 onward Khosrow had been conducting a long war against Justinian, which, although interrupted by several armistices, lasted until the 50 years' peace of 561. Khosrow also extended his power to the Black Sea and inflicted heavy defeats on the Hephthalites. These military successes were in part the result of several reorganizations of the armed forces and the chain of command that were achieved during Khosrow's long reign.

Conflicts with the Turks and Byzantium. About 560 a new nation, that of the Turks, had emerged in the east. By concluding an alliance with a Turkish leader called Sinjibu (Silzibul), Khosrow was able to inflict a decisive defeat on the Hephthalites, after which event a common frontier between the Turkish and Sāsānian empires was established. Inevitably, this alliance became a source of possible friction, and the Turks sometimes acted as an ally of Byzantium against Iran in a second war (572–579).

Revolts of Bahrām Chūbīn and Bestām

Khosrow bequeathed this war to his son Hormizd IV (579–590), who in spite of repeated negotiations failed to reestablish peace between Byzantium and Iran. Hormizd was unable to display the same authority as his father, and he antagonized the Zoroastrian clergy by failing to take action against the Christians. He finally fell victim to a conspiracy headed by the general Bahrām Chūbīn. Hormizd's son, Khosrow II, was set up against his father and forced to acquiesce in Hormizd's execution. New unrest broke out, in which Bahrām Chūbīn—though not of royal lineage—attempted to secure the throne. Simultaneously another pretender, Prince Bestām, decided to try his luck. Khosrow fled to Byzantium, and the emperor Maurice (582–602) undertook his restoration by military force. Bahrām Chūbīn was routed (591) and fled to and was killed by the Turks, and Khosrow again ascended the throne in Ctesiphon. Bestām held out in Media until 596.

Khosrow II (590; 591–628), surnamed Parvīz (the Victorious), achieved unprecedented splendour and material wealth. The assassination of Maurice (602) impelled him to a war against Byzantium, in the course of which his armies penetrated as far as Chalcedon (opposite Con-

stantinople), ravaged Syria, and captured Antioch (611), Damascus (613), and Jerusalem (614); in 619 Egypt was occupied. The Byzantine Empire was, indeed, at its lowest ebb.

It took the great emperor Heraclius, who was crowned in 610, many years to rebuild the nucleus of a new army. This done, however, he set out in 622 and retaliated vigorously against the Persians. Their armies were defeated everywhere. In 624 Heraclius invaded Atropatene (Azerbaijan) and destroyed the great Zoroastrian fire temple; in 627 he entered the Tigris provinces. Khosrow II attempted no resistance; a revolution followed in which he was defeated and slain by his son Kavadh (Qobād) II (628). When Kavadh died a few months later, anarchy resulted. After a succession of short-time rulers, Yazdegerd III, grandson of Khosrow II, came to the throne in 633.

Triumph of the Arabs. All of these prolonged and exhausting hostilities had drastically reduced the powers of both Byzantium and Iran. The door was open to a newly emerging force that challenged both states and religions—the Arabs. After several encounters, the fate of the Sāsānian Empire was decided in the battle of al-Qādisīyah (636/637), on one of the Euphrates canals, not far from al-Ḥīrah, during which the Sāsānian commander in chief, Rustam, was killed. Ctesiphon with its treasures was at the mercy of the victors. Yazdegerd fled to Media, where his generals tried to organize new resistance. The battle fought at Nehāvand (642), south of Hamadān, put an end to their hopes. Yazdegerd sought refuge in one province after another, until at last, in 651, he was assassinated near Merv.

End of the Sāsānian Empire

With the fall of the empire, the fate of its religion was also sealed. The Muslims officially tolerated the Zoroastrian faith, though persecutions were not unknown. Little by little it vanished from Iran, except for a few surviving adherents who remain to the present day in Yazd and a few other places. Other Zoroastrians emigrated to western India, where they are now chiefly concentrated in Bombay. These Parsis (Persians) have preserved only a relatively small portion of their sacred writings. They still number their years by the era of Yazdegerd III, the last king of their faith and the last Sāsānian sovereign of Iran.

(A.D.H.B./M.J.Dr.)

IRAN FROM 640 TO THE PRESENT

The advent of Islām (640–829). The Arab invasion of Iran made a break with the past that affected not only Iran but all of western Asia and resulted in the assimilation of peoples who shaped and vitalized Muslim culture. The Prophet Muḥammad had made Medina, his adopted city, and Mecca, his birthplace, centres of an Arabian movement that Muslim Arabs developed into a world movement by the conquest of Iranian and Byzantine territories. Neither the Iranian nor Byzantine empires had been unfamiliar to those Arabs who were the former's Lakhmid and the latter's Ghassānid vassals, the two empires' frontier guardians against fellow Arabs who roamed deeper in the Arabian Desert. Also, Meccan and Medinese Arabs had established commercial connection with the Byzantines and Sāsānids. The immunity of Mecca's ancient sanctuary against outlawry and outrage had promoted this city's commercial importance. The Kaʿbah, as the sanctuary was called, was cleansed of idols by Muḥammad, who had himself once been engaged in commerce. He made it the sanctuary of a monotheistic faith whose sacred writings were impregnated with the injunctions and prohibitions needed by a business community for secure and stable trading.

Arab tribalism beyond urban fringes was less easily broken than idols. It was embedded in the desert sparsity that led to warfare and careful counting of a tribe's male offspring. After Mecca and Medina had become Muslim, to secure the routes they depended upon necessitated winning the desert Arabs' allegiance. In the process of doing this, wars over water holes, scanty pastures, men-at-arms, and camels were enlarged into international campaigns of expansion.

The vulnerability of Sāsānid Iran assisted the expansionist process. In AD 623 the Byzantine emperor Heraclius reversed Persian successes over Roman arms, namely, the capture of Jerusalem in 614 and a victory at Chalcedon in 617. His victim, Khosrow Parvīz, died in 628 and left Iran a prey to a succession of puppet rulers who were frequently deposed by a combination of nobles and Zoroastrian clergy. Thus, when Yazdegerd III, Iran's last Sāsānid and Zoroastrian sovereign, came to the throne in 632, the year of Muḥammad's death, he inherited an empire weakened by Byzantine wars and internal dissension.

Sāsānid weakness

The former Arab vassals on the empire's southwestern border realized that their moment had arrived, but their raids into Sāsānid territory were quickly taken up by Muḥammad's caliphs, or deputies, at Medina—Abū Bakr and ʿUmar (632–634; 634–644)—to become a Muslim, Pan-Arab attack on Iran.

An Arab victory at al-Qādisīyah in 636/637 was followed by the sack of the Sāsānid winter capital at Ctesiphon on the Tigris. The Battle of Nahāvand in 642 completed the Sāsānids' vanquishment. Yazdegerd fled to the empire's northeastern outpost, Merv, whose *marzbān*, or march lord, Mahūyeh, was soured by Yazdegerd's imperious and expensive demands. Mahūyeh turned against his emperor and defeated him with the help of Hephthalites from Bādghis. The Hephthalites, an independent border power, had troubled the Sāsānids since at least 590, when they had sided with Bahrām Chūbīn, Khosrow Parvīz's rebel general. A miller near Merv murdered the fugitive Yazdegerd for his purse.

The Sāsānid end was ignominious, but it was not the end of Iran. Rather, it marked a new beginning. Within two centuries, Iranian civilization was revived with a cultural amalgam, with patterns of art and thought, with attitudes and a sophistication that were indebted to its pre-Islāmic Iranian heritage—a heritage changed but also stirred into fresh life by the Arab-Muslim conquest.

Abū Muslim's revolution. Less time was needed before the Arabs' assimilation with Iranians in regions they colonized caused a new Islāmic beginning, Abū Muslim's movement, which began in Khorāsān, in 747. This revolution followed years of conspiracy directed from Medina and across to Khorāsān along the trade route that linked the Far East to Merv and thence with the West. Along the route merchants with contacts in the Mesopotamian Arab garrison cities of Kūfah, Wāsiṭ, and Basra acted as intermediaries. Iranians who by converting to Islām had become clients, or *mawālī*, of Arab patrons played direct and indirect parts in the revolutionary movement. The movement also involved Arabs who had partnered Khorāsānian and Transoxanian Iranians in ventures in the great east–west trade and intercity trade of northeastern Iran. The revolution was, nevertheless, primarily an Arabo-Islāmic movement, aimed to supplant a militaristic, tyrannical central government whose fiscal problems made it avid for revenue by one more sympathetic to the needs of the merchants of eastern Islām. Abū Muslim, a revolutionary of unknown origin, was able in Merv to exploit the discontent of the merchant classes as well as that of the Arab and Iranian settlers. The object of attack was the Umayyad government in Damascus.

Background of the revolution

When Muḥammad died in 632, his newly established community in Medina and Mecca needed a guiding counsellor, an imam, to lead them in prayers and an *amīr al-muʾminīn*, "Commander of the Faithful," to ensure proper application of the Prophet's divinely inspired precepts. As the Prophet, Muḥammad could never be entirely succeeded, but it was accepted that men who had sufficient dignity and who had known him could fulfil the functions, as his caliphs (*khulafāʾ*), or deputies, and imams. After Abū Bakr and ʿUmar, ʿUthmān was chosen for this role.

By ʿUthmān's time, factionalism was growing among Arabs, partly the result of the jealousies and rivalries consequent upon acquiring new territories, and partly the result of the competition between first arrivals in them and those who followed. There was also uncertainty over the most desirable kind of imamate. One faction, the Shīʿah, supported ʿAlī, the Prophet's cousin and husband of his favourite daughter, Fāṭimah, for the caliphate, since he had been an intimate of Muḥammad and seemed more capable than the other candidates of expressing Muḥammad's wisdom and virtue as the people's judge. The

desire for such a successor points to disenchantment with ʿUthmān's attempt to strengthen the central government and impose demands on the colonies. His murder in 656 left his Umayyad relatives poised to avenge it, while ʿAlī was raised to the caliphate. A group of his supporters, the Khārijites, desired more freedom than ʿAlī was willing to grant, with a return to the simplest interpretation of the Prophet's revelation in the Qurʾān, along puritanical lines.

The Khārijites

A Khārijite killed ʿAlī in 661. The Shīʿah thenceforth crystallized into the obverse position of the Khārijites, emphasizing ʿAlī's relationship to the Prophet as a means of making him and his descendants by Fāṭimah the sole legitimate heirs to the Prophet, some of whose spiritual power was even believed to have been transmitted to them. This Shīʿism centuries later became the official Islāmic sect of Iran. In the interim, Shīʿism was a rallying point for socially and politically discontented elements within the Muslim community. In addition to the Khārijites, another minority sect was thus formed, hostile from the beginning to the Umayyad government that seized power on ʿAlī's death. The majority of Muslims avoided both the Shīʿite and Khārijite positions but followed instead the *sunnah,* or "tradition," as these orthodox believers conceived the Prophet to have left it and as Abū Bakr, ʿUmar, ʿUthmān, and ʿAlī, too, the "rightly guided" first caliphs, had observed and codified it.

Abū Muslim's revolutionary movement was as much as anything representing Medinese mercantile interests in the Hejaz, dissatisfied with Umayyad inability to shelter Middle Eastern trade under a Pax Islamica. To promote the revolution aimed to destroy Umayyad power, the movement exploited Shīʿite aspirations and other forces of disenchantment. The Khārijites were excluded, since their movement opposed the idea of a caliphate of the kind Abū Muslim's adherents were fighting to establish—one that could command sufficient respect to hold together an Islāmic universal state. A discontented element ready to Abū Muslim's hand in Khorāsān, however, was not a religious grouping but Arab settlers and Iranian cultivators who were burdened by taxation.

In Iran, the first Arab conquerors had concluded treaties with local Iranian magnates who had assumed authority when the Sāsānid imperial government disintegrated. These notables—the *marzbān*s and landlords (*dehqān*s)—undertook to continue tax collection on behalf of the new Muslim power. The advent of Arab colonizers, who preferred to cultivate the land rather than campaign farther into Asia, produced a further complication. Once the Arabs had settled in Iranian lands, they were required to pay the *kharāj,* or land tax, collected by Iranian notables and their agents under treaty with the Muslim government. The Iranian collectors proved extortionate and aroused the hostility of both Arab and Persian.

Taxes levied

Another source of discontent was the head tax, or *jizyah,* which was applied to non-Muslims of the tolerated religions—Judaism, Christianity, and Zoroastrianism. Thus, on conversion to Islām, Iranians expected exemption from this degrading tax. But the Umayyad government, burdened with imperial expenses, often refused exemption to the Iranian converts.

The tax demands of the Damascus government were as distasteful to those urbanized Arabs and Iranians in commerce as they were to those in agriculture; and hopes of easier conditions under the new rulers than under the Sāsānids were not fully realized. The Umayyads ignored Iranian agricultural conditions, which required constant reinvestment to maintain irrigation works and to halt the encroachment of the desert. This no doubt made the tax burden, from which no returns were visible, all the more odious. Furthermore, the regime failed to maintain the peace so necessary to trade. Damascus feared the breaking away of remote provinces where the Arab colonists were becoming assimilated with the local populations. The government, therefore, deliberately encouraged tribal factionalism in order to prevent a united opposition against it.

Thus the revolution set out to establish an Islāmic oecumene above divisions and sectarianism, the Pax Islamica already referred to, which commerce required and which Iranian merchants without status in the Sāsānid social hierarchy looked to Islām to provide. Ease of communication from the Oxus to the Mediterranean was wanted but without what seemed like a nest of robbers calling themselves a government straddling the route at Damascus. In 750 Umayyad power was destroyed and the revolution gave the caliphate to the ʿAbbāsids (see ISLĀMIC WORLD).

Hejazi commercial interests had in a sense overcome the military party among leading Muslim Arabs. Greater concern for the East was manifested by the new caliphate's choice of Baghdad as its capital, situated on the Tigris a short distance north of Ctesiphon and designed as a new city, to be free of the factions of the old Umayyad garrison cities of Kūfah, Wāsiṭ, and Basra.

The ʿAbbāsid caliphate (750–821). The revolution that established the ʿAbbāsids represented a triumph of the Islāmic-Hejazi elements within the empire; the Iranian revival was yet to come. Nevertheless, ʿAbbāsid concern with fostering eastern Islām made the new caliphs willing to borrow the methods and procedures of statecraft employed by their Iranian predecessors. At Damascus the Umayyads had imitated Sāsānid court etiquette, but at Baghdad Persianizing influences went deeper and aroused some resentment among the Arabs, who were nostalgic for the legendary simplicity of human relations among the desert Arabs of yore. Self-conscious schools of manners grew up in the new metropolis, representing the competitive merits of the Arabs' or Persians' ancient ways. Regard for poetry—the Arabs' vehicle of folk memory—increased, and minds and imaginations were quickened. Philosophical enquiry was developed out of the need for precision about the meaning of Holy Writ and for the establishment of the authenticity of the Prophet's dicta, collected as *ḥadīth*s, or sayings traditionally ascribed to him, and recollected and preserved for posterity by his companions. An amalgam known as Islāmic civilization was thus being forged in Baghdad in the 8th and 9th centuries. The Iranian intellect, however, played a conspicuous part in what was still an Arab milieu. Works of Indian provenance were translated into Arabic from Pahlavi, the written language of Sāsānid Iran, notably by Ibn al-Muqaffaʿ (720–756/757). The wisdom of both the ancient East and West was received and discussed in Baghdad's schools. The metropolis's outposts confronted Byzantium as well as infidel marches in Afghanistan and Central Asia. Cultural influences came from both directions. Curiosity in the pursuit of knowledge had been enjoined by the Prophet "even as far as China." This cosmopolitanism was not new to the descendants of the urban Arabs of Mecca or to the Iranians, whose land lay across the routes from the Pacific to the Mediterranean. Both peoples knew how to transmute what was not originally their own into forms that were entirely Islāmic. Islām had liberated men of the scribal and mercantile classes who in Iran had been subject to the dictates of a taboo-ridden and excessively ritualized Zoroastrianism and in Arabia had been inhibited by tribal feuds and prejudices.

Islāmic civilization

Despite the development of a distinctive Islāmic culture, the military problems of the empire were left unsolved. The ʿAbbāsids were under pressure from the infidel on several fronts—the Turks in Central Asia, pagans in India and in the Hindu Kush, and Christians in Byzantium. War for the faith, or *jihād,* against these infidels was a Muslim duty. But while the Umayyads had been expansionists and had seen themselves as heads of a military empire, the ʿAbbāsids were more pacific and saw themselves as the supporters of more than an Arab-conquering militia. Yet rebellions within the imperial frontiers had to be contained and the frontiers protected.

Rebellion within the empire took the form of peasant revolts in Azerbaijan and Khorāsān, coalesced by popular religious appeals centred on men who assumed or were accorded mysterious powers. Abū Muslim, executed in 755 by the second ʿAbbāsid caliph al-Manṣūr, who feared his influence, became one such messianic figure. Another was al-Muquannaʿ, "the Veiled Prophet of Khorāsān," who used Abū Muslim's mystique and whose movement lasted from 777 to 780. The Khorram-dīnān, "Glad Religionists," under the Azerbaijanian Bābak (816–838) also necessitated vigorous military suppression. Not until

Revolts within the empire

two decades had elapsed was Bābak caught. He was brought to Baghdad for torture and death after so long defying the caliph in Azerbaijan and west Persia. These heresiarchs revived such creeds as that of the anti-Sāsānid religious leader Mazdak (died AD 528 or 529), expressive of social and millennial aspirations that were later canalized into Ṣūfism on the one hand and into Shīʿism on the other.

Sīstān, Iran's southeastern border area, had a tradition of chivalry as the ancient homeland of Iranian military champions. Their tales passed to posterity collectively in the deeds of Rostam, son of Zāl, in Ferdowsī's *Shāh-nāmeh* ("Book of Kings"), the Persian national epic. On the route to India, Seistan also had trade. Its agrarian masses were counterbalanced by an urban population whose economy could be bolstered by forays into still non-Muslim areas under the southern Hephthalites, the Zunbīls of the Hindu Kush's southwestern flanks, whose command of trade routes with India had to be contested when partnership in this command broke down.

Early exploitation of the province's agriculture by Arab governors had, however, debilitated the rural life, and Khārijites, who found refuge in Seistan from the Umayyads, organized or attracted bands of local peasants and of vagabonds who had strayed south from Khorāsān. The presence of these groups indicates agricultural depression following the first century of rule by nonagricultural Arabs who had failed to grasp the needs of the Iranian cultivator. Khārijite bands isolated the cities and threatened their supplies. Seistan needed an urban champion who could come to terms with the Khārijites and divert them to what could legitimately be termed *jihād* across the border, forming the gangsters into a well-disciplined loyal army. Such a man appeared in Yaʿqūb ebn Leyṯ, who founded the first purely Iranian dynasty and threatened the Muslim empire with the first resurgence of Iranian independence.

The "Iranian intermezzo" (821–1055). Yaʿqūb ebn Leyṯ's movement differed from Ṭāhir ibn al-Ḥusayn's establishment of a dynasty of Iranian governors over Khorāsān in 821. The latter's rise marks the caliph's recognition, after the difficulties encountered in Iran by Hārūn ar-Rashīd (786–809), that the best way for the imam and *amir al-muʾminīn* at Baghdad to ensure military effectiveness in eastern Islām was by appointing a great general to govern Khorāsān. Ṭāhir had won Baghdad from Hārūn's son al-Amīn in favour of his other son, al-Maʾmūn, in the civil war between the two after their father's death. Ṭāhir was descended from the *mawālī* of the great Arab leader to whom his forbears had become clients in eastern Khorāsān. He was, therefore, of Iranian origin, but, unlike Yaʿqūb, he did not emerge out of his own folk and because of a regional need. He rose as a servant of the caliphate, as whose lieutenant he was, in due course, appointed to govern a great frontier province. He made Nīshāpūr his capital. Though death followed his assumption of the right of having his name mentioned after the caliph's in the *khuṭbah* (the formal sermon at the Friday congregations of Muslims when those with authority over the community were mentioned after the Prophet), his family was sufficiently influential and respected at Baghdad for the governorship to continue in it until the Ṭāhirids were ousted from Nīshāpūr by Yaʿqūb in 873. Thereafter, they retired to Baghdad.

The Ṭāhirids

Discussion of the rise of "independent" Persian dynasties such as the Ṭāhirid in the 9th century has to be qualified not only by consideration of skillful ʿAbbāsid statecraft but also by recognition of the Muslims' need for legality in a juridical-religious setting. The majority of Muslims considered the caliph as the legitimate head of the Faith and the guarantor of the law. Such a guarantee was preeminently the need of merchants in the cities of Sīstān, Transoxania, and central Iran.

In the Caspian provinces of Gīlān and Ṭabaristān (Māzanderān) the situation was different. The Elburz Mountains had been a barrier against the integration of these areas into the caliphate. Small princely families—the Bāvands, including the Kāʾūsīyeh and the Espahbadīyeh (665–1349), and the Mosāferids, also known as Sallārids, or Kangarids (916–*c.* 1090)—had remained independent of Damascus and Baghdad in the mountains of Daylam. When Islām reached these old Iranian enclaves, it was brought by Shīʿite leaders in flight from metropolitan persecution. It was not the Islām of the Sunnī state.

The Ṣaffārids. Yaʿqūb ebn Leyṯ began life as an apprentice *ṣaffār* ("coppersmith"), hence his dynasty's name, Ṣaffārid. Taking to military freebooting, he mustered an army that he disciplined and regularly paid in cash, absorbing many Khārijites into its ranks. This and his extension of Islām into pagan areas of Sind and Afghanistan earned him the *amīr* al-Muʾminīn's gratitude, which Yaʿqūb courted by sending golden idols captured from infidels to be paraded in Baghdad. Yaʿqūb's attitude toward the imam's claiming political subservience was, nevertheless, strikingly similar to that of the caliph-rejecting Khārijites. He turned his attention inward instead of outside the pale of Islām. He seized Baghdad's breadbaskets—Fārs and Khūzestān. He drove out the Ṭāhirid emir from Nīshāpūr. His march on Baghdad itself was halted only by the stratagem devised by the caliph's commander in chief, whereby Yaʿqūb's army was inundated by burst dikes. Yaʿqūb died soon after, in 879. He had made an empire, minted his own coin, fashioned a new style of army loyal to its leader rather than to any religious or doctrinal concept, and required that verses in his praise be put into his own language—Persian—from Arabic, which he did not understand. He began the Iranian resurgence.

Yaʿqūb's empire

The collapse of the Ṭāhirid viceroyalty left Baghdad faced with a power vacuum in Khorāsān and southern Persia. Yaʿqūb's brother ʿAmr was confirmed as governor of Fārs and Khorāsān, albeit reluctantly. The caliph's recognition of the upstart was thrice withdrawn, ʿAmr's authority being disclaimed to the Khorāsānian pilgrims to the Hejaz when they passed through Baghdad. But so long as Khorāsān was victimized by the rebels ʿAbd Allāh al-Khujistānī and, for longer, Rāfiʿ ibn Harthama, ʿAmr was useful to Baghdad. After Rāfiʿ had been finally defeated in 896, ʿAmr's broader ambitions gave the caliph al-Muʿtaḍid (892–902) his chance. ʿAmr conceived designs on Transoxania, but there the Sāmānid ʿAbbāsid vassals held the caliph's license to rule, after having nominally been Ṭāhirid deputies. When ʿAmr demanded and obtained the former Ṭāhirid tutelage over the Sāmānids in 898, Baghdad could leave the Ṣaffārid and Sāmānid to fight each other, and the Sāmānid Esmāʿīl (892–907) won. ʿAmr was sent to Baghdad, where he was put to death in 902. His family survived as Sāmānid vassals in Sīstān and were heard of until the 16th century. Yaʿqūb remains a popular hero in Iranian history.

The Sāmānids. There was nothing of the popular hero in the Sāmānids' orgin. Their eponym was Sāmān-Khudā, a landlord in the district of Balkh and, according to the dynasty's claims, a descendant of Bahrām Chūbīn, the Sāsānid general. Sāmān became Muslim. His four grandsons were rewarded for services to the caliph al-Maʾmūn (813–833) and received the caliph's investiture for areas that included Samarkand and Herāt. They thus gained wealthy Transoxanian and east Khorāsānian entrepôt cities, where they could profit from trade across Asia, even as far as Scandinavia, and the provision of Turkish slaves, much in demand in Baghdad, while they protected the frontiers and provided security for merchants in Bukhara, Samarkand, Khojand, and Herāt. With one transitory exception, they upheld orthodoxy and at each new accession paid a tribute to Baghdad for the tokens of investiture from the caliph whereby their rule represented lawful authority. Thus, legal transactions in Sāmānid realms would be valid, and Baghdad received tribute in return for the insignia prayed over and signed by the caliph. This tribute took the place of regular revenue, so that it represented a solution of those taxation problems and the consequent resentments that had bedeviled the Umayyad regime. In modern assessments of imperial power, Baghdad may seem to have been politically the weaker for this type of arrangement, but ensuring the reign of Islām in peripheral provinces was important to the caliphs. Islām's portals to the Far East were adequately guarded, the supply of Turkish slaves was maintained, and Turkish pagan tribes were converted to Islām under the Sāmānid aura.

Tribute to the caliph

The Iranian renaissance. This Sāmānid aura lasted from 819 until eclipsed in 999. Its supremacy in northeastern Islām began in 875, when the Sāmānid emir, Naṣr I, re-

ceived the license to govern all Transoxania. Sāmānid emirs succeeded the Ṭāhirid-Ṣaffārid power in Khorāsān, and under them, the Iranian renaissance at last came to fruition. Shaped out of the vernacular of northeastern Iranian courts and households and making skillful use of additional Arabic vocabulary, the Persian language emerged as a literary medium. Persian notation had been used in the first Muslim *dīwān*s, or chancelleries, in accountancy, because the first civil servants in the old Iranian areas had been Iranians. In 697 the ruthless Umayyad governor Ḥajjāj ibn Yūsuf had ordered the change to Arabic notation, marking the final dethronement of Pahlavi characters. When "New Persian" began to emerge as a written language two centuries later, its alphabet was Arabic. It emerged as poetry, by which it was disciplined into a most expressive and flexible tongue, with the flexibility resulting from perfect control of a highly formal medium. The discipline was that of Arabic prosody, to which scenes of a verdure unknown to the Arab poet in the desert added, in the words of Iranian poets, a new and lustrous imagery. To rival the Arabs' tales of ancient valour, the Iranian legend was versified under Sāmānid impetus in the *Shāh-nāmeh.* Iran's "national epic" was completed by Ferdowsī of Ṭūs in Khorāsān in 1009/10.

Under the Sāmānids, Bukhara rivalled Baghdad as a cultural capital of Islām. Besides the Persian poet Rūdakī (died 940/941), who had crystallized the language and imagery of Persian lyrical poetry as Ferdowsī (died between 1020 and 1026) was to do for that of the epic, patrons such as Naṣr II (914–943) attracted Arabic poets and scholars to Bukhara, and many were bilingual. A written Persian evolved that has survived with remarkably little change.

The Ghaznavids. Rūdakī, in a poem about the Sāmānid emir's court, describes how "row upon row" of Turkish slave guards were part of its adornment. From these guards' ranks two military families arose—the Sīmjūrids and Ghaznavids—who ultimately proved disastrous to the Sāmānids. The Sīmjūrids received an appanage in the Qohestān region of southern Khorāsān. Alp Tigin founded the Ghaznavid fortunes when he established himself at Ghazna (modern Ghaznī, Afghanistan) in 962. He and Abū ol-Ḥasan Sīmjūrī, as Sāmānid generals, competed with each other for the governorship of Khorāsān and control of the Sāmānid empire by placing on the throne emirs they could dominate. Abū ol-Ḥasan died in 961, but a court party instigated by men of the scribal class—civilian ministers as contrasted with Turkish generals—rejected Alp Tigin's candidate for the Sāmānid throne. Manṣūr I was installed and Alp Tigin prudently retired to his fief of Ghazna. The Sīmjūrids enjoyed control of Khorāsān south of the Oxus but were hard pressed by a third great Iranian dynasty, the Būyids, and were unable to survive the Sāmānid collapse and rise of the Ghaznavids.

The struggles of the Turkish slave generals for mastery of the throne with the help of shifting allegiance from the court's ministerial leaders both demonstrated and accelerated the Sāmānid decline. Sāmānid weakness attracted into Transoxania the Qarluq Turks, who had recently converted to Islām. They occupied Bukhara in 992 to establish in Transoxania the Qarakhanid, or Ilek Khanid, dynasty. Alp Tigin had been succeeded at Ghazna by Sebüktigin (died 997). Sebüktigin's son Maḥmūd made an agreement with the Qarakhanids whereby the Oxus was recognized as their mutual boundary. Thus the Sāmānids' dominion was divided and Maḥmūd was freed to advance westward into Khorāsān to meet the Būyids.

The Būyids. The Būyids share with the Sāmānids the palm for having brought to fruition the Iranian renaissance. They achieved Iranian political reascendancy by doing what Yaʿqūb ibn Leyš had failed to do and what the Sāmānids would probably have considered it illegal to do: they captured Baghdad and made the caliph their puppet. As far east as the city of Rayy, western, central, and southern Iran were once more ruled by an Iranian family. At the peak of the Būyid Empire, second to Baghdad the Būyid base became Fārs, whence the Achaemenids and the Sāsānians had sprung. Politically, the Būyids effected the Iranization of the metropolitan government. Yet by the very fact that they saw in the caliphate an institution of enough purely political significance to merit its dramatic takeover, they paradoxically left the caliphate's political role emphasized by what at first sight might seem to have been deepest humiliation. Spiritually, the caliphate held no appeal for the Būyids, who were Shīʿite. Politically and juridically, as the stabilizing factor over the Islāmic peoples, the Būyids, in spite of their own religious affiliation, maintained the caliphate.

The homeland of the Būyids (or Buwayhids) was Daylam, in the Gīlān uplands in northern Iran. There, at the end of the 9th century, hardy valley dwellers had been stirred into martial activity by a number of factors, among them the rebel Rāfiʿ ibn Harthama's attempt to penetrate the region, ostensibly with Sāmānid support. ʿAmr ebn Leyš had pursued the rebel into the region. Other factors had been the formation of Shīʿite principalities in the area and the continued Sāmānid attempts to subjugate them. After the Ṭāhirid collapse, the lack of stability in northern Iran south of the Elburz Mountains attracted many Daylamite mercenaries into the area on military adventures. Among them, Mākān ebn Kākī served the Sāmānids with his compatriots, the sons of Būyeh, and their allies the Zeyārids under Mardāvīz. Mardāvīz introduced the three Būyid brothers to the Iranian Plateau, where he established an empire reaching as far south as Isfahan and Hamadān. He was murdered in 935 but his Zeyārid descendants sought Sāmānid protection. They adhered to the Sunnī orthodoxy and maintained themselves in the region southeast of the Caspian. The Zeyārid Qābūs ebn Vashmgīr (978–1012) built himself a tomb tower, the Gonbad-e Qābūs (1006–07), which remains one of Iran's finest monuments. Also still extant is his descendant ʿOnṣor ol-Maʿālī Keykāvūs' (1049–90), *Qābūs-nāmeh,* a prose "Mirror for Princes," which is a valuable document on the social and political life of the time.

Foundation of Būyid power

Mardāvīz' expansionism south of the Elburz was taken up by his Būyid lieutenants: the eldest brother, ʿAlī, consolidated power for himself in Isfahan and Fārs and obtained the Caliph's recognition; another brother, Ḥasan, occupied Rayy and Hamadān; the youngest brother, Aḥmad, took Kermān in the southeast and Khuzistan in the southwest. The caliphs al-Muttaqī (940–944) and Mustakfī (944–946) were at the mercy of the Turkish slaves in their palace guard. The generals of the guard competed with each other for the office of *amīr alumarā'* (commander in chief), who virtually ruled Iraq on behalf of the caliphs. When Aḥmad gained Khūzestān, he was close to the scene of the *amīr al-umarā'* contests, which he chose to settle by himself. Aḥmad entered Baghdad in 945 and assumed control of the caliphate's political functions. The caliph became the Būyid's protégé, and conferred on Aḥmad the title of Muʿizz ad-Dawlah. ʿAlī became ʿImād ad-Dawlah, and Ḥasan, Rukn ad-Dawlah. All of these titles implied that the Būyids were the upholders of the Muslim ʿAbbāsid *dawlah,* or state. In practice, however, the *dawlah* became the "Daylamite State." It should be noted that the titles did not include the world *dīn,* or faith, which was still in the caliph's sphere of action.

Later, Būyid titles increased in grandeur. Even the old Achaemenid title of *shāhanshāh,* king of kings, reappeared—a title Aḥmad may have thought appropriate for an Iranian whose family reconquered Iran south of the Elburz. As suggested above, Bùyid titles emphasized political and territorial sovereignty. This sovereignty reached its greatest extent under Rukn ad-Dawlah's son, ʿAḍud ad-Dawlah, who, after the deaths of his father and uncles, ruled an empire that comprised all of Persia west and south of Khorāsān and included Iraq, with Baghdad at its heart. ʿAḍud ad-Dawlah pursued peace negotiations with Byzantium, perhaps to free himself for his cherished project of an Egyptian campaign against the rival caliphate of the Fāṭimid Shīʿites, established in North Africa in 909, which had been relocated in Egypt in 969. ʿAḍud ad-Dawlah's concern with the middle kingdom and its westward extension toward the Mediterranean increased his hostility toward the Fāṭimids, despite his own Shīʿī persuasion. In the north, he drove the Zeyārids out of Ṭabarestān, which struck a blow against the Sāmānids' influence in the Caspian area.

Aḍud ad-Dawlah is celebrated for public works, of which the dam he built across the River Kor near Shīrāz, the "Band-e amīr," Emir's Dam, remains. He embellished the

tomb of ʿAlī at an-Najaf in Iraq, where he himself is buried. He built libraries, schools, and hospitals, and he was the patron of the Arabic poet Mutanabbī. Some Arabic verses of his own are still extant. Although ʿAḍud ad-Dawlah was undoubtedly one of Iran's greatest rulers, his fratricidal wars, conducted with terrible intractability on his way to power, initiated Būyid decline. The descendants of the early Būyids reversed the mutual fidelity of the first three brothers. The power this fidelity had achieved and ʿAḍud ad-Dawlah had made into a world force crumbled after his death in 983.

His base had been Shīrāz, which he beautified and established as a cultural centre, but he died at Baghdad, where he chose to keep close to the caliph, whose daughter he married and from whom he took the title "The Crown of the Community" and the privilege of, like the caliph, having drums beaten at his gate on the calls to prayer. He also had his name mentioned after that of the Caliph aṭ-Ṭaʾiʿ in the *khuṭbah*. The Būyids avoided the policy, which would have disrupted the empire, of favouring the Shīʿīs. Instead, they offered consolations of an emotional sort to the Shīʿīs in the form of public rites on the anniversaries of the Shīʿī martyrs, notably the one commemorating the massacre of ʿAlī's son Ḥusayn and his followers under the Umayyads at Karbalāʾ in Iraq.

ūyid ecline

Although the Būyids were careful to avoid sectarian strife, family quarrels weakened them sufficiently for Maḥmūd of Ghazna to gain Rayy in 1029. But Maḥmūd (998–1030) went no further: his dynasty paid great deference to the caliphate's legitimating power, and he made no bid to contest the Būyids' role as its protectors. Maḥmūd's agreement with the Sāmānids' Ilek Khanid successors, that the Oxus should be their and his boundary, held, but south of the river the Ghaznavids had to contend with their own distant relatives, the Oğuz Turks. Contrary to the sage counsel of Iranian ministers, Maḥmūd and his successor Masʿūd (1031–41) permitted these tribesmen to use Khorāsānian grazing grounds, which they entered from north of the Oxus. United under descendants of an Oğuz leader named Seljuq, between 1038 and 1040 these nomads drove the Ghaznavids out of northeastern Iran. The final encounter was at Dandānqān in 1040.

After their defeat by the Seljuqs, the Ghaznavids, patrons of Islāmic culture and letters, were deflected eastward into India, where Maḥmūd had already conducted successful raids. The raids took the form of *jihād*, or holy war, and the Ghaznavids carried Islām and Perso-Muslim art to the Indian subcontinent. In Iran, it was the Seljuqs' turn to create a new imperial synthesis with the ʿAbbāsid caliphs. Toghrïl Beg, the Seljuq sultan, entered Baghdad in 1055, and Būyid power was terminated, thus ending what Vladimir Minorsky, the great Iranologist, called the "Iranian intermezzo."

The Seljuqs and the Mongols. *The Seljuqs.* Toghrïl I had proclaimed himself sultan at Nīshāpūr in 1038 and had espoused strict Sunnī orthodoxy by which he gained the caliph's confidence and undermined the Būyid position in Baghdad. The Oğuz Turks had accepted Islām late in the 10th century, and their leaders displayed a convert's zeal in their efforts to restore a Muslim polity along orthodox lines. Their efforts were made all the more urgent by the spread of Fāṭimid propaganda by underground means in the eastern caliphate and by the threat posed by the Christian crusaders from the west.

The Būyids' usurpation of the caliph's secular power had given rise to a new theory of state formulated by al-Māwardī (died 1058). Al-Māwardī's treatise partly prepared the theoretical ground for Toghrïl's attempt to establish an orthodox Muslim state in which conflict between the caliph-imam's spiritual-juridical authority on the one side and the secular power of the sultan on the other could be resolved, or at least regulated, by convention. Al-Māwardī reminded the Muslim world of the necessity of the imamate; but the treatise realistically admitted the existence of, and thus accommodated, the fact of military usurpation of power. The Seljuqs' own political theorist al-Ghazālī (died 1111) carried this admission further by explaining that the position of a powerless caliph, overshadowed by a strong Seljuq master, was one in which the latter's presence guaranteed the former's capacity to defend and extend Islām.

The caliph al-Qāʾim (1031–75) replaced the last Būyid's name, al-Malik ar-Raḥīm, in the *khuṭbah* and on the coins with Toghrïl Beg's; and after protracted negotiation ensuring restoration of the caliph's dignity after Shīʿī subjugation, Toghrïl entered Baghdad in December 1055. The caliph enthroned him and married a Seljuq princess. After Toghrïl had campaigned successfully as far as Syria, he was given the title of "King of the East and West." The new situation was justified by the theory that existing practice was legal whereby a new caliph could be instituted by the sultan, who possessed effective power and sovereignty, but that thereafter the sultan owed the caliph allegiance because only so long as the caliph-imam's juridical faculties were recognized could government be valid.

Toghrïl Beg died in 1063. His heir, Alp-Arslan, was succeeded by Malik-Shāh in 1072, and the latter's death in 1092 led to succession disputes out of which Berk-yaruq emerged triumphant to reign until 1105. After a brief reign, Malik-Shāh II was succeeded by Muḥammad I (1105–18). The last "Great Seljuq" was Sanjar (1118–57), who had earlier been governor of Khorāsān. Alp-Arslan nearly annihilated the Byzantine Army at Manzikert in 1071, opening Asia Minor to those dependent tribesmen of the Seljuqs of whom Iran and the world were to hear more in the period of Ottoman power. Transoxania was subdued, the Christians in the Caucasus chastised, and the Egyptians expelled from Syria. An empire was for a short time achieved whose extent and stability enabled Alp-Arslan's and Malik-Shāh's great minister, Niẓām al-Mulk (died 1092), to pay an Oxus ferryman with a draft cashable in Damascus.

Battle of Manzikert

The building and maintenance of such a great empire necessitated a military regime and a vast war machine. The price to be paid later was oppression by military commanders and their units, set free after the machine fell out of the grasp of powerful sultans to compete with each other and harry the land. The soldiers had been remunerated by grants of land called *iqṭāʿs*. The grants later became nuclei out of which petty principalities grew with the decline of the central power. The cultivators were left at the mercy of military overlords in possession of the soil.

The great minister Niẓām al-Mulk was typical of the Iranian bureaucracy, which, in an area prone to invasion, was often called upon to attempt to cushion the impact of the brute military force of invaders and contain it within the bounds of administrative, economic, and cultural feasibility. He wrote a "Book of Government" (*Seyāsat-nāmeh*) for his Turkish masters in which he urged the regulation of court procedures in line with Sāmānid models and the restriction of the arrogance and cupidity of the military fief holders. His book is the measure of the Seljuqs' failure to provide enduring stability and equitable government. Had they done so, such a work would have been unnecessary.

The Ismāʿīlīyah. Of one disruptive force the book, in terms betraying near-panic, is dramatically descriptive. The Seljuqs failed to nip in the bud the power of the Ismāʿīlīyah, originally spread by Fāṭimid propagandists, but later split from the mainstream of events in Egypt to become an underground organization within the Seljuq Empire. This movement exercised power by terrorism, and the name given its adherents by Europeans in the Middle Ages, Assassins (from *ḥashīshī*, denoting a consumer of hashish), has become a common noun in English. The movement was a highly intellectual form of extremist Shīʿism, which recognized only seven of the imams in descent from ʿAlī and Fāṭimah, whereas the Shīʿism of the Būyids and of Iran later on recognized 12. The movement was brought to Iran by Ḥasan-e Ṣabbāḥ, who had been trained in Fāṭimid Egypt. In 1090 Ḥasan gained the castle of Alamūt in the Elburz Mountains, and the order's principal cells were thereafter situated, so far as possible, in similar impregnable mountain strongholds. From these centres, *fidāʾīs*, or devotees ready to sacrifice their lives, issued forth and permeated society, spreading their mission as peddlers and itinerant tailors and gaining influence among the urban artisan and weaving classes. They were often able to win also the confidence of many highly placed ladies and children, whom they could please with novelties of dress or toys. Niẓām al-Mulk himself was

The Assassins

assassinated by one of the *fidā'īs*, but it is possible that this was done with the connivance of Malik-Shāh's wife, whose son the vizier did not support for the succession.

The Ismā'īlīyah were able to puncture Seljuq power but not destroy it. In the end, the Seljuq Empire collapsed where it had begun—in Khorāsān, where Sultan Sanjar ultimately failed to control Turkmen tribes related to him by blood. Sanjar could not rely on military commanders his family had raised to high posts and had rewarded with land and provincial powers. The tribesmen refused to be coerced into paying taxes. In 1153 they captured the old sultan and, although allowing him all the respect of his regal position, kept him captive for three years.

The Khwārezm-Shāhs. Atsiz was the military leader who, after Sultan Sanjar's death in 1157, succeeded in supplanting Seljuq power in northeastern Iran. His ancestor, Anūshtegin, had been keeper of Malik-Shāh's kitchen utensils and had been rewarded with the governorship of Khwārezm on the Oxus, where he founded the Khwārezm-Shāh dynasty (1077–1231). Regions elsewhere in Iran, on the passing of Seljuq supremacy, became independent under *atabeg*s, who were originally proxy fathers and tutors sent with young Seljuq princes when these were deputed to govern provinces. At first the *atabeg*s took power in the names of Seljuq puppets. When this fiction lapsed, *atabeg* dynasties such as the Eldegüzids of Azerbaijan (1137–1225) and Salghurids of Fārs (1148–1270) split Iran into independent rival principalities.

Atabeg rule

The Salghurid court in Shīrāz especially fostered the arts, as parvenu, competitive courts are wont to do. The poet Sa'dī (died 1292) was a contemporary in Shīrāz of the Salghurid *atabeg* Abū Bakr ebn Sa'd ebn Zangī (1231–60), to whom he alludes by name in his *Būstān* ("The Orchard"), a book of ethics in verse. Abū Bakr's father, Sa'd, after whom Sa'dī took his pen name, conferred great prosperity on Shīrāz.

Sa'd ebn Zangī came to terms with the Khwārezm-Shāhs. Their power in Transoxania was secured by acceptance of tributary status to the infidel Karakitais of Central Asia. They endeavoured to emulate the Seljuqs by following an expansionist policy in Iran south of the Oxus. Sa'd ebn Zangī, in his relations with the Khwārezm-Shāh, set the pattern his successor Abū Bakr followed later. These *atabegs* saved Fārs from outright invasion by northern military powers by paying heavy tribute. This tribute was the price of Shīrāz's remaining the peaceful haven of the arts in which Sa'dī and after him Ḥāfeẓ (died 1390) flourished, to continue the Persian literary tradition begun under the Sāmānids and continued under both the Ghaznavids and the Seljuqs.

The collapse of the Karakitai Empire northeast of the Oxus was partly accelerated by the unsuccessful bid of Khwārezm-Shāh 'Alā' ad-Dīn Muḥammad (1200–20) to win Muslim approval while releasing himself from the Khwārezm-Shāhs' humiliating tributary status to an infidel power. But the *coup de grace* to the Karakitai Empire was delivered by its Gurkhan's own vassal from the east, Küchlüg Khan. Thus, the Khwārezm-Shāh was from 1211 onward faced by another hostile opponent in Central Asia—Küchlüg Khan. The Karakitai settlement had been destroyed; but the situation on the Khwārezm-Shāh's eastern border had worsened.

Meanwhile, Sultan 'Alā' ad-Dīn Muḥammad quarrelled with the caliph; he set up an anti-caliph of his own and further antagonized his Muslim subjects, unremittingly suspicious of a regime once subject to the Karakitai infidels and whose Kipchak mercenary militia and brutal commanders brought cruelty and desolation wherever they marched. Sultan Muḥammad Khwārezm-Shāh was unable to control his army leaders, who had tribal connections with such influential people at court as his own mother. The post-Karakitai wars between him and Küchlüg Khan damaged the safety of the Central Asian trade arteries from China to the West. The great Mongol leader Genghis (Chinghiz) Khan took Peking in 1215 and as Lord of China was concerned with Chinese trade outlets. The situation between Küchlüg and Sultan Muḥammad afforded scope as well as a pretext for the Mongols' westward advance, if only to restore the flow of trade.

The Mongol invasion. Misunderstanding of how essentially fragile Sultan 'Alā' ad-Dīn Muḥammad Khwārezm-Shāh's apparently imposing empire was, its distance away from the Mongols' eastern homelands, the strangeness of new terrain, all doubtless induced fear in Mongols, and this might partly account for the terrible events with which Genghis Khan's name has ever since been associated. The terror his invasion brought must also be ascribed to his quest for vengeance. Genghis Khan's first two missions to Khwārezmia had been massacred; but the place of commercial motives in the Mongol's decision to march to the west is indicated by the fact that the first was a trade mission. The massacre and robbery of this mission at Utrār by one of Sultan Muḥammad's governors before it reached the capital made Genghis single out Utrār for especially savage treatment when the murder of his second, purely diplomatic, mission left him no alternative but war.

Reasons for Mongol vengeance

His guides were Muslim merchants from Transoxania. They had to witness one of the worst catastrophes of history. The years 1220 and 1221 saw the razing of Bukhara, Samarkand, Herāt, Ṭūs, and Nīshāpūr, and the slaughter of whole populations. The Khwārezm-Shāh fled, to die on an island off the Caspian coast. His son Jalāl ad-Dīn survived until murdered in Kurdistan in 1231. He had eluded Genghis Khan on the Indus, across which his horse swam, enabling him to escape to India. He returned to attempt the restoration of the Khwārezmian Empire over Iran, but he failed to unite the Iranian regions, although Genghis Khan had withdrawn to Mongolia, where he died in August 1227. Iran was left divided, Mongol agents remaining in some districts, local adventurers profiting from the lack of order in others.

The Il-Khans. A second Mongol invasion began when Genghis Khan's grandson Hülegü Khan crossed the Oxus in 1256 and destroyed the Assassin fortress at Alamūt. With the disintegration of the Seljuq Empire, the caliphate had created a state in the area around Baghdad and in southwestern Iran. In 1258 Hülegü besieged Baghdad, where divided counsels prevented the city's salvation. Al-Musta'ṣim, the last 'Abbāsid caliph of Baghdad, was executed by being kicked to death. Eastern Islām fell to pagan rulers.

Hülegü hoped to consolidate Mongol rule over western Asia and to extend the Mongol Empire as far as the Mediterranean, an empire that would span the earth from China to the Levant. Hülegü made Iran his base, but the Mamlūks of Egypt (1250–1517) prevented him and his successors from achieving their great imperial goal. Instead, a Mongol dynasty, the Il-Khans, or "deputy khans" to the Great Khan in China, was established in Iran to attempt repair of the damage of the first Mongol invasion. But it failed in repeated attempts to reach the Mediterranean. The injuries Iran had suffered went deep, but it would be unfair to attribute them all to Ghengis Khan's invasion, itself the climax to a long period of social and political disarray under the Khwārezm-Shāhs and dating from the decline of the Seljuqs.

The Il-Khanid dynasty made Azerbaijan its centre, and established Tabrīz as its first capital until Solṭānīyeh were built early in the 14th century. At first repair and readjustment of a stricken society were complicated by the collapse of law. The caliphate, as the symbol of Muslim legality, had been eroded by Sultan Muḥammad Khwārezm-Shāh and by its own withdrawal into a temporal state in Iraq and the Tigris–Euphrates estuary region. But it had retained enough vitality for Sultan Muḥammad's action in setting up an anti-caliph to have alienated influential members of his subject people. After 1258 it was gone altogether, while Hülagü Khan showed considerable religious eclecticism and had, in any event, the *yāsā* of Genghis Khan to apply as the law of the Mongol state, in opposition to, or side by side with, the Sharī'ah, the law of Islām.

The Il-Khans' religious toleration released Christians and Jews from their restrictions under the Islāmic regime. Fresh talent thus became available, but competition for new favours marred what good effects this release might have had. It took time for Iranian administrators to resume their normal role after the invasion and to restore some semblance of administrative order and stability. Their process was impeded by the paganism of the new conquerors as well as by jostling for influence among classes of the conquered, not in this instance exclusively Muslim. At the same time,

a shattered agrarian economy was burdened by heavy taxes, those sanctioned by the Sharīʿah being added to by those the *yāsā* provided for, so that the pressure of exploitation was increased by Mongol tax innovations as well as by the invaders' cupidity.

The pressure was increased beyond the economy's endurance: the Il-Khanid government ran into fiscal difficulties. An experiment with paper money, modelled on the Chinese money, failed under Gaykhatu (1291–95). Gaykhatu was followed briefly by Baydu (died 1295), who

he reign
f Ghāzān

was supplanted by the greatest of the Il-Khans, Maḥmūd Ghāzān (1295–1304). Ghāzān abandoned Buddhism, the faith in which his grandfather Abagha, Hülegü's successor (1265–82), had reared him, and adopted Islām. One of his chief ministers was also his biographer, Rashīd ad-Dīn, of Jewish descent. He seems deliberately to have striven to present Ghāzān, whom he styles the "emperor of Islām" (*pādshāh-e islām*), as a ruler who combined the qualities and functions of both the former caliphs and ancient Iranian "great kings."

Ghāzān made strenuous efforts to regulate taxes, encourage industry, bring wasteland into cultivation, and curb the abuses and arrogance of the military and official classes. Facilities for home and foreign merchants were furnished. Buildings were constructed and irrigation channels dug. Medicinal and fruit-bearing plants were imported and the cultivation of indigenous ones was encouraged. Observatories were built and improved—a sure indication of concern with agricultural improvement, for seasonal planning required accurate calendars. He fostered Muslim sentiment by showing consideration for those who claimed descent from the Prophet's family, the Sayyids, and it seems probable that he wished to eradicate or overlay Shīʿī-Sunnī sectarian divisiveness, for Ghāzān's Islām appears to have been designed to appeal equally to both persuasions. Any slight bias in favour of the Shīʿah might be attributed to a desire to capture the emotions and imagination of many of the humble people who had reacted against the Seljuqs' orthodoxy and craved a teaching that included millennial overtones. Shīʿism had been liberated by the fall of the ʿAbbāsid Caliphate. Its belief in the reappearance of the Twelfth Imam, who was to inaugurate peace and justice in the world, satisfied this popular craving for religious solace.

Ghāzān's work was carried on but less successfully by his successor Öljeitü (1304–16). Between 1317 and 1335, though he finally relinquished the expensive campaigns against Egypt for the opening to the Mediterranean, Abū Saʿīd was unable to keep the Il-Khanid regime consolidated, and it fell apart on his death. Ghāzān's brilliant reign survives only in the pages of his historian, Rashīd ad-Dīn. Wars against Egypt and their own Mongol kinsmen in Asia had in fact hampered the Il-Khans in accomplishing a satisfactory reintegration of an Iranian polity.

As the *atabeg*s had done after the Seljuqs, after 1335, Il-Khanid military emirs began to establish themselves as independent regional potentates. At first, two of them, formerly military chiefs in the Il-Khans' service, competed for power in western Iran, ostensibly acting on behalf of rival Il-Khanid puppet princes. Ḥasan the Small of the Chūpānids was eventually defeated by Ḥasan the Tall of the Jalāyirids, who set up the Jalāyirid dynasty over Iraq, Kurdistan, and Azerbaijan; it lasted from 1336 to 1432. In Fārs, Il-Khanid agents, the Injuids, after a spell of power during which Abū Isḥāq Injū had been the poet Ḥāfeẓ's patron, were ousted by Abū Saʿīd's governor of Yazd, Mobārez od-Dīn Moẓaffar. Thus in 1353 Shīrāz became the Moẓaffarid dynasty's capital, which it remained until conquest by Timur (Tamerlane) in 1393.

The Timurids and Turkmen. Timur claimed descent from Genghis Khan's family. The disturbed conditions in Mongol Transoxania gave this son of a minor government agent in the town of Kish the chance to build up a kingdom in Central Asia in the name of the Chagatai Khans, whom he eventually supplanted. He entered Iran in 1380 and in 1393 reduced the Jalāyirids after taking their capital, Baghdad. In 1402 he captured the Ottoman Sultan Bayezid at Ankara. He conquered Syria and then turned his attention to campaigns far to the east of his tumultuously acquired and ill-cemented empire: he died in 1405 on an expedition to China. He left an awesome name and an ambiguous record of flights of curiosity into the realms of unorthodox religious beliefs, history, and every kind of enquiry concerning lands and peoples. He showed interest in Ṣūfism, a form of Islāmic mysticism that varied from a scholastic study of ascetic techniques for mastering the carnal self to complete abandonment of all forms of authority in the belief that faith alone is necessary for salvation. Ṣūfism had increased in the disturbed post-Seljuq era as both the consolation and refuge of desperate people. In Ṣūfism Timur may have hoped to find popular leaders whom he could use for his own purposes. His encounters with such keepers of the consciences of harried, exploited, and ill-treated Iranians proved that they knew him perhaps better than he knew himself. Whatever his motives may have been, the reverse of stability was his legacy to Iran. His division of his ill-assimilated conquests among his sons served to ensure that an integrated Timurid Empire would never be achieved.

Character of Timur

The nearest a Timurid state came to being an integrated Iranian Empire was under Timur's son Shāh Rokh Shāh (1405–47). Shāh Rokh endeavoured to weld Azerbaijan and western Persia to Khorāsān and eastern Persia to form a united Timurid state for a short and troubled period of time, but he only succeeded in loosely controlling western and southern Iran from his beautiful capital at Herāt. Azerbaijan demanded three major military expeditions from this pacific sovereign and even so could not long be held. Herāt he made the seat of a splendid culture, the atelier of great miniature painters (Behzād notable among them), and the home of a revival of Persian poetry, letters, and philosophy. This revival was not unconnected with an effort to claim for an Iranian centre once more the palm of leadership in the propagation of Sunnī ideology: Herāt sent copies of Sunnī canonical works on request to Egypt. The reaction in Shīʿism's ultimate victory under the Ṣafavid shahs of Persia was, however, already in preparation.

"Black Sheep" and "White Sheep"

Western Iran was dominated by the Kara Koyunlu, the "Black Sheep" Turkmen. In Azerbaijan they had supplanted their former masters, the Jalāyirids. Timur had put these Kara Koyunlu to flight, but in 1406 they regained their capital, Tabrīz. On Shāh Rokh's death, Jahān Shāh (reigned *c.* 1438–67) extended Kara Koyunlu rule out of the northwest deeper into Iran at the Timurids' expense. The Timurids relied on their old allies, the Kara Koyunlus' rival Turkmen of the Ak Koyunlu, "White Sheep" clans, who had long been established at Diyarbakir in Turkey. The White Sheep acted as a curb on the Black Sheep, whose Jahān Shāh was destroyed by the Ak Koyunlu Uzun Ḥasan by the end of 1467.

Uzun Hasan (1453–78) achieved a short-lived Iranian Empire and even briefly deprived the Timurids of Herāt. He was, however, confronted by a new power in Asia Minor—the Ottoman Turks. His relationship with the Christian Emperor at Trebizond (Trabzon) through his Byzantine wife, Despina, involved Uzun Ḥasan in attempts to shield Trebizond from the ineluctable Ottoman advance. The Ottomans crushingly defeated him in 1473. Under his son Yaʿqūb (1478–90), the Ak Koyunlu state was subjected to fiscal reforms associated with a government-sponsored effort to reapply rigorous purist principles of orthodox Islāmic rules for revenue collection. Yaʿqūb attempted to purge the state of taxes introduced under the Mongols and not sanctioned by the Muslim canon. But the inquiries made by the orthodox religious authorities antagonized the vested interests, damaged the popularity of the Ak Koyunlu regime, and discredited Sunnī fanaticism.

This attempt to revive religious orthodoxy through revenue reform or to effect the latter under the guise of religion no doubt gave impetus to the spread of Ṣafavid Shīʿī propaganda. Another factor must have been related to the same general economic decline that made Sultan Yaʿqūb's fiscal reforms necessary in the first place. Sheykh Ḥeydar led a movement that had begun as a Ṣūfī order under his ancestor Sheykh Ṣafī od-Dīn of Ardabīl (1252–1334). This order may be considered to have originally represented a puritanical, but not legalistically so, reaction against the sullying of Islām, the staining of Muslim lands, by the Mongol infidels. What began as a spiritual, otherworldly reaction against irreligion and the betrayal of spiritual aspirations developed

into a manifestation of the Shīʿī quest for dominion over a Muslim polity. By the 15th century, the Ṣafavid movement could draw on both the mystical emotional force and the Shīʿī appeal to the oppressed populace to gain a large number of dedicated adherents. Sheykh Ḥeydar inured his numerous followers to warfare by leading them on expeditions from Ardabīl into the nearby Caucasus. He was killed on one of these campaigns. His son Esmāʿīl was to avenge his death and lead his devoted army to a conquest of Iran whereby Iran gained a great dynasty, a Shīʿī regime, and in most essentials its shape as a modern nation state.

Esmāʿīl

In 1501 Esmāʿīl supplanted the Ak Koyunlu in Azerbaijan. Within a decade, he gained supremacy over most of Iran as a ruler regarded as divinely entitled to sovereignty. The Ṣafavids claimed descent, on grounds that modern research has shown to be dubious, from the Shīʿah imams. Islām in Iran, therefore, could regard itself as at last having a legitimate imam-ruler, who, as a descendant of ʿAlī, required no caliph to legitimate his position. Gone were the days of rule by converted and zealous Sunnī Turks or by Mongols of ambiguous spiritual allegiance. Iran's defilement was removed by the swelling tide of Shīʿism, which bore Esmʿīl to the throne his family was to occupy until 1722, in one of the greatest epochs of Iranian history.

The Ṣafavids (1502–1736). The new Iranian empire lacked the resources that had been available to the caliphs of Baghdad in former times through their dominion over Central Asia and the West: Asia Minor and Transoxania were gone, and the rise of maritime trade in the West was detrimental to a country whose wealth had depended greatly on its position on important east-west overland trade routes. The rise of the Ottomans impeded Iranian westward advances and contested with the Ṣafavids' control over both the Caucasus and Mesopotamia. Years of warfare with the Ottomans had imposed a heavy drain on the Ṣafavids' resources. The Ottomans threatened Azerbaijan itself. Finally, in 1639 a treaty gave Yerevan in the southern Caucasus to Iran and Baghdad in Mesopotamia to the Ottomans.

Shāh ʿAbbās I. The Ṣafavids were still faced with the problem of making their empire pay. The silk trade, over which the government held a monopoly, was a primary source of revenue. Esmāʿīl's successor, Shāh Ṭahmāsp (1524–76), encouraged carpet weaving on the scale of a state industry. Shāh ʿAbbās I (1588–1629) established trade contacts directly with Europe, but Iran's remoteness from Europe, behind the imposing Ottoman screen, made maintenance and promotion of these contacts difficult and sporadic. Shāh ʿAbbās I also transplanted a colony of industrious and commercially astute Armenians from Jolfā in Azerbaijan to a new Jolfā adjacent to Isfahan, the city he developed and adorned as his capital. The Ṣafavids had earlier moved their capital from the vulnerable Tabrīz to Kazvin. Since the Uzbek menace from east of the Caspian had been overcome, ʿAbbās could move his capital south to Isfahan, more centrally placed than Kazvin for control over the whole country, and for communication with the trade outlets of the Persian Gulf. ʿAbbās engaged English help to oust the Portuguese from the island of Hormuz in 1622. He also strove to lodge Ṣafavid power strongly in Khorāsān, where he developed the shrine of ʿAlī ar-Riḍā at Mashhad, the eighth Shīʿah imam, as a pilgrim centre to rival Shīʿī holy places in Mesopotamia, where visiting pilgrims took currency out of Ṣafavid into Ottoman territory.

Army reforms

Under Shāh ʿAbbās I Iran prospered. The old Ṣūfī bands, which had been formed into artificial tribal units for military purposes during the dynasty's formative period, were replaced by a standing army trained and equipped on European lines with the advice of Robert Sherley. Sherley was an English adventurer versed in artillery tactics who, accompanied by a party of cannon founders, reached Kazvin with his brother Anthony in 1598. The bureaucracy, too, was carefully reorganized, but the seeds of the sovereignty's weakness lay in the royal house itself, which lacked an established system of inheritance by primogeniture. A reigning shah's nearest and most acute objects of suspicion were his own sons. Among them, brother plotted against brother over who should succeed on their father's death. Intriguers, ambitious for influence in a subsequent reign, supported one prince against another. Shāh ʿAbbās did not adopt the Ottoman sultans' practice of eliminating royal males by murder (as a child he had been within a hair's breadth of being a victim of such a policy). Instead, he instituted the practice of immuring infant princes in palace gardens away from the promptings of intrigue and the world at large. As a result, his successors tended to be indecisive men, easily dominated by powerful religious dignitaries to whom the Ṣafavids had accorded considerable influence in an attempt to make Shīʿism the state religion.

The Afghan interlude. Shāh Solṭān Ḥoseyn (reigned 1694–1722) was of a religious temperament and especially influenced by the divines, whose conflicting advice, added to his own procrastination, sealed the sudden and unexpected fate of the Ṣafavid empire. One Maḥmūd, a former Ṣafavid vassal in Afghanistan, captured Isfahan and murdered Solṭān Ḥoseyn in his cell in the beautiful mosque-school, or *madrasah,* built in his mother's name.

The Afghan interlude was disastrous for Iran. In 1723 the Ottomans took advantage of the disintegration of the Ṣafavid realm and invaded from the west, ravaging western Persia. Nāder, an Afshārid Turkmen from northern Khorāsān, was eventually able to reunite Iran, a process he began on behalf of the Ṣafavid prince Ṭahmāsp II (1722–32), who had escaped the Afghans. After he had cleared the country of Afghans, Nāder was made governor of a large area of eastern Iran.

Nāder Shāh (1736–47). Nāder later dethroned Ṭahmāsp II in favour of another Ṣafavid puppet. His successful military exploits, however, which included victories over rebels in the Caucasus, made it feasible for this stern warrior himself to be proclaimed shah in 1736. He attempted to mollify Perso-Ottoman hostility by establishing in Iran a less aggressive form of Shīʿism, which would be less offensive to Ottoman susceptibilities; but this experiment did not take root. Want of money drove him to embark on his celebrated Indian campaign in 1738–39. His capture of Delhi and of the Mughal emperor's treasure gave Nāder booty in such quantities that he was able to exempt Iran from taxes for three years. His Indian expedition temporarily solved the problem of how to make his empire financially viable.

Attack on India

How large this problem loomed in Nāder's mind is demonstrated by his increasingly morbid obsession with treasure and jewels. After suspecting his son of complicity in a plot against him in 1741, Nāder's mind seems to have become unhinged; his brilliance and courage deteriorated into a meanness and capricious cruelty that could no longer be tolerated. In 1747 he was murdered by a group of his own Afshārs, together with some Qājār chiefs—a sad end to one of Iran's greatest leaders.

Nāder had been the first modern Iranian leader to perceive the importance of having his own navy, and in 1734 he had appointed an "Admiral of the Gulf." Ships were purchased from their British captains, and by 1735 the new Iranian navy had attacked Basra. What really mattered, however, were the land forces. Nāder's reign exemplified the fact that to be successful, a shah of Iran had to prove himself capable of defending his realm's territorial integrity and of extending its sources of wealth and production by conquest. To these ends, Nāder built up a large army comprised of tribal units under their own chiefs, such as his Afshārid kinsmen, and the Qājārs and Bakhtyārīs.

But on Nāder's death, his great military machine dispersed, its commanders bent on establishing their own states. Aḥmad Shāh Durrānī founded a kingdom in Afghanistan based on Qandahār. Shāh Rokh, Nāder's blind grandson, succeeded in maintaining himself at the head of an Afshārid state in Khorāsān, its capital at Mashhad. The Qājār chief Moḥammad Ḥasan took Mazanderan south of the Caspian Sea. Āzād Khān, an Afghan, held Azerbaijan, whence Moḥammad Ḥasan Khān Qājār ultimately expelled him. The Qājār chief, therefore, disposed of this post-Nāder Afghan remnant in northwestern Iran, but was himself unable to make headway against a new power arising in central and southern Iran, that of the Zands.

The Zand dynasty (1750–79). Moḥammad Karīm Khān Zand began his career as an ally of the Bakhtyārī chief ʿAlī Mardān Khān in a bid to oust Shāh Rokh's nominee, Abū ol-Fatḥ Bakhtyārī, from Isfahan. Victory over this representative of the Mashhad regime having been achieved, it was

agreed that Shāh Solṭān Ḥoseyn Ṣafavī's grandson, a boy named Abū Ṭarāb, should be proclaimed Shāh Esmā'īl III; but 'Alī Mardān Khān broke the compact and was killed. Karīm Khān gained supremacy over central and southern Iran and reigned as regent or deputy (*vakīl*) on behalf of the powerless Ṣafavid prince, never arrogating to himself the title of shah. Karīm Khān made Shīrāz his capital and did not contend with Shāh Rokh (1748–95) for the hegemony of Khorāsān. He concentrated on Fārs and the centre, but managed to contain the Qājārs in Mazanderan, north of the Elburz Mountains. He kept Āghā Moḥammad Khān Qājār a hostage at his court in Shīrāz, after repulsing Moḥammad Ḥasan Qājār's bids for extended dominion. (P.W.A.)

Prosperity under Karīm Khān

Karīm Khān's geniality and common sense inaugurated a period of peace and popular contentment, and he strove for commercial prosperity in Shīrāz, a centre accessible to the Persian Gulf ports and the Indian trade. After Karīm Khān's death in 1779, Āghā Moḥammad escaped to the Qājār tribal country in the north, gathered a large force, and embarked upon a war of conquest.

The Qājārs (1779–1925). *Aghā Moḥammad Khān.* Between 1779 and 1789 the Zands fought among themselves over their legacy. In the end, it fell to the gallant Loṭf 'Alī, the Zands' last hope. Āghā Moḥammad Khān relentlessly hunted him down until he overcame and killed him at the southeastern city of Kermān in 1794. In 1796 Āghā Moḥammad assumed the imperial diadem, and later in the same year he took Mashhad. Shāh Rokh died of the tortures inflicted on him to make him reveal the complete tally of the Afshārids' treasure. Āghā Moḥammad was cruel and he was avaricious. Karīm Khān's commercial efforts were nullified by his successors' quarrels. With cruel irony, attempts to revive the Persian Gulf trade were followed by a British mission from India in 1800, which ultimately opened the way for a drain of Persian bullion to India. This drain was made inevitable by the damage done to Iran's productive capacity during Āghā Moḥammad's campaigns to conquer the country.

European penetration. Fatḥ 'Alī Shāh (ruled 1797–1834), in need of revenue, relied on British subsidies but lost the Caucasus to Russia by the treaties of Golestān in 1813 and Turkmanchay in 1828. The last gave Russian commercial and consular agents entrance to Iran, and this began a diplomatic rivalry between Russia and Britain that victimized Iran. This rivalry was eventually resolved when in 1907 an Anglo-Russian convention established, in Iran, Afghanistan, and Tibet, exclusive Anglo-Russian spheres of influence.

Anglo-Russian rivalry

Hājjī Mīrzā Āghāsī, a minister of Moḥammad Shāh (ruled 1834–48), tried to activate the government in a revival of the sources of production and to cement ties with lesser European powers, such as Spain and Belgium, as an alternative to Anglo-Russian dominance, but little was achieved. Nāṣer od-Dīn Shāh (ruled 1848–96) made Iran's last effort to regain Herāt, but 1857 saw the end of such expansionist efforts. Popular and religious antagonism to the Qājār regime increased as Nāṣer od-Dīn Shāh strove to raise funds by selling concessions, ostensibly for the development of his country's resources, to Europeans. The money paid for concessions was squandered by the court and on the Shah's journeys to Europe.

Popular protest and the constitution. The effect of popular protest in bringing about cancellation of a tobacco concession in 1890 demonstrated two factors of crucial significance for the future: first, that unified popular protest could limit despotism's scope; and, second, that a mercantile class of sufficient prosperity existed in Iran to make use, with support from religious orators, of popular feeling. The Shah's suppression of the Bābī and Bahā'ī heterodoxies had not, in spite of its ugly severity, ingratiated the regime with the orthodox *'ulamā'*.

The "Tobacco Riots" were the prelude to the constitutional revolution of the reign of Moẓaffar od-Dīn Shāh (ruled 1896–1907). Iran remained on silver after the failure of bimetallism and the world slump in silver values from the 1870s onward. Silver bullion drained out of the country and copper money proliferated, causing bread riots. In 1906 the ailing Shah granted a constitution. In October of that year the first National Consultative Assembly (the Majles) was opened. In 1908, under Moḥammad 'Alī Shāh (ruled 1907–09), the Majles was suppressed with the aid of the Persian Cossack Brigade. The Majles was revived after a civil war that culminated in Moḥammad 'Alī's deposition.

Constitution of 1906

Rise of Reza Khan. During World War I, Iran was the scene of rival intrigues by pro-British and pro-German groups among the notables, a class that had succeeded in gaining control of the Majles. The disruption was exacerbated by famine and national bankruptcy.

In 1919 the Majles refused a British offer of financial and military assistance, and British financial and military experts were withdrawn from the country. Salvation came from another quarter, in the person of an Iranian officer, Reza Khan, of the Persian Cossack Brigade. In collaboration with a political writer, Sayyid Zia od-Din Tabataba'i, he staged a coup d'etat in 1921 and took control of all the military forces. Reza Khan's efforts between 1921 and 1925 as, successively, war minister and prime minister under Aḥmad Shāh resulted in the formation of an army loyal to him, the achievement of order, and finally, in 1925, the deposition of the last Qājār shah and the transference of sovereignty to himself.

The Pahlavi Dynasty (1925–79). *Reza Shah.* During the reign of Reza Shah Pahlavi (Reza Khan's imperial style), educational and juridical reforms were effected that lessened the influence of the religious classes and laid the basis of a modern state. Women were freed from the veil and divorce laws were modified in their favour. The nation's independence from foreign political interference was restored. In 1933 improved terms were gained for Iran on the oil concession granted to a British company in 1901. Iran's first railroad was inaugurated in 1938. Unfortunately, trade necessities, fear of the Soviet grip on the routing of Iranian goods to Europe, and fear of British influence in the south and of Soviet influence in the north made Reza Shah turn to Nazi Germany. His refusal to abandon what he conceived to be obligations toward numerous Germans in Iran in 1941 occasioned an Anglo-Soviet invasion of the country, to ensure the safe passage of American supplies to the Soviet front through Iran. In September 1941 Reza Shah abdicated and left Iran.

Mohammad Reza Shah. In wartime conditions, Reza Shah's son, Mohammad Reza Shah Pahlavi, succeeded to the throne. The Soviet Union and a Communist "popular" regime were removed from Azerbaijan (then the Northwestern Province of Iran) in 1946. In 1951 the Majles passed an act, introduced by Mohammad Mosaddeq, nationalizing Iranian oil; the British Oil Company withdrew. The disturbed political situation during Mosaddeq's premiership turned his nationalization triumph into a Pyrrhic victory. His period in office ended in turmoil in 1953.

By 1961 the Shah was able finally to take the initiative. Dissolving the 20th Majles in May of that year, he cleared the way for the first Land Reform Law, enacted in 1962. The landed minority had to give up its lands to the government for redistribution to cultivators. (Among those stripped of land was the Shī'ah Muslim religious establishment.) Profit sharing in industry was introduced, and the former landlords could receive compensation for their holdings in the form of shares in industries. Cultivators and workers were given more voice in national affairs, and cooperatives in rural areas began to replace the former landowners as sources of capital for irrigation and agrarian maintenance and development. A campaign was organized to reduce illiteracy, and education was further removed from the control of the clergy. The country's power structure was radically changed in a program termed the "White Revolution."

The White Revolution

On January 26, 1963, the White Revolution was overwhelmingly endorsed by the nation. By 1971, when land distribution ended, about 2,500,000 families, comprising a farming population of more than 12,000,000, were estimated to have benefitted from the reforms. During 1960–72 the percentage of owner-occupied farmland in Iran rose from 26 to 78 percent.

Per capita income rose from about $176 in 1960 to about $2,500 in 1978. During 1970–77 the gross national product was reported to have increased to an average annual

rate of 7.8 percent. Oil revenues financed the extraordinary growth of investment under Iran's successive development plans, rising from $68,455,000 under the 1949–56 plan to $9,802,300,000 (50 percent from oil) under the Fourth Plan of 1968–73 and to some $68,600,000,000 (more than two-thirds from oil) under the Fifth Plan of 1973–78.

Improvement and construction after 1961 were accompanied by an "independent national policy" in foreign relations. The main principles of this foreign policy were support for the United Nations and peaceful coexistence, with a positive approach in cementing friendly and mutually beneficial ties with other nations. Relations with the United States remained close. Iran played a major role in the Central Treaty Organization (Cento) and Regional Cooperation for Development (RCD) with Turkey and Pakistan. It also embarked upon trade and cultural relations with eastern Europe, the Soviet Union, France, Germany, and Scandinavia.

Cento and RCD

Iranian oil development was accelerated after 1954 by a multinational Western consortium led by British Petroleum. The National Iranian Oil Company (NIOC) embarked on all-around expansion. It formed a petrochemical subsidiary and concluded agreements, mainly on the basis of equal shares, with several international companies for exploitation of oil outside the area of the consortium.

National Iranian Oil Company

In July 1973 a new 20-year oil agreement with the consortium of Western firms was concluded. It had the effect of initiating a plain seller-buyer relationship between Iran and the oil companies and of giving Iran control over field operations. These operations, along with all facilities and installations previously worked by the consortium, were vested in NIOC. The consortium agreement was cancelled in 1979, leaving NIOC in control of the Iranian oil industry.

World monetary instability and fluctuations in Western oil consumption dangerously threatened an economy rapidly expanded after 1954 and directed on an unprecedented scale to high-cost development programs. Acutely aware of the danger of dependence on a diminishing asset, the Shah pursued a policy of diversification. Assembly of motor vehicles started in 1957, and in the early 1970s Iran became an exporter of motor vehicles to Egypt and Yugoslavia. Copper reserves were exploited, and in early 1972 Iran's first steel mill began producing structural steel. Iran purchased a major share in the Krupp firm of West Germany in 1974 and continued to press for barter agreements for the marketing of its oil and gas.

Population continued to rise, and the flight from rural areas to towns, particularly to the capital and other northern centres, presented a serious problem. Employment was artificially increased by loans and credits, while businesses were obliged to offer 49 percent of their stock for sale to workers. In 1975 a government-sponsored war on high prices resulted in arrests and fines of traders and manufacturers, injuring the market's confidence.

The Islāmic republic. *Revolution of 1978–79.* The sense that in both the agricultural and industrial spheres too much had been attempted too rapidly and that mistakes had been made and expectations disappointed was manifested in demonstrations against the government in 1978; many people were killed, and martial law was imposed in the major cities in September. This ended the relaxation of government controls, begun in 1977, that had encouraged protests and that had led to the emergence of religious activists allied with extremist "Dedicated Fighter" groups, the Mujāhedīn; these groups were opposed to the influx of foreigners, particularly Americans, and to a westernization they saw as threatening to those traditional values subsumed under the cloak of Shī'ite Islām.

Ruhollah Khomeini, an Islāmic jurist accorded the honorific title of ayatollah ("sign of God"), coordinated an upsurge of opposition—first from Iraq, where he had gone into exile in 1964, and then from France, after 1978 demanding the Shah's abdication. On January 16, 1979, under constitutional procedures, the Shah left Iran, and he died in Egypt on July 27, 1980.

The Regency and Supreme Army Councils established for the Shah's absence proved unable to function, and Prime Minister Shahpur Bakhtiar was unable to effect compromise with his former National Front colleagues or with Khomeini. Crowds in excess of 1,000,000 demonstrated in Tehrān, proving the wide appeal of Khomeini, who arrived in Iran amid wild rejoicing on February 1, 1979. Ten days later Bakhtiar went into hiding, eventually to find exile in Paris.

The republic. On April 1, after a landslide victory in a national referendum, Khomeini declared an Islāmic republic, subsequently invested with a new constitution reflecting his ideals of Islāmic government. Fundamentalist measures followed, with rules for the comportment and veiling of women and the repeal of modern divorce laws. Efforts were made to suppress Western influence, and many of the Western-educated elite fled the country.

In November 1979 supporters of the revolution took control of the U.S. embassy in Tehrān, seized 66 U.S. citizens there and at the foreign ministry, and, with the exception of 14 who were granted early release, held them hostage until January 20, 1981. Also in November 1979, the republic's first prime minister, Mehdi Bazargan, resigned. The republic's first president, Abolhassan Bani-Sadr, opposed the holding of the U.S. embassy. He was forced to flee to France, together with opposition leader Massoud Rajavi of the outlawed Mujāhedīn-e Khalq (Fighters for the People) faction. The Mujāhedīn stepped up a campaign of sporadic and highly demoralizing bombing in Tehrān, including the bombing on June 28, 1981, of the headquarters of the ruling Islāmic Republican Party.

War with Iraq

In September 1980 the president of Iraq ordered the invasion of the southwestern oil-producing Iranian province of Khūzestān, but a supposedly weakened Iranian army achieved surprising defensive success. By summer 1982, Iraq's initial territorial gains had been recaptured by Iranian troops that were stiffened with Revolutionary Guards and, to a disturbing degree, with young village boys attracted by a promise of rewards after life. Prolongation of the war caused Arab and international anxiety, centred on the possible threat to the oil-producing countries on the Persian Gulf.

Factionalism and corruption pointed to increased disillusionment among Iranians. The numbers of extremely young casualties in the Iran–Iraq war affected public morale, but the government was careful to provide for the poor through mosque-based *komitehs* (committees). Imports of foodstuffs were still high and, coupled with war expenditures, created a drain on an economy already stricken by a declining world demand for oil and by severe war damage to oil installations. Unrelenting executions on sometimes trivial allegations, rumours of torture, persecution of Bahā'īs, arbitrary arrests, and bad prison conditions tarnished the reputation of the republic's leaders.

For later developments in the political history of Iran, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (P.W.A./Ed.)

BIBLIOGRAPHY

General. Prominent among histories that are also informative on contemporary Iranian cultures are JOSEPH M. UPTON, *The History of Modern Iran: An Interpretation* (1960, reprinted 1970); and DONALD N. WILBER, *Iran: Past and Present: From Monarchy to Islamic Republic,* 9th ed. (1981). CHARLES ISSAWI (ed.), *The Economic History of Iran, 1800–1914* (1971), contains documents, statistical data, and commentary on economic conditions prior to World War I; and TAREQ Y. ISMAEL, *Iraq and Iran: Roots of Conflict* (1982), is a well-documented history of Iran–Iraq relations, covering events from the conflict between the Ottoman and Persian empires over the territory of modern-day Iraq through the agreement reached by Iranian and Iraqi leaders in Algiers in 1975. ROBIN WRIGHT, *Sacred Rage: The Crusade of Modern Islam* (1985), recounts Iran's efforts to export its revolution to other Islāmic countries; see also NIKKI R. KEDDIE (ed.), *Religion and Politics in Iran: Shi'ism from Quietism to Revolution* (1983). JOHN I. CLARKE, *The Iranian City of Shiraz* (1963); and PAUL WARD ENGLISH, *City and Village in Iran: Settlement and Economy in the Kirman Basin* (1966), are two of the best books on modern Iranian cities. See also LUDWIG ADAMEC (ed.), *Historical Gazetteer of Iran* (1976–).

Land and people. W.B. FISHER (ed.), *The Cambridge History of Iran,* vol. 1, *The Land of Iran* (1968), is perhaps the single most comprehensive and informative work on geography and

social ecology, and Fisher's *Middle East: A Physical, Social, and Regional Geography,* 7th rev. ed. (1978), includes a brief survey of Iran. Also useful is the section on Iran in *The Middle East and North Africa* (annual), a country survey with up-to-date statistical data. Notable among the books on pastoral nomads are FREDRIK BARTH, *Nomads of South-Persia* (1961, reissued 1965); and MERIAN C. COOPER, *Grass* (1925), a dramatic eyewitness account of the spring migration of a Bakhtyārī group over the Zagros Mountains in 1924.

Economic conditions. M. FROOZĀN, M. SHIRĀZI, and I. EBTEHĀJ-SAMI'I, "The Development of the Gas Industry in Iran," *Tahqiqāt-e Eqtesādi: Quarterly Journal of Economic Research,* 7(19–20):25–47 (Summer and Autumn 1970), the text of a paper delivered to the 11th International Gas Congress; ANN K.S. LAMBTON, *The Persian Land Reform, 1962–1966* (1969), a study based on research in the field; DANIEL CRAIG, "The Impact of Land Reform on an Iranian Village," *Middle East Journal,* 32(2):141–154 (Spring 1978), a discussion of the problems created by the monarchy's land reform; and NIKKI R. KEDDIE, "Stratification, Social Control, and Capitalism in Iranian Villages: Before and After Land Reform," in RICHARD ANTOUN and ILIYA HARIK (eds.), *Rural Politics and Social Change in the Middle East,* pp. 364–402 (1972), also a field research study.

Government and social conditions. WILLIAM GREEN MILLER, "Political Organization in Iran: From Dowreh to Political Party," *Middle East Journal,* 23(2):159–167 (Spring 1969), and 23(3):343–350 (Summer 1969); E.A. BAYNE, *Persian Kingship in Transition: Conversations with a Monarch Whose Office Is Traditional and Whose Goal Is Modernization* (1968), a discussion with Mohammad Reza Shah Pahlavi of the political process; RUHOLLAH KHOMEINI, *Islamic Government* (1979; originally published in Fārsī, 1979), Khomeini's concept of the nature and function of an Islāmic state; and HAMID ALGAR (trans.), *Constitution of the Islamic Republic of Iran* (1980; originally published in Fārsī, 1979). Three works on the modern family, fertility, mortality, and related matters are ALI A. PAYDARFAR and MAHMOOD SARRAM, "Differential Fertility and Socioeconomic Status of Shirazi Women: A Pilot Study," *Journal of Marriage and the Family,* 32(4):692–699 (1970); PAUL VIEILLE, "Birth and Death in an Islamic Society," *Diogenes,* 57:101–127 (Spring 1967); and *Some Demographic Aspects of the Population of Iran* (1968), a collection of articles written mostly by Iranian social scientists. Two books that systematically consider many aspects of Iranian culture in depth and in the context of change are NORMAN JACOBS, *The Sociology of Development: Iran as an Asian Caste Study* (1966); and EHSAN YAR-SHATER (ed.), *Iran Faces the Seventies* (1971). A. REZA ARASTEH, *Education and Social Awakening in Iran, 1850–1968,* 2nd rev. ed. (1969), is a critical study of Iranian education.

Cultural life. S.H. NASR, "Ithnā 'Ashariyya," in *The Encyclopaedia of Islam,* new ed., vol. 4, pp. 277–279 (1978), an article in English on the beliefs of the Twelvers; 'ALLĀMAH SAYYID MUḤAMMAD ḤUSAYN ṬABĀṬABĀ'Ī, *Shī'ite Islam,* trans. from Fārsī, 2nd ed. (1977), an authoritative monograph on the origins and growth of Shī'ism; FAZLUR RAHMAN, *Major Themes of the Qur'ān* (1980); PETER J. CHELKOWSKI (ed.), *Ta'ziyeh: Ritual and Drama in Iran* (1979), a collection of essays by both Iranian and Western scholars; ARTHUR UPHAM POPE, *Masterpieces of Persian Art* (1945, reissued 1970); and HANS E. WULFF, *The Traditional Crafts of Persia: Their Development, Technology, and Influence on Eastern and Western Civilizations* (1966), a well-documented study with illustrations.

History. *Iran to the end of the Achaemenid period:* RICHARD N. FRYE, *The Heritage of Persia,* 2nd ed. (1976); ROMAN GHIRSHMAN, *Iran from the Earliest Times to the Islamic Conquest* (1954), a useful introduction; and GEOFFREY HANDLEY-TAYLOR, *Bibliography of Iran: Coronation Edition,* rev. ed. (1967). (*Prehistory*): R.H. DYSON, JR., "Problems in the Relative Chronology of Iran, 6000–2000 B.C.," in ROBERT W. EHRICH (ed.), *Chronologies in Old World Archaeology* (1965), an authoritative summary of prehistoric Iranian chronology; "The Archaeological Evidence of the Second Millennium B.C. on the Persian Plateau," *Cambridge Ancient History,* 3rd ed., vol. 2, part 1 (1973); and L. VANDEN BERGHE, *Archéologie de l'Iran ancien* (1959), a fairly complete survey by province and period of Iranian archaeology for all periods to the Sāsānian, with a good bibliography. (*Elam*): PIERRE AMIET, *Elam* (1966), a survey of Elamite art, containing good pictures and a good bibliography; GEORGE G. CAMERON, *History of Early Iran* (1936, reprinted 1976), a good general survey; WALTER HINZ, "Persia *c.* 2400–1800 B.C.," *Cambridge Ancient History,* rev. ed., fascicle 19 (1963); "Persia *c.* 1800–1550 B.C.," *Cambridge Ancient History,* 3rd ed., vol. 2, part 1 (1973); and RENÉ LABAT, "Elam *c.* 1600–1200 B.C.," *ibid.,* vol. 2, part 2 (1975), authoritative reviews. (*Protohistory*): R. LABAT, "Elam and Western Persia *c.* 1200–1000 B.C.," *ibid.*, a fine survey of Elamite history in the period; T. CUYLER YOUNG, JR., "A Comparative Ceramic Chronology for Western Iran, 1500–500 B.C.," *Iran,* vol. 3 (1965), an attempt to combine archaeological and written sources in a historical reconstruction. (*The Achaemenids*): ANDREW R. BURN, *Persia and the Greeks: The Defense of the West, c. 546–478 BC* (1962), a fairly balanced view of the Achaemenids in their dealings with the Greeks; *Cambridge Ancient History,* vol. 4, *The Persian Empire and the West* (1939), a standard summary of events but in some respects out-of-date; ALBERT T. OLMSTEAD, *History of the Persian Empire* (1948, reissued 1978), a good introduction.

The Parthian period: The publications of M.E. MASSON and of G.A. PUGACHENKOVA are indispensable for an acquaintance with this period, as are the works of G.A. KOCHELENKO. A detailed Russian bibliography may be found in GREGOIRE FRUMKIN, *Archaeology in Soviet Central Asia* (1970). See also FRANZ ALTHEIM, *Weltgeschichte Asiens im griechischen Zeitalter,* 2 vol. (1947–48), a brilliant study of Hellenism in the Orient, with particular attention to the role of the Iranians; FRANZ ALTHEIM and RUTH STIEHL, *Geschichte Mittelasiens im Altertum* (1970); N. DEBEVOISE, *A Political History of Parthia* (1938, reprinted 1969); JOSEF DOBIAS, "Les Premiers Rapports des Romains avec les Parthes et l'occupation de la Syrie," *Archiv Orientální,* 3:215–256 (1931), well-documented research; ALFRED VON GUTSCHMID, *Geschichte Irans und seiner Nachbarländer von Alexander dem Grossen bis zum Untergang der Arsaciden* (1888); and GEORGE RAWLINSON, *The Sixth Great Oriental Monarchy; or, The Geography, History and Antiquities of Parthia* (1873), both old but still quite valuable; MIKHAIL ROSTOVTSEV, "The Sarmatae and Parthians," *Cambridge Ancient History,* vol. 11 (1936), of capital importance for Parthian history; WERNER SCHUR, "Parthia," *Pauly-Wissowa Real-Encyclopädie,* vol. 36, col. 1987–2029 (1949); W.W. TARN, "Parthia," *Cambridge Ancient History,* vol. 9 (1932), and *The Greeks in Bactria and India,* 2nd ed. (1951), both works based on vast expertise, although certain details are still not universally accepted; JOSEF WOLSKI, "The Decay of the Iranian Empire of the Seleucids and the Chronology of the Parthian Beginnings," *Berytus,* 12:35–52 (1956–57); K.H. ZIEGLER, *Die Beziehungen zwischen Rom und dem Partherreich* (1964), a valuable work with a bias toward Rome; and G.L. RIDER, *Suse sous les Séleucides et les Parthes* (1965), providing an important listing of coins.

The Sāsānian period: (*History*): THEODOR NOLDEKE, *Geschichte der Perser und Araber zur Zeit der Sasaniden aus der arabischen Chronik des Ṭabarī* (1879), German trans. of at-Ṭabarī's (9th–10th century) history, with commentary, still essential; ARTHUR CHRISTENSEN, *L'Iran sous les Sassanides,* 2nd ed. rev. (1944), the only comprehensive history; FRANZ ALTHEIM and RUTH STIEHL, *Feudalismus unter den Sasaniden und ihren Nachbarn* (1954), and *Finanzgeschichte der Spätantike,* pp. 5–114 (1957); N. PIGULEVSKAJA, *Les Villes de l'état iranien aux époques parthe et sassanide* (1963; originally published in Russian, 1956); JEAN GAGE, *La Montée des Sassanides et l'heure de Palmyre* (1964), with translations from several original sources; GEO WIDENGREN, *Der Feudalismus im alten Iran* (1969). (*Chronology*): KURT ERDMANN, "Die Entwicklung der sasanidischen Krone," *Ars Islamica,* 15:87–123 (1951); ROBERT GOBL, *Sasanidische Numismatik* (1968), tables 1–14. (*Inscriptional materials*): ERNST HERZFELD, *Paikuli: Monument and Inscription of the Early History of the Sassanian Empire,* 2 vol. (1924), the first, now somewhat superseded, interpretation of the inscription of Narses; and MARTIN SPRENGLING, *Third Century Iran: Sapor and Kartir* (1953), the first interpretation of inscriptions of Shāpūr I and Kartēr.

Iran since c. AD 600: SIR PERCY M. SYKES, *A History of Persia,* 3rd ed., 2 vol. (1969), a general history; ALESSANDRO BAUSANI, *The Persians* (1971, originally published in Italian, 1962); E. DENISON ROSS, *The Persians* (1931); WILHELM BARTHOLD, *Turkestan Down to the Mongol Invasion,* 4th ed. (1977; originally published in Russian, 1898–1900), the essential survey of northeastern Iranian history from *c.* AD 600 to the 13th century; later research is represented by MUHAMMAD A. SHABAN, *The 'Ābbasid Revolution* (1970). The *Cambridge History of Iran,* vol. 5 (1968), covers the 12th–14th centuries; MARTIN B. DICKSON, *Shāh Tahmāsb and the Úzbeks* (1958), on the history of the first half of the 16th century; LAURENCE LOCKHART, *Nadir Shah* (1938, reprinted 1973), and *The Fall of the Safavī Dynasty and the Afghan Occupation of Persia* (1958), deal with the period 1722–46; PETER W. AVERY, *Modern Iran* (1965), is a general work on the 19th and 20th centuries to 1963. See also JULIAN BHARIER, *Economic Development in Iran: 1900–1970* (1971); EDWARD G. BROWNE, *The Persian Revolution of 1905–1909* (1910, reprinted 1966); R.N. FRYE, *The Heritage of Persia,* 2nd rev. ed. (1976); and AMIN BANANI, *The Modernization of Iran: 1921–1941* (1961). L.P. ELWELL-SUTTON, *Modern Iran* (1941), covers the period 1924–41, and his *Guide to Iranian Area Study* (1952) has a useful chronological table

and bibliography. Dynastic tables and essays on different aspects of Iranian history and culture are in A.J. ARBERRY (ed.), *The Legacy of Persia* (1953, reissued 1968). Sound modern research has produced articles on Iran in *The Cambridge History of Islam*, 2 vol. (1970, reissued 1977). An essential reference work is *The Encyclopaedia of Islam*, 4 vol. and suppl. (1913–38; new ed., 1960–). VLADIMIR F. MINORSKY, *Iranica: Twenty Articles* (1964), *Taḏhkirat al-Mulūk* (1943), and *Persia in A.D. 1478–1490* (1957), are recommended, as are ANN K.S. LAMBTON, *Landlord and Peasant in Persia* (1953, reprinted 1969), and *The Persian Land Reform, 1962–1966* (1969); GEORGE LENCZOWSKI (ed.), *Iran Under the Pahlavis* (1978); ROUHOLLAH K. RAMAZANI, *The Foreign Policy of Iran: A Developing Nation in World Affairs, 1500–1941* (1966), and *Iran's Foreign Policy, 1941–1973* (1975); SHAHRAM CHUBIN and SEPEHR ZABIH, *The Foreign Relations of Iran* (1974), a prerevolutionary source; FIRUZ KAZEMZADEH, *Russia and Britain in Persia 1864–1914* (1968); MARVIN ZONIS, *The Political Elite of Iran* (1971); SEPEHR ZABIH, *The Communist Movement in Iran* (1966); ERVAND ABRAHAMIAN, *Iran Between Two Revolutions* (1982); SHAHROUGH AKHAVI, *Religion and Politics in Contemporary Iran: Clergy-State Relations in the Pahlavī Period* (1980); JAMES A. BILL, *The Politics of Iran: Groups, Classes and Modernization* (1972); ROBERT GRAHAM, *Iran: The Illusion of Power*, rev. ed. (1980); FRED HALLIDAY, *Iran: Dictatorship and Development* (1979); AMIN SAIKHAL, *The Rise and Fall of the Shah* (1980); MICHAEL M.J. FISCHER, *Iran: From Religious Dispute to Revolution* (1980); NIKKI R. KEDDIE, *Roots of Revolution: An Interpretive History of Modern Iran* (1981); MICHAEL A. LEDEEN and WILLIAM LEWIS, *Debacle: The American Failure in Iran* (1981); HOMA KATOUZIAN, *The Political Economy of Modern Iran* (1980); LOIS BECK and N.R. KEDDIE (eds.), *Women in the Muslim World* (1978); MICHAEL E. BONINE and N.R. KEDDIE (eds.), *Modern Iran: The Dialectics of Continuity and Change* (1981); and FEREIDUN FESHARAKI, *Development of the Iranian Oil Industry* (1976).

Iraq

Iraq (al-Jumhūrīyah al-ʻIrāqīyah) is an independent Arab country of the Middle East, situated at the northwest end of the Persian Gulf. It is bounded on the north by Turkey, on the east by Iran, on the southeast by the Persian Gulf, on the west by Syria and Jordan, and on the south by Kuwait and Saudi Arabia. It covers an area of 168,878 square miles (437,393 square kilometres). Baghdad is the national capital. Called Mesopotamia (the Land Between the [Two] Rivers) by the classical world, the country became known as Iraq in the 7th century.

Although basically an agricultural nation, Iraq is internationally important as a major source of the world's oil. Its considerable wealth of oil reserves has enabled the country to embark upon economic development and land-reform programs. It has, however, been subject to political instability since the overthrow of the monarchy in 1958. The army and the Baʻth Party have progressively increased their hold over the political and economic structures of the country. Iraq, part of the Arab world, has attempted to affirm its unique national culture, stressing its past and its own identity. (I.A.A.-L.)

The article is divided into the following sections:

Physical and human geography 900
  The land 900
    Relief and drainage
    Soils
    Climate
    Plant and animal life
    Settlement patterns
  The people 902
    Ethnic and religious characteristics
    Demography
  The economy 904
    Resources
    Agriculture, forestry, and irrigation
    Industry
    Finance and trade
    Transportation
  Administrative and social conditions 905
    Government
    Justice
    Armed forces
    Education
    Health and welfare
  Cultural life 906
History 906
  The origins of Mesopotamian history 906
    The background
    The emergence of Mesopotamian civilization
  Sumerian civilization 912
    The Sumerians to the end of the Early Dynastic period
    Sumer and Akkad in 2350–2000 BC
    The 3rd dynasty of Ur
  The Old Babylonian period 918
    Isin and Larsa
    Early history of Assyria
    The Old Babylonian Empire
    The Hurrians
  The Kassites, the Mitanni, and the rise of Assyria 922
    The Kassites in Babylonia
    Kingdom of the Hurrians and the Mitanni
    The rise of Assyria
  Assyria and Babylonia at the end of the 2nd millennium 924
    Babylonia under the 2nd dynasty of Isin (*c.* 1156–*c.* 1025)
    Assyria between 1200 and 1000
  Assyria and Babylonia *c.* 1000–*c.* 750 BC 925
    Assyria and Babylonia until Ashurnasirpal II
    Shalmaneser III and Shamshi-Adad V of Assyria
    Adad-nirari III and his successors
  The Neo-Assyrian Empire (746–609) 926
    Tiglath-pileser III and Shalmaneser V
    Sargon II (721–705) and Marduk-apal-iddina of Babylonia
    Sennacherib
    Esarhaddon
    Ashurbanipal (668–627) and Shamash-shum-ukin (668–648)
    Decline of the Assyrian Empire
  The Neo-Babylonian Empire (626–539) 930
    Nebuchadrezzar (Nabu-kudurri-usur) II
    The last kings of Babylonia
  Mesopotamian art and architecture 931
    Sumerian period
    Akkadian period
    Sumerian revival
    Assyrian period
    Neo-Babylonian period
  Assyro-Babylonian literature 938
  Mesopotamia under the Persians 938
  Babylonia under Alexander, the Seleucids, and the Parthians 939
  Iraq from the Arab conquest to 1918 939
    The Arab conquest
    The Umayyads
    The ʻAbbasids
    The Buyids
    The Seljuqs
    The Mongols
    The Ottomans
  Iraq since 1918 944
    British occupation and the mandatory regime
    Independence, 1932–39
    World War II and British intervention, 1939–45
    Postwar reconstruction and social upheavals, 1945–58
    The revolution of 1958
    Recurrence of military coups, 1963–68
    The Arab Baʻth Socialist regime since 1968

Physical and human geography

THE LAND

The Tigris and Euphrates

Relief and drainage. The fundamental elements in Iraq's physiography are the twin valleys of the Tigris and Euphrates rivers. Joined in the south but diverging in the north, they are separated by a tongue of higher land stretching southward from the Anatolian foothills to just north of Baghdad and known as al-Jazīrah (the Island).

Lower Iraq. The region begins at the ridge between ar-Ramādī and Baghdad and extends about 330 miles (530 kilometres) south-eastward to the Persian Gulf. The Tigris and Euphrates diverge east and west, respectively, below Baghdad, which is therefore in a narrower waist of lowland. Here the Tigris has a particularly tortuous course, and the river is marked by large numbers of meander channels and oxbow lakes. Because of the natural levees (raised embankments on either side of the river) and the difficult entry for tributary streams, there is much marsh. Many subsidiary rivers empty into the Hawr ("lake" or "marsh") as-Suwayqīyah depression, and navigation of the mainstream is difficult because of sharp bends and strong current. Below al-Kūt the course of the Tigris is much straighter and flatter. In its lowest section, especially from al-'Amārah southward, much of the waters are dispersed into distributary channels and marshes.

From ar-Ramādī southward the Euphrates flows at first in a well-defined channel. Natural levees occur, as on the Tigris. Between al-Musayyib and as-Samāwah the river discharges into various distributaries; the remnants of a previous swamp include old channels, levees, cutoffs, and lakes. From as-Samāwah there is a well-defined channel as far as an-Nāṣirīyah, or even Sūq ash-Shuyūkh, lying within an open plain that is now largely arid and almost featureless. Below an-Nāṣirīyah distributary channels take off the whole of the Euphrates water, which ultimately drains into Hawr al-Ḥammār, a vast expanse of reedy swamp. From al-Qurnah, the point of junction between the waters of Hawr al-Ḥammār and the Tigris, the river is known as the Shatt al-Arab. Away from the rivers, patches of swamp and marsh are separated by ridges.

Upper Iraq. This region comprises the valleys of the two main rivers, together with the somewhat higher ground that stretches southeastward between them. At its entry into Iraq, the Euphrates has cut a broad, flat valley with steep sides. Rejuvenation of the Euphrates led to many meanders being incised into the flat floor, giving a second set of terraces. Irregularities in rock strata have produced rapids and shoals, and the river tends to hug the western bank, producing a marked escarpment on the west and more open country to the east. The Tigris on leaving Turkey has cut an irregular winding passage among low foothills and ridges. Where the river flows southeastward, it tends to have a wide, shallow bed, with steep-sided low hills on the western (right) bank and open rolling country to the east. Separating these reaches are defiles that run from north to south, the narrowest of which is at Bāghlo-ja, about 14 miles below the confluence with the Nahr al-Khābūr on the Turkish–Iraqi frontier. Here the Tigris is only 120 feet wide, with a series of rapids that are almost cascades. Such conditions continue as far as Mosul, after which relief becomes much less accentuated, ultimately falling away to open steppe through which the Tigris flows in a broader channel. There the Tigris is joined by the Great and Little Zab rivers. One last, isolated defile occurs where the river breaks through the Jabal Ḥamrīn (Ḥamrīn Mountains) above Ba'ījī at a gorge that is now used to carry the Iraqi oil pipelines from Kirkūk to the Mediterranean. From Tikrīt southward the Tigris wanders on a flat plain, breaking up into several arms.

MAP INDEX

Political subdivisions

al-Anbār 33·45n 42·40e
al-Baṣrah 30·30n 47·47e
al-Muthannā 30·30n 45·00e
al-Qādisiyah 31·35n 45·00e
an-Najaf 31·59n 44·20e
as-Sulaymānīyah 35·40n 45·30e
at-Ta'mīm 31·35n 44·35e
Bābil 32·40n 44·35e
Baghdād 34·00n 43·16e
Dahūk 37·00n 43·15e
Dhī Qār 31·00n 46·00e
Diyālā 34·00n 45·00e
Irbīl 36·10n 44·00e
Karbalā' 32·15n 43·07e
Maysān 32·00n 47·00e
Nīnawā 36·10n 42·35e
Ṣalāh ad-Dīn 34·30n 43·49e
Wāsiṭ 32·45n 45·25e

The names of the political subdivisions do not appear on the map when they are the same as the names of their capital cities.

Cities and towns

Abū al-Khaṣīb 30·27n 47·59e
ad-Dīwānīyah 31·59n 44·56e
'Afak 32·04n 45·15e
Ain Ghazal 36·14n 41·59e
al-'Amādīyah 37·06n 43·29e
al-'Amārah 31·50n 47·09e
al-Bādī 35·56n 41·32e
al-Baṣra, see Basra
al-Baṭḥā' 31·06n 45·53e
al-Fallūjah 33·20n 43·46e
al-Fāw 29·58n 48·29e
al-Ḥadīthah 34·07n 42·23e
al-Ḥaḍr 35·35n 42·44e
al-Ḥalfāyah 31·49n 47·26e
al-Ḥayy 32·10n 46·03e
al-Ḥillah 32·29n 44·25e
al-Hindīyah 32·33n 44·13e
'Ali al-Gharbī 32·27n 46·41e
al-Kāzimīyah 33·22n 44·20e
al-Khāliṣ 33·49n 00·00e
al-Kifl 32·14n 44·22e
al-Kūt 32·25n 45·59e
al-Ma 'ānīyah 30·44n 43·00e
al-Maḥmūdīyah 33·03n 44·21e
al-Mawṣil, see Mosul
al-Miqdādīyah 33·59n 44·56e
al-Musayyib 32·47n 44·18e
al-Qā'im 34·21n 41·07e
al-Qurnah 31·00n 47·26e
Āltūn Kūprī 35·45n 44·09e
al-'Uzayr 31·19n 47·25e
Amarah, see al-'Amārah
'Ānah 34·28n 41·56e
an-Najaf 31·59n 44·20e
an-Nāṣirīyah 31·02n 46·16e
an-Nu'mānīyah 32·32n 45·25e
'Aqrah 36·45n 43·54e
ar-Ramādī 33·25n 43·17e
ar-Rumaythah 31·32n 45·12e
ar-Rustamīyah 33·17n 44·31e
ar-Ruṭbah 33·02n 40·17e
ash-Shabakah 30·49n 43·39e
ash-Shāmīyah 31·57n 44·36e
ash-Sharqāṭ 35·27n 43·16e
ash-Shaṭrah 31·25n 46·10e
ash-Shināfīyah 31·35n 44·39e
ash-Shu'aybah 30·25n 47·38e
as-Sa'dīyah 34·11n 45·07e
as-Salmān 30·30n 44·32e
as-Samāwah 31·18n 45·17e
as-Sulaymānīyah 35·33n 45·26e
'Ayn Zālah 36·43n 42·33e
az-Zubayr 30·23n 47·43e
Badrah 33·06n 45·58e
Baghdad 33·21n 44·25e
Ba'ījī 34·56n 43·29e
Balad 34·01n 44·09e
Ba'qūbah 33·45n 44·38e
Baṣra 30·30n 47·47e
Dahūk (Dhok) 36·52n 43·00e
Dokan, see Dukān
Dukān (Dokan) 35·56n 44·58e
Ḥalabjah 35·10n 45·59e
Hilla, see al-Ḥillah
Hīt 33·38n 42·49e
Irbīl 36·11n 44·01e
Jadīdah 34·01n 42·28e
Jalībah 30·35n 46·32e
Jalūlā' 34·16n 45·10e
Jāsimīyah 33·45n 44·41e
Karbalā' 32·36n 44·02e
Khān al-Baghdādī 33·51n 42·33e
Khānaqīn 34·21n 45·22e
Kibāsī 30·34n 47·50e
Kifrī 34·42n 44·58e
Kirkūk 35·28n 44·28e
Kūt al-Imārah, see al-Kūt
Kūysanjaq 36·05n 44·38e
Mandalī 33·45n 45·32e
Mosul (al-Mawṣil) 36·10n 42·35e
Naft Khāneh 34·02n 45·28e
Nasiriya, see an-Nāṣirīyah
Nukhayb 32·02n 42·15e
Qal'at Ṣāliḥ 31·31n 47·16e
Qal'at Sukkar 31·51n 46·05e
Qayyārah 35·48n 43·17e
Ramadi, see ar-Ramādī
Rāniyah 36·15n 44·53e
Rāwah 34·28n 41·55e
Rawāndūz 36·37n 44·31e
Rumaylah 30·47n 47·37e
Ṣafwān 30·07n 47·43e
Sāmarrā' 34·12n 43·52e
Shaykh Sa'd 32·34n 46·17e
Shithāthah 32·33n 43·29e
Sinjār 36·19n 41·52e
Sulaymaniyah, see as-Sulaymānīyah
Sūq ash-Shuyūkh 30·53n 46·28e
Tall 'Afar 36·22n 42·27e
Tall Kayf 36·29n 43·08e
Ṭaqṭaq 35·53n 44·35e
Ṭāwūq 35·08n 44·27e
Tikrīt 34·36n 43·42e
Tukayyid 29·47n 45·36e
Tūz Khurmātū 34·53n 44·38e
Zākhū 37·08n 42·41e

Physical features and points of interest

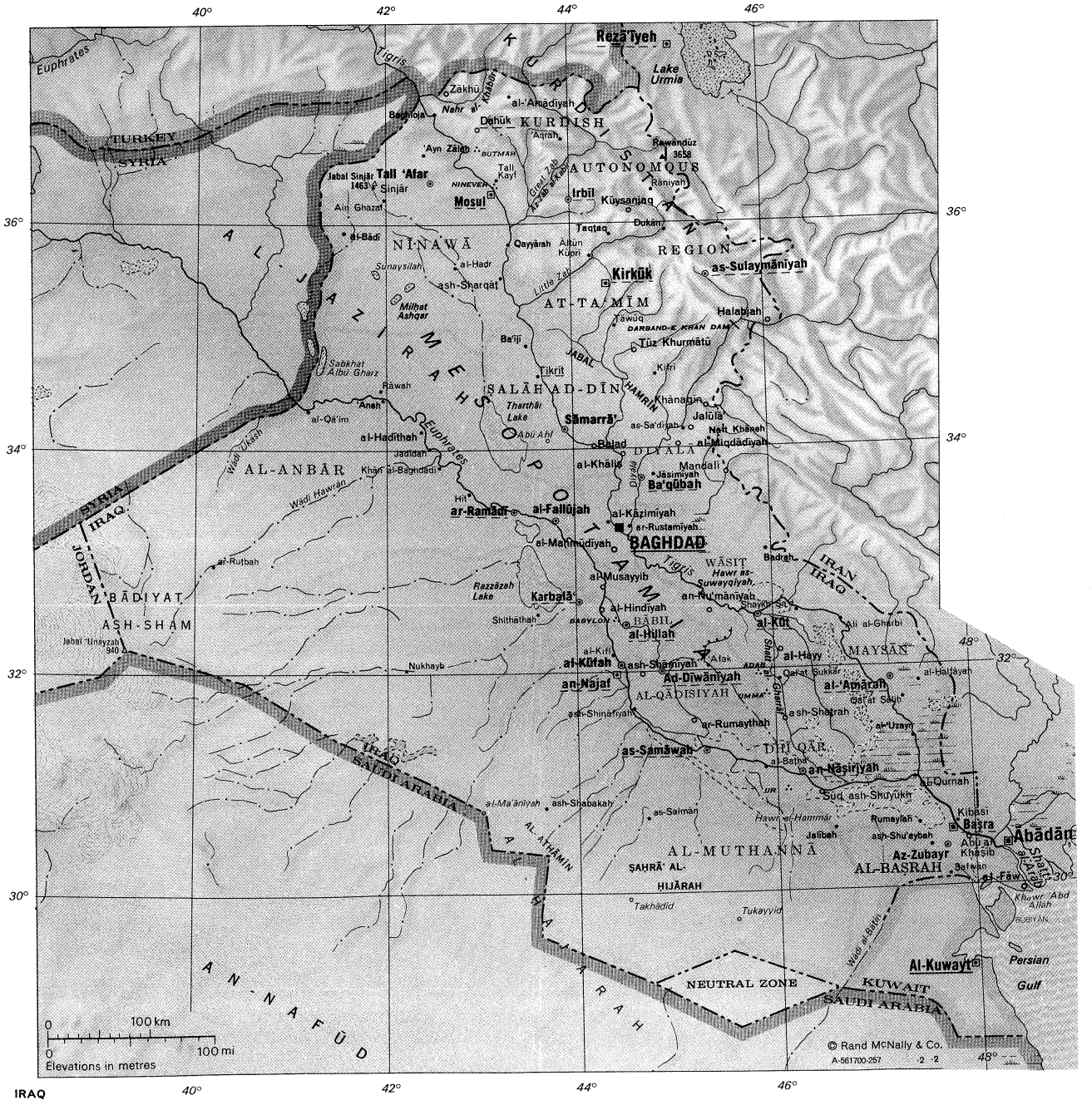
'Abd Allāh, Khawr, *channel* 29·50n 48·20e
Abū Ahl, *well* 34·06n 43·45e
Adab, *historic site* 31·59n 45·45e
al-Athāmīn, *mountains* 30·22n 43·40e
Ālbū Gharz, Sabkhat, *salt lake* 34·45n 41·15e
al-Furāt, see Euphrates
al-Ḥajarah, *desert* 30·00n 44·00e
al-Jazīrah, *historic region* 35·10n 42·00e
Arab, Shatt al-, *river* 31·00n 47·29e
az-Zāb al-Kabīr, see Great Zab
az-Zāb aṣ-Ṣaghīr, see Little Zab
Babylon, *ruins* 32·33n 44·24e
Bāṭin, Wādī al-, *watercourse* 29·35n 47·00e
Butmah, *oil deposit* 36·43n 42·42e
Darband-e Khan Dam 35·18n 45·47e
Dijlah, see Tigris
Diyālā, *river* 33·14n 44·31e
Euphrates (al-Furāt), *river* 31·00n 47·25e
Gharrāf, Shatt al-, *river* 32·30n 45·48e
Great Zab (az-Zāb al-Kabīr) *river* 36·00n 43·21e
Ḥabbānīyah, Hawr al-, *lake* 33·17n 43·29e
Ḥammār, Hawr al-, *lake* 30·50n 47·10e
Hamrīn, Jabal, *mountains* 34·30n 44·30e
Ḥawrān, Wādī, *watercourse* 33·58n 42·34e
Ḥijārah, Ṣaḥrā' al-, *desert* 30·25n 44·30e
Khābūr, Nahr al-, *river* 37·07n 42·21e
Kurdistan, *historic region* 37·00n 44·00e
Little Zab (az-Zāb aṣ-Ṣaghīr) *river* 35·12n 43·25e
Mesopotamia, *physical region* 34·00n 44·00e
Nineveh, *historic site* 36·25n 43·10e
Persian Gulf 29·00n 49·00e
Razzāzah, *lake* 33·45n 43·45e
Sham, Bādiyat ash-, *desert* 32·00n 40·00e
Sinjār, Jabal, *mountain* 36·23n 41·52e
Sunaysilah, *salt lake* 35·35n 41·53e
Suwayqīyah, Hawr as-, *lake* 32·40n 46·03e
Takhādīd, *well* 29·59n 44·30e
Tharthār, *lake* 34·00n 43·15e
Tigris (Dijlah), *river* 31·00n 47·25e
'Ukāsh, Wādī, *river* 34·15n 40·35e
Umma, *ruins* 31·40n 45·52e
Ur, *ruins* 30·57n 46·09e

Al-Jazīrah region

Between the Euphrates and Tigris and bounded on the north by the Jabal Sinjār lies al-Jazīrah, a region of undulating steppe, small-scale ranges, and closed drainage basins. The warping of the whole area has produced closed drainage basins, the largest of which is the Wādī ath-Tharthār (a long defile running more or less parallel to the Tigris about 20 to 40 miles west of the river), now part of the Tigris flood-control project.

The northeastern region. Fronting the Tigris on the east, between the Turkish frontier and the broad Diyālā Valley paralleling the Iran border, is an upland area roughly rectangular in shape that rises in steps eastward from river level. The first step is the Jabal Ḥamrīn, and behind it lies an undulating territory of river basins, rolling plateaus, and irregular hills that ultimately pass into the main Zagros range in Iran. The lower (western) part, broken by the valleys of the Great and Little Zab, is the ancient region of Assyria. Farther east, as the mountain zone begins, there is an alternation of high ridges aligned northwest to southeast and separated by river basins. This is often spoken of as Iraqi Kurdistan. Here, the mountain ridges have a summit average of about 8,000 feet (2,440 metres). They attain elevations of 10,000 to 11,000 feet in a few places, culminating in Rawāndūz (12,001 feet; 3,658 metres), the highest point in Iraq. Between the ridges are trough valleys usually occupied by rivers that have eroded out basins. All the streams weave a tortuous way south and southwest toward the lowlands and are marked by gorge and defile sections (*darband*) where they skirt or cut through ridges. In a few places, such as as-Sulaymānīyah, al-ʿAmādīyah, and Rawāndūz town, patches of lowland allow cultivation, but much of Kurdistan is given over to nomadism.

The western desert. Lying to the west of the riverine lowlands, this region extends from Kuwait and Saudi Arabia on the south as far as Jordan and Syria. In the region of the Jordan frontier is the Bādiyat ash-Shām (Syrian Desert). The desert area slopes gently upward from the Euphrates lowland and is formed of ancient rocks that are often exposed as irregular pavement. In the extreme south, part of the frontier with Kuwait is formed by Wādī al-Bāṭin, a rift structure three miles wide running inland from az-Zubayr as far as central Arabia. To the northwest lies Ṣaḥrāʾ al-Ḥijārah, in the main a region of shallow, mud-lined depressions, rocky outcrops, and loose sand.



IRAQ

An extensive swamp lies northeast of this area, paralleling the Euphrates.

Farther north lies the Wādīyah region, with steeper slopes. Because of its steepness, normal drainage developed, and there is a network of dry valleys aligned northeastward to the Euphrates that carry floodwater for an hour or two each year. The region takes its name from these valleys, or *wādīs*. Since there is no fringe of continuous sand or salt marsh, as there is farther south, the Wādīyah offers more practicable routes westward and is now used for oil pipelines.

Soils. The soils fall into two contrasted groups—heavy alluvial deposits associated with the main rivers and very light soils in the rest of the country. The alluvial soils contain a significant proportion of humus and clay and dry out to a hard, tenacious mass that can be used for building. The light soils are deficient or entirely lacking in humus and have a low clay content, but they may contain a proportion of wind-deposited minerals, some of which are valuable plant nutrients.

The problem of soil salinity

For both groups of soils, there is an acute problem of salinity induced by conditions of sharply seasonal rainfall followed by a hot, arid season. This produces evaporation of soil moisture and consequent accumulation of alkaline salts at or near the surface in quantities sufficient to inhibit cultivation. On entering Iraq, the Tigris and Euphrates contain about 30 parts per 100,000 of salts; in their lower courses this proportion trebles. It was estimated in 1949 that 60 percent of all irrigated land in Iraq had become salinated, and an overall assessment is that 1 percent of the total cultivated area is abandoned each year because of salinity.

Climate. Because of the simplicity and regularity of relief, Iraq has a straightforward climatic regime. In the lowlands two contrasting seasons occur: a dry and intensely hot summer from May to October and a relatively cool, humid winter from December to March, with short transitional periods. Altitude exerts a considerable effect in the mountain zones, where winters can be severe.

From May onward the predominating climatic element is a persistent and regular northwesterly wind (*shamāl*) over the whole of Iraq. Because these air currents come from land areas, they are dry and few clouds form, so that for several weeks or even months the sun beats down uninterruptedly, producing extremely high temperatures. At Baghdad, July and August daily temperature means are about 95° F (35° C) and 88° F (31° C), respectively; mean daily maxima are about 110° F (43° C). Some drop is apparent close to water surfaces because of high evaporation, but the effect is hardly improved because of greater relative humidity. Strong winds produce blowing dust or sandstorms; July is the worst month, with an average of five storms at Baghdad and eight at ash-Shu'aybah. Usually, no rain falls between May and the end of October.

Conditions are more variable during the winter season, when the *sharqī*—a cool, damp southeasterly wind—brings rain. At other times a cold air mass from the interior of Asia may extend westward, bringing fine but cold conditions. Hot air masses from the south may also affect Iraq; January temperatures of more than 81° F (27° C) are known, but frost may occur anywhere, even at Basra. Such extremes, however, are rare. The northeast has its own climatic conditions; as much as 30 to 40 inches (760 to 1,000 millimetres) of rain occur on the highest mountains, with 15 to 25 inches in the northeastern piedmont (region at the base of mountains).

Two climatic provinces

There are therefore two climatic provinces within Iraq; the hot, arid lowland and the damper northeast, where rainfall is sufficient for cultivation without irrigation, summer temperatures are slightly lower, and a colder, continental winter brings up to three months of snow in some places. The northeast and Kurdistan apart, the main features of Iraqi climate are aridity and overwhelming summer heat, the difficulties of which are intensified by high atmospheric humidity near the rivers.

Plant and animal life. Iraq is a meeting place of two major plant groups, both of which show the predominating influence of drought. Toward the north and east, there is a steppe vegetation of perennial bushes and low shrubs, with mugwort (*Artemisia*), goosefoot (Chenopodiaceae), and grass species dominant, together with a "flush" of short-lived creepers and grasses that amounts to about two-thirds of the total vegetation. Toward the south and west, vegetation is often restricted to thorns (chiefly tamarisk and species of *Haloxylon,* commonly called saxual), a few shrubs such as rhanterium, papposum, and other salt-resistant plants. Poplar, willow, licorice, and tamarisk flourish in small clumps near the rivers. Below al-Qurnah, the marshland is an area of reeds, tall grasses, and sedge. The open oak forest on the Zagros Mountains, at altitudes between 2,000 and 6,000 feet, has been mostly reduced to scrub by intensive cutting and unrestricted grazing. Above this region is an Alpine zone of cushion (clumped and compacted) plants and dwarf species broadly similar to those of Alpine Europe.

Animal life has been much reduced. The oryx (a kind of gazelle), ostrich, and wild ass are now practically extinct, and the last known lion was killed around 1910. Bats, rats of various species, jackals, hyenas, and wildcats are the most common mammals, with wild pigs and gazelles in remoter regions. Reptiles are numerous and include lizards, the largest reaching four feet; snakes, mainly "racer" species and sand snakes; and two types of tortoise, of which one is peculiar to Iraq. Many kinds of fish occur, including the "Tigris salmon," a barbel that can attain seven feet in length. Iraq is also visited by vast numbers of migratory birds.

Settlement patterns. Traditionally, the country may be divided into the valleys of the two large rivers, which are relatively densely populated, and the deserts of the west, where nomadism is important. In the northeast (including Kurdistan), where rainfall is sufficient for cultivation without irrigation, a pattern of mixed agriculture and herding is found; this region therefore contrasts sharply with the riverine and desert areas.

Traditional regions

Towns play an important part in Iraq. Baghdad, the capital, is by far the largest; other large cities are Basra, Mosul, Kirkūk, al-Ḥillah, Irbīl, Karbalā', and an-Najaf. The rest are spread fairly evenly throughout the riverine areas. Special features of the settlement pattern are the number and even spread of towns of relatively small size (from 10,000 to 25,000 inhabitants) and the extent to which the populations of even large towns are partly agricultural. There is much less of the sharp distinction between town and country that is characteristic of Western countries.

Nomadism is important not only in the western deserts but also between the two rivers, both in al-Jazīrah and the drier and sandier parts as far south as the Shaṭṭ al-Gharrāf. Nomads are also an important element in the towns, which they frequent in large numbers at certain seasons. Religious pilgrimages to Karbalā', an-Najaf, and Sāmarrā' swell the local population. In Kurdistan, seasonal migration mainly from winter (lowland) to summer (upland) pastures is widely practiced. Some Kurdish tribes cross the frontiers into Iran and Turkey. (W.B.Fi.)

Nomadism

THE PEOPLE

Ethnic and religious characteristics. *Cultural origins.* The population, while increasingly assuming greater homogeneity, still reflects a certain degree of ethnolinguistic and religious diversity. The ancient peoples of Iraq, the Babylonians and the Assyrians, who were Semitic, have long been assimilated and absorbed by successive waves of migration and settlement.

Arabs and their descendants, along with the Arabized population, are to be found throughout Iraq, and they constitute an overwhelming majority in the central and southern regions. They are estimated to constitute slightly more than three-quarters of the population.

Throughout the centuries, Iraq's contiguity with Iran has brought constant streams of Persian migrants who settled principally in the east. From 50,000 to 60,000 of these, however, were deported to Iran in late 1971 and early 1972. Migrations from Central Asia also brought successive waves of Mongols, Turks, Turkmens, Kurds, and Circassians. Although the Mongols lost their special character as a result of fusion with the dominant population, this was less true of the Kurds, Turks, and Turkmens.

The Kurdish minority. The Kurds, inhabiting the mountainous regions of the north and northeast, retain their separate communal and individual qualities, continue to speak their own language, and are the most visible and articulate of the ethnic minorities. After World War I, the Kurds were persuaded to live under the Iraqi political system. But neither the Arabs nor the Kurds were prepared to merge into one nationality and create a new Iraqi nation. The Arabs were divided into two unequal communities on religious grounds (Sunnīs and Shīʿīs). The merging of the Kurds—a solidly Sunnī community—with the Sunnī Arabs was considered by the Shīʿīs to weaken their position and was resisted. Moreover, the Sunnī Arabs themselves were reluctant to merge with the Kurds because of the upsurge of Arab nationalism, which, in turn, encouraged Kurdish nationalism.

Kurdish nationalism

The Kurds welcomed the revolution of 1958, which brought Abdul Karim Kassem (Qassim) to power, and formed a Kurdish Democratic Party (KDP), advocating Kurdish nationalism and Communism. The Kassem regime (1958–63), however, did little or nothing to improve the social and economic conditions of the Kurds whose leaders despaired of improvement under the new regime.

Mustafa al-Barzani, a Kurdish leader exiled under the monarchy, returned in 1958. He soon found Kassem instigating rival leaders against him, and in June 1961 a protracted war against Kassem broke out. Unable to defeat Barzani's forces, the successor Baʿth government agreed that a democratic regime would be established and that the Kurds would enjoy autonomy but not independence. Negotiations led to no agreement, however, even after Barzani's acceptance of administrative decentralization instead of autonomy.

After achieving power in 1968, the Baʿth government began to take practical steps to settle the queston. The Baʿth Party and the Kurdish Democratic Party (KDP) laid down proposals agreed upon by both and embodying essential Kurdish national demands. These proposals, formally accepted by the Baʿth Revolutionary Command Council (RCC) and endorsed by the Progressive National Front (PNF) were issued in the form of a manifesto on March 11, 1970. The principal points may be summed up as recogniton of the legitimate existence of the Kurdish nationality; recognition of the Kurdish language and approval of a university in as-Sulaymānīyah and a Kurdish academy of letters; a provincial law embodying the principle of decentralization of local administration and creation of a new province, Dahūk (Dʿhok); and a general amnesty for all civil and military personnel who had taken part in "violent acts."

When the terms of the March Manifesto were put to the test, however, the Baʿth and Barzani found themselves going in different directions. Two attempts on Barzani's life (1971 and 1972) and arrests and shootings on both sides created the impression that the March Manifesto was but a temporizing measure. On the fourth anniversary of the proclamation of the March Manifesto, the Iraqi government declared that it would enforce the law governing self-rule and invited the Kurds to cooperate in its enforcement. Barzani, however, warned the Kurds against collaboration with the Baʿth regime, and in 1974 the Iraqi Army moved slowly into the Kurdish areas. In 1975 the agreement reached directly between the Shah of Iran and Saddam Ḥusayn on the Shatt al-Arab dispute left the Kurdish forces at the mercy of the Iraqi Army. Iran agreed to encourage Barzani to accept a cease-fire (March 13, 1975), and Iraq declared that amnesty would be granted to all Kurds who would surrender by April 1. The Iraqi Army took complete control of the Kurdish area by the end of March, and Barzani surrendered his power to other Kurdish leaders.

Turks, Turkmens, and other groups. Turks and Turkmens inhabit central and northwest Iraq, particularly those areas contiguous to Turkey; these elements in the population resulted from a lengthy and massive migration, with the first settlements dating from the 9th century AD. The ultimate incorporation of Iraq into the Ottoman Empire in the 16th century resulted in additional Turkish and Turkmen settlements. The Muslim affiliation of these groups facilitated their intermixture with the Arabs and with other Iraqi groups, so ethnic distinctions lost much of their significance.

Other ethnic groups are quite small. They include Yazīdīs living west and north of Mosul, who speak a Kurdish dialect. Mandaeans are to be found in Baghdad, while the Lurs, a Persian-speaking group, are to be found chiefly along the Iraqi–Iranian frontier.

Linguistic groups. The predominant language of the entire population is Arabic, although Turkish and Kurdish are still used in certain areas. Kurdish in particular is widely spoken by the Kurds, although the educated are bilingual. Turkish is spoken only by members of the oldest generation. The Turkmens speak Turkmen, an Altaic language. Persian is widely spoken in eastern Iraq, particularly in areas adjacent to Iran; Persian speakers tend to be bilingual.

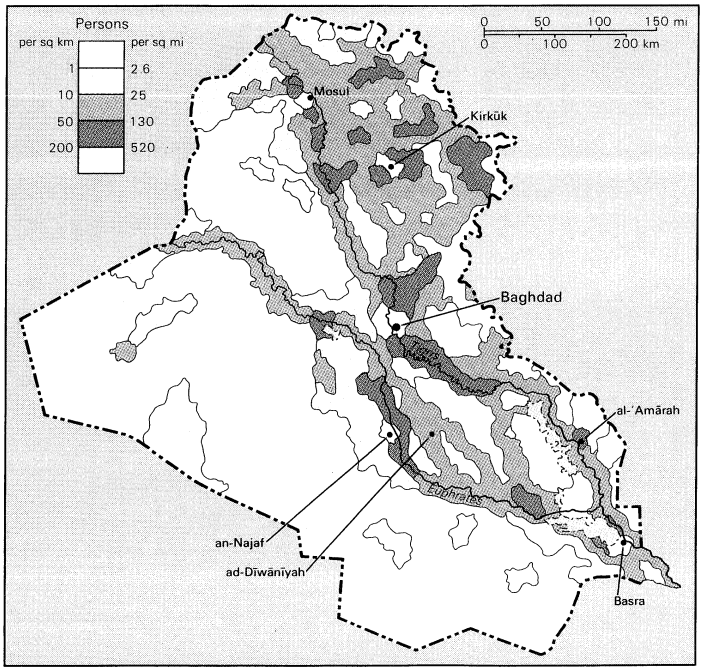
Religious affiliations. Religious variation cuts across ethnic and linguistic divisions. Iraq is predominantly an Islāmic country. Jewish and Christian communities, however, are still to be found. Some among the Jewish community date their origin from the time of the Babylonian Exile (586–516 BC). Throughout the early period of statehood, Jews constituted about 2 percent of the total population. With the emergence of Zionism and the eventual establishment of Israel in 1948, the major part of the Jewish community emigrated to Israel, the United States, Canada, and Europe.

Religious minorities

The Christian communities are chiefly descendants of the ancient population that was not converted to Islām when it took root in Iraq in the 7th century. They are subdivided among various Christian sects, including Chaldeans affiliated with the Catholic Church, Jacobites, Nestorian Christians (Assyrians), and members of the Eastern Orthodox Church. They are to be found throughout the country.

Muslims are divided almost evenly between the two major sects of Islām—Sunnī and Shīʿah. The Shīʿīs predominate in the east, although they are found throughout the country. Ethnically and linguistically, the Shīʿīs are Arabs, with a very small community of Turkmens, while the Sunnīs include Arabs, Turks, Turkmens, and Kurds. The religious schism has had important impact on the politics of the country, much in the same way that ethnic divisions have affected stability. The Iraqi regime has attempted to centralize, fuse, modernize, and secularize Iraqi society in order to reduce the importance of the schism.

Demography. The annual growth rate results almost entirely from natural increase. Infant mortality, while still



Population density in Iraq.

relatively high, is declining rapidly as a result of vigorous health programs implemented since the 1960s. Immigration and emigration are minimal. The few migrants entering the country are chiefly nomadic Bedouins from Saudi Arabia and the Syrian Desert. The only significant emigration in recent times was that of many Iraqi Jews to Israel.

Growth of the urban areas

As is true in other developing countries, there is considerable internal migration. Migrants are chiefly rural people attracted by better employment opportunities in the cities. As a result, Iraqi cities and towns have experienced considerable growth since the late 1950s.

Roughly three-fifths of the Iraqi population live in cities and towns, while the rest are considered rural. The ratio has altered in favour of the urban sector as a result of increasing industrial and commercial activities.

By conventional standards, and especially in comparison with other countries of the region, Iraq has a relatively low ratio of people to land. Its developing economy is therefore expected to meet the demands of its increasing population, and, in that sense, future prospects are encouraging.

THE ECONOMY

Iraq had an almost exclusively agricultural economy until expansion in manufacturing, service industries, and nonagricultural activities began in the 1950s. The successive governments have committed themselves to industrial growth. Increasing income from oil production has enabled the Iraqi government to devote a considerable percentage of the national budget to industrial development and to modernization of its agricultural base. Land-distribution reforms initiated in 1959 and amended by the 1970 Agricultural Act, cooperative rural associations, and the decline of the landed aristocracy contributed to improved agricultural productivity and to a more equitable distribution of income.

Resources. With the exception of oil, Iraq has modest mineral resources. They include iron ore, chromite, copper, lead, and zinc deposits in the north. Limestone, gypsum, phosphates, and sulfur are abundant. The most important single mineral, however, is oil, which is exploited in the north, northeast, and south. Other oil deposits are known to exist in central Iraq but are uneconomic to exploit.

Oil extraction

Although oil was first discovered at Naft Khāneh in 1909, the first commercial oil field was developed at Kirkūk between the Zagros Mountains and the Mesopotamian plain in 1927. It has remained the country's main source of oil. The second largest field is located at Zubayr-Rumaylah near Basra. The Mosul field became operative in 1952 and produces lower grade oil, while the small field at Naft Khāneh, near Mandalī, produces oil for domestic consumption. Fields have also been developed at 'Ayn Zālah and Buṭmah, north of Mosul, and elsewhere.

Oil from the Zubayr field is shipped by tanker through the Persian Gulf. Production from Kirkūk, however, is transported by overland pipeline to Tripoli, Lebanon, and Bāniyās, Syria. The inauguration in 1977 of the oil pipeline between Kirkūk and Dörtyol, Turkey, and the completion in 1976 of a pipeline from Kirkūk to Basra minimized Iraq's previous dependence on Syrian pipeline facilities.

Until 1961, the prospecting, extracting, and exporting of oil were monopolized by the Iraq Petroleum Company (IPC), or its affiliates; mostly British owned, it was finally nationalized on July 1, 1972. In 1967 and 1969 agreements were reached with French and Soviet companies. Further agreements were reached in 1972 with Italy, France, and the Soviet Union for the purchase of nationalized oil. The National Petroleum Authority is the supreme agency concerned with the oil industry.

The country has suffered from the absence of strict controls over its biological resources. Trees have been greatly exploited as a form of wealth and as sources of fuel. Afforestation projects have effected minimal recovery. Important wild plants include licorice, which grows along the main rivers, and the valonia oak, which grows in the northeastern mountains and yields a tanning extract. Gum tragacanth is derived from the *Astragalus* (loco) bush, which occurs in the dry steppe regions.

The Tigris and Euphrates rivers constitute Iraq's main sources of hydroelectric power. They are limited, however, by seasonal flow and the largely level valleys through which they flow. The Tigris tributaries that flow from the Zagros Mountains—the Little Zab and Diyālā rivers—have been developed for power generation, as has the Tigris. Thermal power resources are also limited; deforestation has led to the importation of wood, and coal resources are small. Oil for domestic use is supplied by a field near Mandalī and is refined at Khānaqīn.

Agriculture, forestry, and irrigation. Two types of agricultural activity predominate—the one dependent upon rainfall in the north; the other dependent upon irrigation in the south. The best known agricultural product, and the largest source of foreign-currency revenue after oil, is the date, of which Iraq is one of the world's largest producers. Dates are grown chiefly around the Shatt al-Arab area in the south. Other crops include barley, wheat, linseed, lentils, and beans, as well as rice, sesame, maize (corn), and millet. Cotton production is hampered by soil salinity.

The diversified landscape has proved hospitable to livestock in greater varieties and quantities than is possible in many other Arab countries. Sheep, goats, camels, cattle, and donkeys are abundant. Trade in animals is traditional, and milk, wool, meat, and other animal products are extensively used. Wool is exported to Europe, and animals are exported to neighbouring Arab countries. Fishing is important domestically.

Forestry is negligible as an industry; building timber must be imported.

The problems of irrigation are complex because flooding usually occurs before many crops have completed their growing season, and excess water must be retained within the river between high banks. Variation in water levels allows for free-flow irrigation in some areas, while in others pumping is necessary. Stagnant water must be led back to the rivers as rapidly as possible, in order to avoid salination.

The Wādī ath-Tharthār project

The largest river-control project is that of the Wādī ath-Tharthār, completed in 1956. A canal 40 miles long leads from the Tigris at Sāmarrā' to the basin of the *wādī*, which thus accumulates floodwater. The project is also being used for the generation of electricity. A smaller project is located on the Euphrates at ar-Ramādī, where water is led into two basins at Hawr al-Ḥabbānīyah and Hawr Abū Dibs. Other dams, primarily for irrigation, with flood control and electricity generation as subsidiary features, were planned for construction during the fourth quarter of the century.

The irrigation works at Hindīyah directs the Euphrates water into an older channel running through al-Ḥillah. Also on the Euphrates are the Musayyib project and smaller works at the head of Hawr al-Ḥammār. At al-Kūt, on the Tigris, a barrage that is three times the size of the Hindīyah dam will ultimately improve irrigation to the south and southeast, but there is a restricted volume of water available in summer. The entire low-season flow of the Diyālā River below the Jabal Ḥamrīn is diverted for irrigation. Other irrigation works exist throughout the country on a smaller scale. There are also land-drainage projects to counter salination and windbreaks to prevent the silting up of irrigation and drainage channels.

The industrial sector

Industry. Textile factories are becoming increasingly important for the Iraqi market; several were built, especially at Mosul, with assistance from the Soviet Union. Steel mills are in operation at Baghdad and Khawr az-Zubayr. Power plants, water-supply facilities, breweries, cigar factories, spinning and weaving factories, and brick and cement establishments are located in the Baghdad metropolitan area. A bitumen plant at Qayyārah, south of Mosul, and two cement plants produce a sufficiency of asphalt and cement, respectively, for the home market. Sugar plants around Mosul were constructed with the assistance of the Soviet Union, and two additional plants were constructed at Missan, near Karbalā', and at as-Sulaymānīyah. There are also fertilizer factories and a paperboard factory around Basra. Approximately half of the existing manufacturing plants are to be found in Baghdad.

The petroleum and natural-gas industries support domestic and export-based refining and petrochemical in-

dustries, with major refineries at Basra and Dawrah, and five smaller facilities elsewhere. Petrochemicals, including fertilizers, urea, sulfuric acid, and ammonia gas, are produced well in excess of domestic agricultural needs at Abulfiluz, south of Basra.

The production of hydroelectric power is centred upon the Tigris River and its tributaries. Power dams are located on the Tigris at Wādī ath-Tharthār, on the Little Zab River at Dukān, and on the Upper Diyālā River at Darband-e Khān.

The country's first nuclear reactor, under construction near Baghdad, was destroyed by an Israeli air strike in June 1981; the government of Israel alleged that the facility was intended for nuclear weapons production. Stricter monitoring and controls for future installations were proposed by France, supplier of the technology.

Finance and trade. All banks and insurance companies, both domestic and foreign, were nationalized in 1964. The Central Bank of Iraq has the sole right of issuing currency. There are two commercial banks—the Commercial Bank of Iraq and the Rafidain Bank—and specialized banks for agriculture, construction, real estate, and industry.

The chief imports are manufactured goods, machinery, engines, automobiles, and pharmaceuticals. Iraq also imports some foodstuffs such as grains, sugar, and tea. After oil and dates, the chief exports are raw cotton, wool, hides and skins, fertilizers, essential oils and perfumes, and cement. Although Iraq's external trade was previously chiefly with western Europe, an increasing emphasis in the early 1970s on diversification of trading partners resulted in some alteration in the trade patterns. The Soviet Union and other Socialist states, the Far East, and neighbouring Arab countries now account for considerable trade. Iraq is a member of the Arab Common Market and has a special customs pact with Kuwait.

Transportation. Increasing modernization is reflected in changes of emphasis given to different parts of the transport system. Until the mid–20th century, transport was either by water or by animal. River transport had developed over the centuries to carry goods and persons, particularly from central to southern Iraq. The use of camels, horses, and horse-drawn carts complemented this system. Both of these modes were adequate for a traditional economy, but increasing commercialization and industrialization and the increasing scale of economic interaction between Iraq and its neighbours brought about fundamental changes.

Approximately one-third of the roads in Iraq are paved. Also, the desert land is sufficiently hard to allow cars, trucks, and buses to travel across it. The most developed parts of the road system are in central, northern, and southern Iraq. The opening of the Basra bridge in 1978 provided a direct link between Iraq's road network and the Asian Highway.

Three major railway lines connect the principal cities, and they all converge on Baghdad and Basra. Two narrow-gauge lines connect Baghdad with Basra, and link Baghdad with the northeast provinces east of the Tigris, and a standard-gauge railway connects Baghdad with Mosul and other northern cities and then continues on to Syria, Turkey, and beyond.

Five ports

Iraq's connection with the outside world through the seas is limited to five ports—one at Basra, 85 miles upstream on the Tigris from the Persian Gulf on the Shatt al-Arab; another at Umm Qaṣr, five miles from the Kuwait border; and three marine oil terminals at Fāw, Khor al-Amya, and Mina al-Bakr. All are accessible to oceangoing vessels; the facilities at Mina al-Bakr can accommodate supertankers. The Tigris is navigable by steamer from Basra to Baghdad; smaller power craft can travel upstream as far as Mosul. Only small local rivercraft use the Euphrates. Large rafts, however, carry goods downstream on both rivers.

Baghdad has two airports, one of which is international, and there is an international airport at Basra. The nation's one airline is controlled and owned by the Iraqi government and operates flights to airports in Asia, the Middle East, and Europe. Various Middle East and some major international airlines have landing privileges at Baghdad airport. Internal air traffic is insignificant.

ADMINISTRATIVE AND SOCIAL CONDITIONS

Government. While Iraq has had a series of constitutions, for all practical purposes Iraq is governed by an ad hoc constitution. It is officially a republic, with legislative power theoretically vested in an elected legislature, executive power in a president and a Council of Ministers, and judicial power in an independent judiciary. Because of the revolutionary nature of Iraqi politics since the overthrow of the monarchy in 1958, however, the political system operates with little reference to constitutional provisions. All legislative and executive powers are in fact exercised by the Council of Command of the Revolution, which since the early 1970s has been composed of 22 members, all of whom were members of the Regional Command of the Ba'th Arab Socialist Party and many of whom were army officers. The chairman of the council is also president of the republic, prime minister, and the commander of the armed forces. The president and the Council of Ministers are accountable to the Council of Command of the Revolution. Actual implementation and execution of broad policies are the function of each relevant minister. While each ministry is headed by a minister, an administrative officer, usually known as director general, is the actual day-to-day administrator. All employees of ministries are appointed civil servants.

The Council of Command of the Revolution

The country is divided for administrative purposes into 18 *muḥāfaẓāt* (governorates), three of which encompass the Kurdish Autonomous Region. Each *muḥāfaẓah* has a governor, or *muḥāfiẓ*, appointed by the minister of the interior, as chief administrative officer. Constitutional provision is made for an elective general council to assist the governor, but this provision met with obstacles to its implementation. Each *muḥāfaẓah* is administratively subdivided into districts headed by a *qā'im-maqām;* each district is then divided into *nāḥīyah*s headed by a *mudīr,* or director. The Kurdish Autonomous Region was formed by government decree in 1974. It contains the predominantly Kurdish provinces of Dahūk, as-Sulaymānīyah, and Irbīl, and is governed by a legislative council that meets in Irbīl city.

Towns and cities have their own municipal councils, each of which is directed by a *ra'īs al-baladīyah,* or mayor. The capital has special status and its own governor.

Political parties were outlawed with the overthrow of the monarchy in 1958. After the overthrow of Abdul Karim Kassem in 1963 by supporters of the Ba'th (an independent branch of the Syrian-based Ba'th Arab Socialist Party), Iraq became a one-party state. In 1970 the Democratic Kurdistan Party obtained permission to operate, primarily in the north. While the country is actually governed more by the armed forces than by the party, members of the Ba'th have nevertheless penetrated the armed forces. Nominally, all governing institutions espouse the Ba'thist ideology, which includes Socialism and a pan-Arab ideal. Constant turmoil and attempts by successive regimes to consolidate power have been used as justifications for perennial delay in resuming the conduct of normal political and constitutional processes. As a consequence, although Iraqis are assured of their rights to office and to voting, no election was held from 1958 until 1980, when a national legislature was elected.

Justice. Theoretically, Iraq has an independent judiciary, which combines French, Islāmic, and revolutionary elements. The system is supervised by a Judicial Council, headed by the minister of justice. There are three types of jurisdiction with corresponding court systems—the civil, religious, and special courts.

Dual system of justice

Civil courts deal with civil litigation, as well as all criminal cases. The courts of sessions hear serious criminal cases and serve as appellate courts in all criminal cases. Judgments in civil cases are appealed to the courts of appeal. The Court of Cassation is the highest court of the land and has original jurisdiction in crimes committed by public officials.

Each of the religious communities has its own religious court, which is governed by its respective religious law and which deals with all questions of personal status, marriage, and inheritance.

Special courts were formed in pursuance of the National

Security Act of 1965. They have original jurisdiction in all cases involving state security. Usually the majority of the judges are military officers.

Armed forces. The Iraqi Army is essentially the creation of the ousted Hashimite monarchy in its attempt to increase the power of the central government. It is a professional army whose officers are trained both in Iraq and in other countries. Since 1934 Iraqi males have been liable to a compulsory two-year service.

The army's military performance has been affected by the intervention of its officers in politics; since the mid–1930s, there has been a series of coups d'etat. These have generated tensions and have resulted in periodic dismissals of large numbers of officers. While the army is the most important branch of the defense force, there is a well-equipped air force and a small navy consisting of a few submarine chasers and torpedo boats. The Iraqi Army has seen action in the government's attempts to stem the secessionist activities of the Kurds. There are also an air force, a paramilitary People's Army, and a security force.

Education. Before 1921 the educational system was minimal, except for traditional, and particularly religious, instruction. Educational facilities were gradually expanded under the monarchy, and rapid expansion followed the revolution in 1958. The commitment to the ideal of universal compulsory education has fostered an increase in the percentage of females enrolled at all levels.

Higher education

Higher education is available at six Iraqi universities, located in Baghdad (three), Mosul, Basra, and as-Sulaymānīyah, and a technical institute. An agricultural college at Irbīl and various teachers' and technical colleges provide specialized instruction. The entire educational system is controlled closely by the Ministry of Education and the Ministry of Higher Education and Scientific Research, which also send considerable numbers of students to European and American universities for graduate training.

Health and welfare. The increasing Socialist commitment of the government has resulted in the expansion of welfare services, including health facilities and housing projects for the poor, particularly in Baghdad. The Law of Social Security of 1956 provides benefits for old age and illness. Other legislation provides for unemployment benefits and minimum wages and has established cooperative societies and industrial trade unions.

Trade unions

Iraq has had a vigorous although somewhat checkered trade-union movement since the 1930s. The strongest unions are those of workers in the oil industry. All the various trade unions are organized in the General Federation of Iraqi Trade Unions, which is a member of both the World Federation of Trade Unions and the International Confederation of Arab Trade Unions.

Employers' associations are relatively weak; they include the Iraq Federation of Industries and the Baghdad Chamber of Commerce. Similarly, peasant unions are somewhat less vigorous than urban-based trade unions. The General Federation of Peasant Societies was established in 1959. Other organizations include a teachers' union, a lawyers' association, and numerous urban and rural cooperative societies.

Social changes

The Iraqi social system—characterized in the past by rigid social and economic differentiation and ethnic, linguistic, and religious schisms—is being altered fundamentally by new, largely Western and Soviet ideologies, institutions, and cultural pursuits. The rapid expansion of the educational system, basic alterations in the system of landholding, and a greater commitment of the state to economic development are affecting Iraqi society and institutions. Poverty, though still prevalent, is declining. The gap is being narrowed between the rich few, who controlled the principal means of wealth and the political superstructure, and the many poor, primarily the peasants and the urban workers. An Iraqi middle class comprising the professional sector has emerged and is exercising a levelling effect on society. It controls the political system through the army and the Ba'th Party and is supported by the peasantry and workers.

CULTURAL LIFE

Traditional arts and crafts, including the weaving of world-famous rugs and carpets, derived their impetus from a rich, predominantly Arab, cultural legacy. Iraqis, like other Arabs, showed particular concern with poetry and prose literature; in that sense Iraq was but one part of Arab culture as a whole from which it derived and to which it contributed. Until the 20th century, cultural pursuits centred around religious schools and mosques. Folk tradition and dancing were also prevalent, particularly in the village and mountain areas of the country.

Increasing commercial, industrial, and urban activities and rapid changes in education and cultural orientation affected the traditional culture significantly. An increasing concern with Western-inspired models of culture, literature, and the arts began to take hold, and traditional arts experienced a decline.

Government support of traditional culture

The government, anxious to preserve some of the traditional creativity of the culture, assumed an increasingly active role in preserving and strengthening traditional folk arts and crafts. The Ministry of Culture and Information sponsored theatre and art groups whose purpose was to support traditional, as well as modern, cultural art and forms. The National Group for Acting, the Rashīd National Group, the Folklore Group, and the United Artists were established between 1965 and 1967 specifically to encourage the development of a "national" Iraqi culture. Both traditional and modern artistic and literary expressions are juxtaposed in contemporary Iraq, and the population supports each, according to its degree of modernization.

The government and several societies patronize the arts; magazines and radio and television stations periodically present performances or reviews of important cultural events. The Ministry of Culture supervises the country's excellent historical museums, such as the Iraqi Museum and the Mosul Museum, as well as the National Library in Baghdad and the public libraries. An Institute of Fine Arts, established in 1939 in Baghdad and strengthened after 1968, trains Iraqis in music, drama, and the plastic arts. Training in ballet and music is also offered by the School of Ballet and Music, the Institute of Rhythmical Studies, and the Institute of Fine Arts. Iraq has produced several excellent poets, such as Jawahiri and Nazik al-Malaa'ikah, and sculptors and painters, such as Khaled ar-Rahhal, Jawad Salim, Akram Shukri, and Hafidh ad-Durubi. For statistical data, see the "Britannica World Data" section in the current *Britannica Book of the Year.*

(I.A.A.-L./M.Kh./Ed.)

History

THE ORIGINS OF MESOPOTAMIAN HISTORY

The background. In the narrow sense, Mesopotamia is the area between the Euphrates and Tigris rivers, north or northwest of the bottleneck at Baghdad; it is al-Jazīrah (The Island) of the Arabs. South of this lies Babylonia, named after the city of Babylon. But in the broader sense the name Mesopotamia has come to be used for the area bounded on the northeast by the Zagros Mountains and on the southwest by the edge of the Arabian Plateau and that stretches from the Persian Gulf in the southeast to the spurs of the Anti-Taurus Mountains in the northwest. Only from the latitude of Baghdad do the Euphrates and Tigris truly become twin rivers, the *rāfidān* of the Arabs, which have constantly changed their courses over the millennia. The low-lying plain of the River Kārūn in Persia has always been closely related to Mesopotamia; but it is not considered part of Mesopotamia as it forms its own river system.

Mesopotamia, south of ar-Ramādī (about 70 miles, or 110 kilometres, west of Baghdad) on the Euphrates and the bend of the Tigris below Sāmarrā' (about 70 miles north-northwest of Baghdad), is flat alluvial land. Between Baghdad and the mouth of the Shatt al-Arab (the mouth of the Tigris and Euphrates) there is a difference in height only of about 30 feet (or 10 metres). As a result of the slow flow of the water, there are heavy deposits of silt; and the river-beds are raised. Consequently, the rivers often overflow their banks (and may even change their course) when they are not protected by high dikes or—in recent times—can be regulated above Baghdad by

the use of escape channels with overflow reservoirs. The extreme south is a region of extensive marshes and reed swamps, *hawr*s, which, probably since early times, have served as an area of refuge for oppressed and displaced peoples. The supply of water is not regular: as a result of the high average temperatures and a very low annual rainfall, the ground of the plain of latitude 35° S is hard and dry and unsuitable for plant cultivation for at least eight months in the year. Consequently, agriculture without risk of crop failure, which seems to have begun in the higher rainfall zones and in the hilly borders of Mesopotamia in the 10th millennium BC, began in Mesopotamia itself, the real heart of the civilization, only when artificial irrigation had been invented, bringing water to large stretches of territory through a widely branching network of canals. Since the ground is extremely fertile and, with irrigation and the necessary drainage, will produce in abundance, southern Mesopotamia became a land of plenty that could support a considerable population. The cultural superiority of north Mesopotamia, which may have lasted until *c.* 4000 BC, was finally overtaken by the south when men there had responded to the challenge of their situation.

The impetus of artificial irrigation

The present climatic conditions are fairly certainly similar to those of 8,000 years ago. An English survey of ruined settlements in the area 30 miles—about 50 kilometres—around ancient Hatra (180 miles northwest of Baghdad) has shown that the southern limits of the zone in which agriculture is possible without artificial irrigation has remained unchanged since the first settlement of al-Jazīrah.

The availability of raw materials is a historical factor of great importance, as is the dependence on those materials that had to be imported. In Mesopotamia, agricultural products and those from stock breeding, fisheries, datepalm cultivation, and reed industries—in short, corn, vegetables, meat, leather, wool, horn, fish, dates, and reed and plant-fibre products—were available in plenty and could easily be produced in excess of home requirements to be exported. There are bitumen springs at Hīt (90 miles west of Baghdad) on the Euphrates (the Is of Herodotus). On the other hand, wood, stone, and metal were rare or even entirely absent. The date palm—virtually the national tree of Iraq—yields a wood suitable only for rough beams and not for finer work. Stone is completely lacking in southern Mesopotamia, although limestone is quarried in the desert about 35 miles to the west; and "Mosul marble" is found not far from the Tigris in its middle reaches. Metal can only be obtained in the mountains, and the same is true of precious and semiprecious stones. Consequently, southern Mesopotamia in particular was destined to be a land of trade from the start. Only rarely could "empires" extending over a wider area guarantee themselves imports by plundering or by subjecting neighbouring regions.

The importance of clay

The raw material that epitomizes Mesopotamian civilization is clay: in the almost exclusively mud-brick architecture and in the number and variety of clay figurines and pottery artifacts, Mesopotamia bears the stamp of clay as does no other civilization; and nowhere in the world but in Mesopotamia and the regions over which its influence was diffused was clay used as the vehicle for writing. Such phrases as cuneiform civilization, cuneiform literature, and cuneiform law can apply only where men had held the idea of using soft clay not only for bricks and jars and for the jar stoppers on which a seal could be impressed as a mark of ownership but also as the vehicle for impressed signs to which established meanings were assigned—an intellectual achievement that amounted to nothing less than the invention of writing.

The character and achievements of ancient Mesopotamia. Questions as to what ancient Mesopotamian civilization did and did not accomplish, how it influenced its neighbours and successors, and what its legacy has transmitted of value are posed from the standpoint of 20th-century civilization and are in part coloured by ethical overtones, so that the answers can only be relative. Modern scholars assume the ability to assess the sum total of an "ancient Mesopotamian civilization"; but, since the publication of an article by Benno Landsberger, an Assyriologist, on "Die Eigenbegrifflichkeit der babylonischen Welt," it has become almost a commonplace to call attention to the necessity of viewing ancient Mesopotamia and its civilization as an independent entity.

Unique variety of the Mesopotamian civilization

Mesopotamia has many languages, many cultures, is broken up into many historical periods, has no real geographical unity, and above all no permanent capital city, so that by its very variety it stands out from other civilizations with greater uniformity, particularly that of Egypt. The script and the pantheon constitute the unifying factors, but in these also Mesopotamia shows its predilection for multiplicity and variety: written documents turned out in quantities and a pantheon consisting of more than 1,000 deities, even though many names may apply to different manifestations of a single god. Throughout 3,000 years of Mesopotamian civilization, each century gave birth to the next. Thus classical Sumerian civilization influenced that of the Akkadians; and the Ur III Empire, which itself represented a Sumero-Akkadian synthesis, exercised its influence on the first quarter of the 2nd millennium BC. With the Hittites, large areas of Anatolia were infused with the culture of Mesopotamia from 1700 BC onward. The contacts between Syrian and Palestinian scribal schools and Babylonian civilization during the Amarna period (14th century BC) must have led to the enrichment of the literature of the west Semitic lands: this, and not direct borrowing, is the explanation of the similarity of themes in cuneiform literature and the Old Testament, such as the story of the Flood or the motif of the righteous sufferer.

The world of mathematics and astronomy owes much to the Babylonians: for instance, the sexagesimal system for the calculation of time and angles, which is still practical because of the multiple divisibility of the number 60; the Greek day of 12 "double-hours"; and the zodiac and its signs. But often the origins and routes of borrowings are obscure, as in the problem of the survival of ancient Mesopotamian legal theory.

The achievement of the civilization itself may be expressed in terms of its best points—moral, aesthetic, scientific, and, not least, literary. Legal theory flourished and was sophisticated early on, being expressed in several collections of legal decisions, the so-called codes, of which the best known is the Code of Hammurabi. Throughout these codes recurs the concern of the ruler for the weak, the widow, and the orphan—even if, sometimes, the phrases were regrettably only literary clichés. The aesthetics of art are too much governed by subjective values to be assessed in absolute terms; but certain peaks stand out above the rest, notably the art of Uruk IV, the seal engraving of the Akkad period, and the relief sculpture of Ashurbanipal. Nonetheless, there is nothing in Mesopotamia to match the sophistication of Egyptian art. Science the Mesopotamians had, of a kind, though not as the word is understood of Greek science. From its beginnings in Sumer before the middle of the 3rd millennium BC, Mesopotamian science was characterized by endless, meticulous enumeration and ordering into columns and series, with the ultimate ideal of including all things in the world but without the wish or ability to synthesize and reduce the material to a system. Not a single general law has been found and only rarely the use of analogy. Nevertheless, it remains a highly commendable achievement that Pythagoras' law (that the sum of the squares on the two shorter sides of a right-angled triangle equals the square on the longest side), even though it was never formulated, was being applied as early as the 18th century BC. Technical accomplishments were perfected in the building of the ziggurats (temple towers like pyramids), with their huge bulk, and in irrigation, both in practical execution and in theoretical calculations beforehand. At the beginning of the 3rd millennium BC an artificial stone in use at Uruk (160 miles south-southeast of modern Baghdad) constitutes a forerunner of concrete, but the secret of its manufacture apparently was lost in subsequent years.

Scientific achievements

Writing pervaded all aspects of life and gave rise to a highly developed bureaucracy—one of the most tenacious legacies of the ancient East. Remarkable organizing ability was required to administer huge estates, in which, under the 3rd dynasty of Ur, for example, it was not unusual to prepare accounts for thousands of cattle or

tens of thousands of bundles of reeds. Above all, the literature of Mesopotamia is one of its finest cultural achievements. Though there are many modern anthologies and chrestomathies (compilations of useful learning), with translations and paraphrases of Mesopotamian literature, as well as attempts to write its history, it cannot truly be said that "cuneiform literature" has been resurrected to the extent that it deserves. There are partly material reasons for this: many clay tablets survive only in a fragmentary condition, and duplicates that would restore the texts have not yet been discovered, so that there are still large gaps. A further reason is the inadequate knowledge of the languages: insufficient acquaintance with the vocabulary and, in Sumerian, major difficulties with the grammar. Consequently, another generation of Assyriologists will pass before the great myths, epics, lamentations, hymns, "law codes," wisdom literature, and pedagogical treatises can be presented to the reader in such a way that he can fully appreciate the high level of literary creativity of those times.

The classical and medieval views of Mesopotamia; its rediscovery in modern times. Before the first excavations in Mesopotamia, around 1840, nearly 2,000 years had passed during which knowledge of the ancient Near East was derived from three sources only: the Bible, Greek and Roman authors, and the excerpts from the writings of Berosus, a Babylonian who wrote in Greek. In 1800 very little more was known than in AD 800, although these sources had served to stir the imagination of poets and artists, down to *Sardanapalus* by the 19th-century English poet Lord Byron. Apart from the building of the Tower of Babel, the Old Testament mentions Mesopotamia only in those historical contexts in which the kings of Assyria and Babylonia affected the course of events in Israel and Judah: in particular Tiglath-pileser III, Shalmaneser V, and Sennacherib, with their policy of deportation, and the Babylonian Exile introduced by Nebuchadrezzar II. Of the Greeks, Herodotus of Halicarnassus (5th century BC, a contemporary of Xerxes I and Artaxerxes I), was the first to report on "Babylon and the rest of Assyria"; at that date the Assyrian Empire had been overthrown for more than 100 years. The Athenian Xenophon took part in an expedition, during 401–399 BC, of Greek mercenaries who crossed Asia Minor, made their way down the Euphrates as far as the vicinity of Baghdad, and returned up the Tigris after the famous Battle of Cunaxa. In his *Cyropaedia* Xenophon describes the final struggle between Cyrus II and the Neo-Babylonian Empire. Later, the Greeks adopted all kinds of fabulous tales about King Ninos, Queen Semiramis, and King Sardanapalus. These stories are described mainly in the historical work of Diodorus Siculus (1st century BC), who based them on the reports of a Greek physician, Ctesias (405–359 BC). Herodotus saw Babylon with his own eyes, and Xenophon gave an account of travels and battles. But all later historians wrote at second or third hand, with one exception, Berosus (born *c.* 340 BC), who emigrated at an advanced age to the Aegean island of Cos, where he is said to have composed his three books of *Babylōniaka.* Unfortunately, only extracts from them survive, prepared by one Alexander Polyhistor (1st century BC), who, in his turn, served as a source for the church father Eusebius (who died AD 342). Berosus derided the "Greek historians" who had so distorted the history of his country. He knew, for example, that it was not Semiramis who founded the city of Babylon; but he was himself the prisoner of his own environment and cannot have known more about the history of his land than was known in Babylonia itself in the 4th century BC.

Greek views of Mesopotamia

Berosus' first book dealt with the beginnings of the world and with a myth of a composite being, Oannes, half fish, half man, who came ashore in Babylonia at a time when men still lived like the wild beasts. Oannes taught them the essentials of civilization: writing, the arts, law, agriculture, surveying, and architecture. The name Oannes must be derived from the cuneiform U'anna (Sumerian) or Umanna (Akkadian), a second name of the mythical figure Adapa, the bringer of civilization. The second book of Berosus contained the Babylonian king list from the beginning to King Nabonassar (Nabu-naṣir, 747–734 BC), a contemporary of Tiglath-pileser III. Berosus' tradition, beginning with a list of primeval kings before the Flood, is a reliable one; it agrees with the tradition of the Sumerian king list, and even individual names can be traced back exactly to their Sumerian originals. Even the immensely long reigns of the primeval kings, which lasted as long as "18 sars" (= 18 × 3,600 = 64,800) of years, are found in Berosus. Furthermore, he was acquaninted with the story of the Flood, with Cronus as its instigator and Xisuthros (or Ziusudra) as its hero, and with the building of an ark. The third book is presumed to have dealt with the history of Babylonia from Nabonassar to the time of Berosus himself.

Diodorus made the mistake of locating Nineveh on the Euphrates, and Xenophon gave an account of two cities, Larissa and Mespila, which, it is suspected, represented the ancient Kalakh (20 miles southeast of modern Mosul) and Nineveh (just north of modern Mosul), Mespila probably being nothing more than the word of the local Aramaeans for ruins: there can be no clearer instance of the rift that had opened between the ancient Near East and the classical West. In sharp contrast, the East had a tradition that the ruins opposite Mosul (in north Iraq) conceal ancient Nineveh. When a Spanish rabbi from Navarre, Benjamin of Tudela, was travelling in the Near East between 1160 and 1173, Jews and Muslims alike knew the position of the grave of the prophet Jonah. The credit for the rediscovery of the ruins of Babylon goes to an Italian, Pietro della Valle, who correctly identified the vast ruins north of modern al-Ḥillah (60 miles south of Baghdad); and he must have seen there the large rectangular tower that represented the ancient ziggurat. Previously, other travellers had sought the Tower of Babel in two other monumental ruins: Birs Nimrūd, the massive brick structure of the ziggurat of ancient Borsippa (near al-Ḥillah), vitrified by lightning, and the ziggurat of the Kassite capital, Dūr Kurigalzu, at Burj ʿAqarqūf, 22 miles west of Baghdad. Pietro della Valle also brought back to Europe the first specimens of cuneiform writing, stamped brick, of which highly impressionistic reproductions were made. Thereafter, European travellers visited Mesopotamia with increasing frequency, among them Carsten Niebuhr (an 18th-century German traveller); Claudius James Rich (a 19th-century Orientalist and traveller); and Ker Porter (a 19th-century traveller).

The work of Pietro della Valle

In modern times, a third Near Eastern ruin drew visitors from Europe—Persepolis, in the land of Persia east of Susiana near modern Shīrāz. In 1602 reports had filtered back to Europe of inscriptions that were not in Hebrew, Arabic, Aramaic, Georgian, or Greek. In 1700 an Englishman, Thomas Hyde, coined the term "cuneiform" for these inscriptions, and by the middle of the 18th century it was known that the Persepolis inscriptions were related to those of Babylon. Carsten Niebuhr distinguished three separate alphabets (Babylonian, Elamite, and Old Persian cuneiform). The first promising attempt at decipherment was made by G.F. Grotefend in 1802, by use of the kings' names in the Old Persian versions of the trilingual inscriptions, although his later efforts led him up a blind alley. Thereafter, the efforts to decipher cuneiform gradually developed in the second half of the 19th century into a discipline of ancient Oriental philology, which was based on results established through the pioneering work of Emile Burnouf, Edward Hincks, Sir Henry Rawlinson, and many others.

Today this subject is still known as Assyriology, because at the end of the 19th century the great majority of cuneiform texts came from the Assyrian city of Nineveh, in particular, from the library of King Ashurbanipal in the mound of Kouyunçik at Nineveh.

Modern archaeological excavations. One hundred and thirty years separate the first excavations in Mesopotamia—adventurous expeditions involving great personal risks and far from the protection of helpful authorities—from those of the present day with their specialist staffs, modern technical equipment, and objectives wider than the mere search for valuable antiquities. The progress of five generations of excavators has led to a situation in which less is recovered more accurately; in other words,

the finds are observed, measured, and photographed as precisely as possible. At first digging was unsystematic, with the consequence that although huge quantities of clay tablets and large and small antiquities were brought to light, the findspots were rarely described with any accuracy. Not until the beginning of the 20th century did excavators learn to isolate the individual bricks in the walls that had previously been erroneously thought to be nothing more than packed clay; the result was that various characteristic brick types could be distinguished, and successive architectural levels established. Increased care in excavation does, of course, carry with it the risk that the pace of discovery will slow down. Moreover, the eyes of the local inhabitants are now sharpened and their appetite for finds is whetted, so that clandestine diggers have established themselves as the unwelcome colleagues of the archaeologists.

Problems of mud-brick buildings

A result of the technique of building with mud brick (mass production of baked bricks was impossible because of the shortage of fuel) was that the buildings were highly vulnerable to the weather and needed constant renewal; layers of settlement rapidly built up, creating a *tall,* a mound of occupation debris that is the characteristic ruin form of Mesopotamia. The word itself appears among the most original vocabulary of the Semitic languages and is attested as early as the end of the 3rd millennium BC. Excavation is made more difficult by this mound formation, since both horizontal and vertical axes have to be taken into account. Moreover, the depth of each level is not necessarily constant, and foundation trenches may be dug down into earlier levels. A further problem is that finds may have been removed from their original context in antiquity. Short-lived settlements that did not develop into mounds mostly escape observation, but aerial photography can now pick out ground discolorations that betray the existence of settlements. Districts with a high water level today, such as the reed marshes (*hawr*s), or ruins that are covered by modern settlements, such as Arbīl, some 200 miles north of Baghdad, or that are surmounted by shrines and tombs of holy men are closed to archaeological research altogether.

Excavations in Mesopotamia have been almost exclusively national undertakings (France, England, the United States, Germany, Iraq, Denmark, Belgium, Italy, Japan, and the Soviet Union) and the "joint expedition" first sent to Ur (190 miles south-southeast of Baghdad) in the 1920s has remained a rare exception. The history of archaeological research in Mesopotamia falls into four categories, represented by phases of differing lengths: the first, and by far the longest, begins with the French expedition to Nineveh (1842) and Khorsabad (the ancient Dur Sharrukin, 20 miles northeast of modern Mosul; 1843–55) and that of the English to Nineveh (1846–55) and Nimrūd (ancient Kalakh, biblical Calah; 1845; with interruptions until 1880). This marked the beginning of the "classic" excavations in the important ancient capitals, where spectacular finds might be anticipated. The principal gains were the Assyrian bull colossuses and wall reliefs and the library of Ashurbanipal from Nineveh, whereas the ground plans of temples and palaces were quite as valuable. While these undertakings had restored to life the remains of the Neo-Assyrian Empire from the 1st millennium BC, new French initiatives in Tello (Arabic Tall Lūḥ), 155 miles southeast of Baghdad, from 1877 onward reached almost 2,000 years further back into the past. There they rediscovered a people whose language had already been encountered in bilingual texts from Nineveh—the Sumerians. Tello (ancient Girsu) yielded not only inscribed material that, quite apart from its historical interest, was critical for the establishment of the chronology of the second half of the 3rd millennium BC but also a large number of artistic masterpieces. Thereafter, excavations in important cities spread to form a network, including Susa, 150 miles west of Isfahan in Iran (France; 1884 onward); Nippur, 90 miles southeast of Baghdad (the United States; 1889 onward); Babylon, 55 miles south of Baghdad (Germany; 1899–1917 and again from 1957 onward); Ashur, modern ash-Sharqāṭ, 55 miles south of Mosul (Germany; 1903–14); Uruk (Germany; 1912–13 and from 1928 onward); and at Ur (England and the United States; 1918–34). Mention should also be made of the German excavations at Boğazköy in central Turkey, the ancient Hattusas, capital of the Hittite Empire, which have been carried on, with interruptions, since 1906.

Investigations after 1925

The second phase began in 1925 with the commencement of U.S. excavations at Yorgan Tepe (ancient Nuzu), 140 miles north of Baghdad, a provincial centre with Old Akkadian, Old Assyrian, and Middle Assyrian/Hurrian levels. There followed, among others, French excavations at Arslan Tash (ancient Hadatu; 1928); Tall al-Aḥmar (ancient Til Barsib; 1929–31); and above all at Tall Ḥarīrī (ancient Mari; 1933 onward); and U.S. excavations in the Diyālā region, 90 miles northeast of Baghdad; Tall al-Asmar (ancient Eshnunna), Khafājī, and other sites. Thus, excavation in Mesopotamia had moved away from the capital cities to include the "provinces." Simultaneously, it expanded beyond the limits of Mesopotamia and Susiana and revealed outliers of "cuneiform civilization" on the Syrian coast at Ras Shamra (ancient Ugarit; France, 1929 onward) and on the Orontes of northern Syria at alʿAṭshānah (ancient Alalakh; England, 1937–39, 1947–49), whereas Danish excavations on the islands of Bahrain and Faylakah, off the Tigris-Euphrates Delta, since 1954 have disclosed staging posts between Mesopotamia and the Indus Valley civilization.

In its third phase, archaeological research in Mesopotamia and its neighbouring lands has probed back into prehistory and protohistory. The objective of these investigations, initiated by U.S. archaeologists, was to trace as closely as possible the successive chronological stages in the progress of man from hunter and gatherer to settled farmer and, finally, to city dweller. These excavations are strongly influenced by the methods of the prehistorian, and the principal objective is no longer the search for texts and monuments. Apart from the United States investigations, Iraq itself has taken part in this phase of the history of investigation, as has Japan since 1956.

Finally, the fourth category, which runs parallel with the first three phases, is represented by "surveys," which do not concentrate on individual sites but attempt to define the relations between single settlements, their positioning along canals or rivers, or the distribution of central settlements and their satellites. Since shortage of time, money, and an adequate task force preclude the thorough investigation of large numbers of individual sites, the method employed is that of observing and collecting finds from the surface. Of these finds, the latest in date will give a rough termination date for the duration of the settlement, but, since objects from earlier, if not the earliest, levels work their way to the surface with a predictable degree of certainty or are exposed in rain gullies, an intensive search of the surface of the mound allows conclusions as to the total period of occupation with some degree of probability. If the individual periods of settlement are marked on superimposed maps, a very clear picture is obtained of the fluctuations in settlement patterns, of the changing proportions between large and small settlements, and of the equally changeable systems of riverbeds and irrigation canals—for when points on the map lie in line, it is a legitimate assumption that they were once connected by water courses. During the four phases outlined, the objectives and methods of excavation have broadened and shifted. At first the chief aim was the recovery of valuable finds suitable for museums, but at the same time there was, from early on, considerable interest in the architecture of Mesopotamia, which won for it the place it deserves in architectural history. Alongside philology, art history also made great strides, building up a chronological framework by the combination of evidence from stratigraphical and stylistic criteria, particularly in pottery and cylinder seals. The discovery of graves and a variety of burial customs threw new light on the history of religion, stimulated by the interest of Old Testament studies. Later, an increasing interest from neighbouring disciplines emerged. While pottery had been collected previously for purely aesthetic motives or from the point of view of art history, attention came to be paid increasingly to the everyday wares, and greater insight into social and economic history was based on knowledge of the distribution and frequency of shapes and materials.

Finally, the observation and investigation of animal bones and plant remains (pollen and seed analysis) supplied invaluable information on the process of domestication, the conditions of animal husbandry, and the advances in agriculture; such studies demand the cooperation of both zoologists and paleobotanists.

The emergence of Mesopotamian civilization. *The Late Neolithic period and the Chalcolithic period.* Between about 10,000 BC and the genesis of large permanent settlements, the following stages of development are distinguishable, some of which run parallel: (1) the change to sedentary life, or the transition from continual or seasonal change of abode, characteristic of hunter-gatherers and the earliest cattle breeders, to life in one place over a period of several years or even permanently; (2) the transition from experimental plant cultivation to the deliberate and calculated farming of grains and leguminous plants; (3) the erection of houses and the associated "settlement" of the gods in temples; (4) the burial of the dead in cemeteries; (5) the invention of clay vessels, made at first by hand, then turned on the wheel and fired to ever greater degrees of hardness, at the same time receiving almost invariably decoration of incised designs or painted patterns; (6) the development of specialized crafts and the distribution of labour; and (7) metal production (the first use of metal—copper—marks the transition from the Late Neolithic to the Chalcolithic period).

Pre-urban stages of development

These stages of development can only rarely be dated on the basis of a sequence of levels at one site alone. Instead, an important role is played by the comparison of different sites, starting with the assumption that what is simpler and technically less accomplished is older. In addition to this type of dating, which can be only relative, over the last 20 years the carbon-14 method has been of great assistance. By this method the known rate of decay of the radioactive carbon isotope (carbon-14) in wood, horn, plant fibres, and bones allows the time that has elapsed since the "death" of the material under examination to be calculated. Although a plus/minus discrepancy of up to 200 years has to be allowed for, this is not such a great disadvantage in the case of material 6,000 to 10,000 years old. Even when skepticism is necessary because of the use of an inadequate sample, carbon-14 dates are still very welcome as confirmation of dates arrived at by other means.

The first agriculture, the domestication of animals, and the transition to sedentary life took place in regions in which animals that were easily domesticated, such as sheep, goats, cattle, and pigs, and the wild prototypes of grains and leguminous plants, such as wheat, barley, bitter vetch, pea, and lentil, were present. Such centres of dispersion may have been the valleys and grassy border regions of the mountains of Iran, Iraq, Anatolia, Syria, and Palestine; but they could also be, for example, the northern slopes of the Hindu Kush. As settled life, which caused a drop in infant mortality, led to the increase of the population, settlement spread out from these centres into the plains—although it must be remembered that this process, described with the phrase Neolithic revolution, in fact took thousands of years.

Greek views of Mesopotamia

Representative of the first settlements on the borders of Mesopotamia are the adjacent sites of Zawi Chemi and Shanidar, which lie northwest of Rawāndūz. They date from the transition from the 10th to the 9th millennium BC and are classified as "pre-pottery." The finds included querns (primitive mills) for grinding grain (whether wild or cultivated is unknown), the remains of huts about 13 feet in diameter, and a cemetery with grave goods. The presence of copper beads is evidence of acquaintance with metal, though not necessarily with the technique of working it into tools; and the presence of obsidian (volcanic glass) is indicative of the acquisition of nonindiginous raw materials by means of trade. The bones found testify that sheep were already domesticated at Zawi Chemi-Shanidar.

At Karīm Shahir, a site that cannot be accurately tied chronologically to Shanidar, clear proof was obtained both of the knowledge of grain cultivation, in the form of sickle blades showing sheen from use, and of the baking of clay, in the form of lightly fired clay figurines. Still in the hilly borders of Mesopotamia, a sequence of about 3,000 years can be followed at the site of Qal'at Jarmo, east of Kirkūk, some 150 miles north of Baghdad. The beginning of this settlement can be dated to around 6750 BC; and excavations uncovered 12 archaeological levels of a regular village, consisting of about 20–25 houses built of packed clay, sometimes with stone foundations, and divided into several rooms. The finds included types of wheat (emmer and einkorn) and two-row barley, the bones of domesticated goats, sheep, and pigs, and obsidian tools, stone vessels, and, in the upper third of the levels, clay vessels with rough painted decorations, providing the first certain evidence for the manufacture of pottery. Jarmo must be roughly contemporary with the sites of Jericho (13 miles east of Jerusalem) and of Çatalhüyük in Anatolia (central Turkey). Those sites, with their walled settlements, seem to have achieved a much higher level of civilization; but too much weight must not be placed on the comparison, as no other sites in and around Mesopotamia confirm the picture deduced from Jarmo alone.

About 1,000 years later are two villages that are the earliest so far discovered in the plain of Mesopotamia: Ḥassūna, near Mosul, and Tall Ṣawwān, near Sāmarrā'. At Ḥassūna the pottery is more advanced, with incised and painted designs; but the decoration is still unsophisticated. One of the buildings found may be a shrine, to judge from its unusual ground plan. Apart from emmer there occurs, as the result of mutation, six-row barley, which was later to become the chief grain crop of southern Mesopotamia. In the case of Tall Ṣawwān, it is significant that the settlement lay south of the boundary of rainfall agriculture; thus, it must have been dependent on some form of artificial irrigation, even if this was no more than the drawing of water from the Tigris. This, therefore, gives a date after which the settlement of parts of southern Mesopotamia would have been feasible.

The emergence of cultures. For the next millennium, the 5th, it is customary to speak in terms of various "cultures" or "horizons," distinguished in general by the pottery, which may be classed by its colour, shape, hardness, and, above all, its decoration. The name of each horizon is derived either from the type site or from the place where the pottery was first found: Sāmarrā' on the Tigris, Tall Halaf in the central Jazīrah, Ḥassūna Level V, al-'Ubaid near Ur, and Ḥājj Muḥammad on the Euphrates, not far from as-Samāwah (some 150 miles south-southeast of Baghdad). Along with the improvement of tools, the first evidence for water transport (a model boat from the prehistoric cemetery at Eridu, in the extreme south of Mesopotamia, *c.* 4000 BC), and the development of terra-cottas, the most impressive sign of progress is the constantly accelerating advance in architecture. This can best be followed in the city of Eridu, which in historical times was the centre of the cult of the Sumerian god Enki.

The work of Pietro della Valle

Originally a small, single-roomed shrine, the temple in the Ubaid period consisted of a rectangular building, measuring 80 by 40 feet (24 by 12 metres), that stood on an artificial terrace. It had an "offering-table" and an "altar" against the short walls, aisles down each side, and a facade decorated with niches. This temple, standing on a terrace probably originally designed to protect the building from flooding, is usually considered the prototype of the characteristic religious structure of later Babylonia, the ziggurat. The temple at Eridu is in the very same place as that on which the Enki-ziggurat stood in the time of the 3rd dynasty of Ur (*c.* 2112–*c.* 2004 BC), so the cult tradition must have existed on the same spot for at least 1,500–2,000 years before Ur III itself. Remarkable as this is, however, it is not justifiable to assume a continuous ethnic tradition. The flowering of architecture reached its peak with the great temples of Uruk, built around the turn of the 4th to 3rd millennium BC (Levels Uruk VI to IV).

In extracting information as to the expression of mind and spirit during the six millennia preceding the invention of writing sketched above, it is necessary to take account of four major sources: decoration on pottery, the care of the dead, sculpture, and the designs on seals. There is, of

course, no justification in assuming any association with ethnic groups.

The most varied of these means of expression is undoubtedly the decoration of pottery. It is hardly coincidental that in regions in which writing had developed, high-quality painted pottery was no longer made. The motifs in decoration are either abstract and geometrical or figured, although there is also a strong tendency to geometric stylization. An important question is the extent to which the presence of symbols, such as the bucranium (a sculptured ornament representing an ox skull), can be considered as expressions of specific religious ideas, such as a bull cult, and, indeed, how much the decoration was intended to convey meaning at all.

It is not known how ancient is the custom of burying the dead in graves nor whether its intention was to maintain communication (by the cult of the dead) or to guard against the demonic power of the unburied dead left free to wander. A cemetery, or collection of burials associated with grave goods, is first attested at Zawi Chemi-Shanidar. The presence of pots in the grave indicates that the bodily needs of the dead person were provided for, and the discovery of the skeleton of a dog or, at Eridu, of a model boat in the cemetery suggests that it was believed that the activities of life could be pursued in the hereafter.

Early statuettes

The earliest sculpture takes the form of very crudely worked terra-cotta representations of women; the "Ubaid horizon," however, has figurines of both women and men, with very slender bodies, protruding features, arms akimbo and the genitals accurately indicated, and also of women suckling children. It is uncertain whether it is correct to describe these statuettes as idols, whether the figures were cult objects, such as votive offerings, or whether they had a magical significance, such as fertility charms, or, indeed, what purpose they did fulfill.

Seals are first attested in the form of stamp seals at Tepe Gawra, north of Mosul. Geometrical designs are found earlier than scenes with figures, such as men, animals, conflict between animals, copulation, or dance. Here again, it is uncertain whether the scenes are intended to convey a deeper meaning. Nevertheless, unlike pottery, a seal has a direct relationship to a particular individual or group; for the seal identifies what it is used to seal (a vessel, sack, or other container) as the property or responsibility of a specific person. To that extent, seals represent the earliest pictorial representations of persons. The area of distribution of the stamp seal was northern Mesopotamia, Asia Minor, and Iran. Southern Mesopotamia, on the other hand, was the home of the cylinder seal, which was either an independent invention or was derived from stamp seals engraved on two faces. The cylinder seal, with its greater surface area and more practical application, remained in use into the 1st millennium BC. Because of the continuous changes in the style of the seal designs, cylinder seals are among the most valuable of chronological indicators for archaeologists.

In general, the prehistory of Mesopotamia can only be described by listing and comparing human achievements, not by recounting the interaction of individuals or peoples. There is no basis for reconstructing the movements and migrations of peoples unless one is prepared to equate the spread of particular archaeological types with the extent of a particular population or the change of types with a change of population or the appearance of new types with an immigration.

The only certain evidence for the movement of men beyond their own territorial limits is provided at first by material finds that are not indigenous. The discovery of obsidian and lapis lazuli at sites in Mesopotamia or in its neighbouring lands postulates the existence of trade, whether it consisted of direct caravan trade or of a succession of intermediate stages.

Just as no ethnic identity is recognizable, so nothing is known of the social organization of prehistoric settlements. It is not possible to deduce anything of the "government" in a village nor of any supraregional connections that may have existed under the domination of one centre. Constructions that could only have been accomplished by the organization of men in large numbers are first found in Uruk Levels VI to IV: the dimensions of these buildings suggest that they were intended for gatherings of hundreds of people. As for artificial irrigation, which was indispensable for agriculture in south Mesopotamia, the earliest form was probably not the irrigation canal. It is assumed that at first floodwater was dammed up to collect in basins, near which the fields were located. Canals, which led the water farther from the river, would have become necessary when the land in the vicinity of the river could no longer supply the needs of the population.

Mesopotamian protohistory. Attempts have been made by philologists to reach conclusions as to the origin of the flowering of civilization in southern Mesopotamia by the analysis of Sumerian words. It has been thought possible to isolate an earlier, non-Sumerian substratum from the Sumerian vocabulary by assigning certain words on the basis of their endings to either a Neolithic or a Chalcolithic language stratum. These attempts are based on the phonetic character of Sumerian at the beginning of the 2nd millennium BC, which is at least 1,000 years later than the invention of writing. Quite apart, therefore, from the fact that the structure of Sumerian words themselves is far from adequately investigated, the enormous gap in time casts grave doubt on the criteria used to distinguish between Sumerian and "pre-Sumerian" vocabulary.

Analyses of Sumerian words

The earliest peoples of Mesopotamia who can be identified from inscribed monuments and written tradition—people in the sense of speakers of a common language—are, apart from the Sumerians, Semitic peoples (Akkadians or pre-Akkadians) and Subarians (the predecessors or near relatives of the Hurrians, who appear in northern Mesopotamia around the end of the 3rd millennium BC). Their presence is known, but no definite statements about their past or possible routes of immigration are possible.

At the turn of the 4th to 3rd millennium BC, the endlessly long span of prehistory is over, and the threshold of the historical era is gained, captured by the existence of writing. Names, speech, and actions are fixed in a system that is composed of signs representing complete words or syllables. The signs may consist of realistic pictures, abbreviated representations, and perhaps symbols selected at random. Since clay is not well suited to the drawing of curved lines, a tendency to use straight lines rapidly gained ground. When the writer pressed the reed in harder at the beginning of a stroke, it made a triangular "head," and thus "wedges" were impressed into the clay. It is the Sumerians who are usually given the credit for the invention of this, the first system of writing in the Near East. As far as they can be assigned to any language, the inscribed documents from before the dynasty of Akkad (*c.* 2334–*c.* 2154 BC) are almost exclusively in Sumerian. The extension, moreover, of the writing system to include the creation of syllabograms by the use of the sound of a logogram (sign representing a word), such as *gi,* "a reed stem," used to render the verb *gi,* "to return," can only be explained in terms of the Sumerian language. It is most probable, however, that Mesopotamia of the 4th millennium BC, just as in later times, was composed of many races. This makes it likely that, apart from the Sumerians, the interests and even initiatives of other language groups may have played their part in the formation of the writing system. A strong argument in favour of this is afforded by the system of numbers and measures, which, in contrast to the rudimentary methods of writing words, was very precise and complex from the beginning. In addition to the sexagesimal system indigenous to Sumerian, which has a progression of alternating tens and sixes (1, 10, 60, 600, 3,600, 36,000, and so on), there is the decimal system (1, 10, 100, and so on). Since the number 100 had no special significance for the Sumerians but was expressed in words as 60+40, the decimal system was probably intended to do justice to the numerals of a second, non-Sumerian language.

Sumerian is an agglutinative language: prefixes and suffixes, which express various grammatical functions and relationships, are attached to a noun or verb root in a "chain." Attempts to identify Sumerian more closely by comparative methods have as yet been unsuccessful

and will very probably remain so, as languages of a comparable type are only known from AD 500 (Georgian) or AD 1000 (Basque)—that is, 3,000 years later. Over so long a time the rate of change in a language, particularly one that is not fixed in a written norm, is so great that one can no longer determine whether apparent similarity between words goes back to an original relationship or is merely fortuitous. Consequently, it is impossible to obtain any more accurate information as to the language group to which Sumerian may once have belonged.

Urban development in the 4th millennium BC

The most important development in the course of the 4th millennium BC was the birth of the city. There were precursors such as the unwalled prepottery settlement at Jericho of *c.* 7000 BC, but the beginning of cities with a more permanent character came only later. There is no generally accepted definition of a "city." In this context, it means a settlement that serves as a centre for smaller places, one that possesses one or more shrines of one or more major deities, has extensive granaries, and, finally, displays an advanced stage of specialization in the crafts.

The earliest cities of southern Mesopotamia, as far as their names are known, are Eridu, Uruk, Bad-tibira, Nippur, and Kish (35 miles south-southeast of Baghdad). The archaeological surveys of Robert McCormick Adams and Hans Nissen have shown how the relative size and number of the settlements gradually shifted: the number of small or very small settlements was reduced overall, whereas the number of larger places grew. The clearest sign of urbanization can be seen at Uruk, with the almost explosive increase in the size of the buildings. Uruk Levels VI to IV had rectangular buildings covering areas as large as 275 by 175 feet; they are described as temples, since the ground plans are comparable to those of later buildings whose sacred character is beyond doubt.

The major accomplishments of the period Uruk VI to IV, apart from the first inscribed tablets (Level IV B), are masterpieces of sculpture and of seal engraving and also of the form of wall decoration known as cone mosaics. Together with the everyday pottery of black or red burnished ware, there is a very coarse type known as the *Glockentopf* ("bell pot"). These are vessels of standard size whose shape served as the original for the sign *sila,* meaning "litre." It is not too rash to deduce from the mass production of such standard vessels that they served for the issue of rations. This would have been the earliest instance of a system that remained typical of the southern Mesopotamian city for centuries: the maintenance of part of the population by allocations of food from the state.

Historians usually date the beginning of history, as opposed to prehistory and protohistory, from the first appearance of usable written sources. If this is taken to be the transition from the 4th to the 3rd millennium BC, it must be remembered that this only applies to part of Mesopotamia: the South, the Diyālā region, Susiana (with a later script of its own invented locally), and the district of the Middle Euphrates. It does not apply to regions that remained without writing for a longer time, such as the Middle Tigris region (Assyria), Syria, Asia Minor, and Iran.

SUMERIAN CIVILIZATION

The Sumerians to the end of the Early Dynastic period. Despite the Sumerians' leading role, the historical role of other races should not be underestimated. While with prehistory only approximate dates can be offered, historical periods require a firm chronological framework, which, unfortunately, is not yet established for the first half of the 3rd millennium BC. The basis for the chronology after about 1450 BC is provided by the data in the Assyrian and Babylonian king lists, which can often be checked by dated tablets and the Assyrian lists of eponyms (annual officials whose names served to identify each year). But it is still uncertain how much time separated the middle of the 15th century BC from the end of the 1st dynasty of Babylon, which is therefore variously dated to 1594 BC ("middle"), 1530 BC ("short"), or 1730 BC ("long" chronology). As a compromise, the middle chronology will be used. From 1594 BC several chronologically overlapping dynasties reach back to the beginning of the 3rd dynasty of Ur, *c.* 2112 BC. From this point to the beginning of the dynasty of Akkad (*c.* 2334 BC) the interval can only be calculated to within 20–30 years, via the ruling houses of Lagash and the rather uncertain traditions regarding the succession of Gutian viceroys. With Ur-Nanshe (*c.* 2520 BC), the first king of the 1st dynasty of Lagash, there is a possible variation of 50–60 years, and earlier dates are a matter of mere guesswork: they depend upon factors of only limited relevance, such as the computation of occupation or destruction levels, the degree of development in the script (paleography), the character of the sculpture, pottery, and cylinder seals, and their correlation at different sites. In short, the chronology of the first half of the 3rd millennium is largely a matter for the intuition of the individual author. Carbon-14 dates are at present too few and far between to be given undue weight. Consequently, the turn of the 4th to 3rd millennium is to be accepted, with due caution and reservations, as the date of the flourishing of the archaic civilization of Uruk and of the invention of writing.

Problems in chronological reconstruction

In Uruk and probably also in other cities of comparable size, the Sumerians led a city life that can be more or less reconstructed as follows: temples and residential districts; intensive agriculture, stock breeding, fishing, and date-palm cultivation, forming the four mainstays of the economy; highly specialized industries, carried on by sculptors, seal engravers, smiths, carpenters, shipbuilders, potters, and workers of reeds and textiles. Part of the population was supported with rations from a central point of distribution, which relieved people of the necessity of providing their basic food themselves, in return for their work all day and every day, at least for most of the year. The cities kept up active trade with foreign lands.

That organized city life existed is chiefly demonstrated by the existence of inscribed tablets. The earliest tablets contain figures with the items they enumerate and measures with the items they measure, as well as personal names and, occasionally, probably professions. This shows the purely practical origins of writing in Mesopotamia: it did not begin as a means of magic or as a way for the ruler to record his achievements, for example, but as an aid to memory for an administration that was ever expanding its area of operations. The earliest examples of writing are very difficult to penetrate because of their extremely laconic formulation presupposing a knowledge of the context and because of the still very imperfect rendering of the spoken word. Moreover, many of the archaic signs were pruned away after a short period of use and cannot be traced in the paleography of later periods, so that they cannot be identified.

Complexities in unravelling the nature of the power structure

One of the most important questions that has to be met when dealing with "organization" and "city life" is that of social structure and the form of government. But it can only be answered with difficulty, and the use of evidence from later periods carries with it the danger of anachronisms. The Sumerian word for ruler *par excellence* is *lugal,* which etymologically means "big man." The first occurrence comes from Kish around 2700 BC, since an earlier instance from Uruk is uncertain because it could simply be intended as a personal name: "Monsieur Legrand." In Uruk the ruler's special title was *en.* In later periods this word (etymology unknown), which is also found in divine names such as En-lil and En-ki, has a predominantly religious connotation that is translated, for want of a better designation, as "en-priest, en-priestess." *En,* as the ruler's title, is encountered in the traditional epics of the Sumerians (Gilgamesh is the "*en* of Kullab," a district of Uruk) and particularly in personal names, such as "The-*en*-has-abundance," "The-*en*-occupies-the-throne," and many others.

It has often been asked if the ruler of Uruk is to be recognized in artistic representations. A man feeding sheep with flowering branches, who is a prominent personality in seal designs, could thus represent the ruler or a priest in his capacity as administrator and protector of flocks. The same question may be posed in the case of a man depicted on a stele aiming an arrow at a lion. But these questions are purely speculative: even if the "protector of flocks" were identical with the *en,* there is no ground for

seeing in the ruler a person with a predominantly religious function.

Literary and other historical sources. The picture offered by the literary tradition of Mesopotamia is clearer but not necessarily historically relevant. The Sumerian king list has long been the greatest focus of interest. This is a literary composition, dating from Old Babylonian times, that describes kingship (*nam-lugal* in Sumerian) in Mesopotamia from primeval times to the end of the 1st dynasty of Isin. According to the theory—or rather ideology—of this work, there was officially only one kingship in Mesopotamia, which was vested in one particular city at any one time; hence the change in dynasties brought with it the change of the seat of kingship:

Kish–Uruk–Ur–Awan–Kish–Hamazi–Uruk–Ur–
Adab–Mari–Kish–Akshak–Kish–Uruk–Akkad–
Uruk–Gutians–Uruk–Ur–Isin.

The king list gives as coming in succession several dynasties that now are known to have ruled simultaneously. It is a welcome aid to chronology and history, but, so far as the regnal years are concerned, it loses its value for the time before the dynasty of Akkad: for here the lengths of reign of single rulers are given as more than 100 and sometimes even several hundred years. One group of versions of the king list has adopted the tradition of the Sumerian flood story, according to which Kish was the first seat of kingship after the Flood, whereas five dynasties of primeval kings ruled before the Flood in Eridu, Bad-tibira, Larak, Sippar, and Shuruppak. These kings all allegedly ruled for multiples of 3,600 years (the maximum being 64,800 or, according to one variant, 72,000 years). The tradition of the Sumerian king list is still echoed in Berosus.

The Sumerian king list

It is also instructive to observe what the Sumerian king list does not mention. It lacks all mention of a dynasty as important as the 1st dynasty of Lagash (from King Ur-Nanshe to Uru'inimgina, formerly known as Urukagina) and appears to retain no memory of the archaic florescence of Uruk at the beginning of the 3rd millennium BC.

Besides the peaceful pursuits reflected in art and writing, the art also provides the first information about violent contacts: cylinder seals of the Uruk IV Level depict fettered men lying or squatting on the ground, being beaten with sticks or otherwise maltreated by standing figures. They may represent prisoners of war being killed in order to dispose of them. It is unknown from where these captives came or what form "war" would have taken or how early organized battles were fought. Nevertheless, this does give the first, albeit indirect, evidence for the wars that are henceforth one of the most characteristic phenomena in the history of Mesopotamia; and it makes it all the more surprising that the Mesopotamian cities at the beginning of the 3rd millennium BC were not walled. The earliest city wall so far known, that of Uruk, dates only to 2700–2650 BC.

Just as with the rule of man over man, with the rule of higher powers over man it is difficult to make any statements about the earliest attested forms of religion or about the deities and their names without running the risk of anachronism. Excluding prehistoric figurines, which provide no evidence for determining whether men or anthropomorphic gods are represented, the earliest testimony is supplied by certain symbols that later became the cuneiform signs for gods' names: the "gatepost with streamers" for Inanna, goddess of love and war, and the "ringed post" for the moon god Nanna. A scene on a cylinder seal—a shrine with an Inanna symbol and a "man" in a boat—could be an abbreviated illustration of a procession of gods or of a cultic journey by ship. The constant association of the "gatepost with streamers" with sheep and of the "ringed post" with cattle may possibly reflect the area of responsibility of each deity. The Sumerologist Thorkild Jacobsen sees in the pantheon a reflex of the various economies and modes of life in ancient Mesopotamia: fishermen and marsh dwellers, date-palm cultivators, cowherds, shepherds, and farmers all have their special groups of gods.

Early religious beliefs

Both Sumerian and non-Sumerian languages can be detected in the divine names and place-names. Since the pronunciation of the names is known only from 2000 BC or later, conclusions as to their linguistic affinity are not without problems. Several, for example, have been reinterpreted in Sumerian by popular etymology. It would be particularly important to isolate the Subarian components (related to Hurrian), whose significance was probably greater than has hitherto been assumed. For the south Mesopotamian city HA.A (the noncommittal transliteration of the signs) there is a pronunciation gloss shubari, and non-Sumerian incantations are known in the language of HA.A, which have turned out to be Subarian.

There have always been in Mesopotamia speakers of Semitic languages (which belong to the Hamito-Semitic group and also include ancient Egyptian, Berber, and various African languages). This element is easier to detect in ancient Mesopotamia, but whether people began to participate in city civilization in the 4th millennium BC or only during the 3rd is unknown. Over the last 4,000 years Semites (Amorites, Canaanites, Aramaeans, and Arabs) have been partly nomadic, ranging the Arabian fringes of the Fertile Crescent, and partly settled; and the transition to settled life can be observed in a constant, though uneven, rhythm. There are, therefore, good grounds for assuming that the Akkadians (and other pre-Akkadian Semitic tribes not known by name) also originally led a nomadic life to a greater or lesser degree. Nevertheless, they can only have been herders of domesticated sheep and goats, which require changes of pasturage according to the time of year and can never stray more than a day's march from the watering places. The traditional nomadic life of the Bedouin makes its appearance only with the domestication of the camel at the turn of the 2nd to 1st millennium BC.

The spread of writing

The question arises as to how quickly writing spread and by whom it was adopted about 3000 BC or shortly thereafter. At Kish, in north Babylonia, almost 120 miles northwest of Uruk, clay tablets have been found with the same repertoire of archaic signs as at Uruk itself. This fact demonstrates that intellectual contacts existed between north and south Babylonia. The dispersion of writing in an unaltered form presupposes the existence of schools in various cities that worked according to the same principles and adhered to one and the same canonical repertoire of signs. It would be wrong to assume that Sumerian was spoken throughout the area in which writing had been adopted: it is linguistically significant that even the earliest of the Uruk texts use a sexagesimal and a decimal system side-by-side. Moreover, the use of cuneiform for a non-Sumerian language can be demonstrated with certainty from the 27th century BC.

First historical personalities. The specifically political events in Mesopotamia after the flourishing of the archaic culture of Uruk cannot be pinpointed. Not until about 2700 BC does the first historical personality appear—historical because his name, Enmebaragesi (Me-baragesi), was preserved in later tradition. It has been assumed, although the exact circumstances cannot be reconstructed, that there was a rather abrupt end to the high culture of Uruk Level IV. The reason for the assumption is a marked break in both artistic and architectural traditions: cylinder seals were replaced by others with an entirely new style; the great temples were abandoned—flouting the rule of a continuous tradition on religious sites—and on a new site a shrine was built on a terrace, which was to constitute the lowest stage of the later Eanna ziggurat. On the other hand, since the writing system developed organically and was continually refined by innovations and progressive reforms, it would be overhasty to assume a revolutionary change in the population.

In the quarter or third of a millennium between Uruk Level IV and Enmebaragesi, southern Mesopotamia became studded with a complex pattern of cities, many of which were the centres of small independent city-states, to judge from the situation in about the middle of the millennium. In these cities, the central point was the temple, sometimes encircled by an oval boundary wall (hence the term temple oval); but nonreligious buildings, such as palaces serving as the residences of the rulers, could also function as centres.

Enmebaragesi, king of Kish, is the oldest Mesopotamian

Enmebaragesi

ruler about whom there are authentic inscriptions. These are vase fragments, one of them found in the temple oval of Khafājī. In the Sumerian king list, Enmebaragesi is listed as the penultimate king of the 1st dynasty of Kish; a Sumerian epic, *Gilgamesh and Agga of Kish,* describes the siege of Uruk by Agga, son of Enmebaragesi. The discovery of the original inscriptions was of great significance because it enabled scholars to ask with somewhat more justification whether Gilgamesh, the heroic figure of Mesopotamia who has entered world literature, had actually been a historical personage. The indirect synchronism notwithstanding, the possibility exists that even remote antiquity knew its "Ninos" and its "Semiramis," figures onto which a rapidly fading historical memory projected all manner of deeds and adventures. Thus, though the historical tradition of the early 2nd millennium believes Gilgamesh to have been the builder of the oldest city wall of Uruk, such may not have been the case. The palace archives of Shuruppak (modern Tall Fa'rah, 125 miles southeast of Baghdad), dating presumably from shortly after 2600, contain a long list of divinities, including Gilgamesh and his father Lugalbanda. More recent tradition, on the other hand, knows Gilgamesh as judge of the nether world. However that may be, an armed conflict between two Mesopotamian cities such as Uruk and Kish would hardly have been unusual in a country whose energies were consumed, almost without interruption from 2500 to 1500, by clashes between various separatist forces. The great "empires," after all, formed the exception, not the rule.

Emergent city-states. Kish must have played a major role almost from the beginning. After 2500, southern Babylonian rulers, such as Mesannepada of Ur and Eannatum of Lagash, frequently called themselves king of Kish when laying claim to sovereignty over northern Babylonia. This is not to agree with some recent histories in which Kish is represented as an archaic "empire." It is more likely to have figured as a representative of the north, calling forth perhaps the same geographical connotation later evoked by "the land of Akkad."

Although the corpus of inscriptions grows richer both in geographical distribution and in point of chronology in the 27th and increasingly so in the 26th century, it is still impossible to find the key to a plausible historical account. Besides, history cannot be written on the sole basis of archaeological findings. Unless clarified by written documents, these findings contain as many riddles as they seem to offer solutions. This applies even to as spectacular a discovery as that of the royal tombs of Ur with their hecatombs (large-scale sacrifices) of retainers who followed their king and queen to the grave, not to mention the elaborate funerary appointments with their inventory of tombs. It is only from *c.* 2520 to the beginnings of the dynasty of Akkad that history can be written within a framework, with the aid of reports about the city-state of Lagash and its capital of Girsu and its relations with its neighbour and rival, Umma.

Sources for this are, on the one hand, an extensive corpus of inscriptions relating to nine rulers, telling of the buildings they constructed, of their institutions and wars, and, in the case of Uru'inimgina, of their "social" measures. On the other hand, there is the archive of some 1,200 tablets—insofar as these have been published—from the temple of Baba, the city goddess of Girsu, from the period of Lugalanda and Uru'inimgina (first half of the 24th century). For generations, Lagash and Umma contested the possession and agricultural usufruct of the fertile region of Gu'edena. To begin with, some two generations before Ur-Nanshe, Mesilim (another "king of Kish") had intervened as arbiter and possibly overlord in dictating to both states the course of the boundary between them; but this was not effective for long. After a prolonged struggle Eannatum forced the ruler of Umma, by having him take an involved oath to six divinities, to desist from crossing the old border, a canal. The text that relates this event, with considerable literary elaboration, is found on the Stele of Vultures. These battles, favouring now one side, now the other, continued under Eannatum's successors, in particular Entemena, until, under Uru'inimgina, great damage was done to the land of Lagash and to its holy places. The enemy, Lugalzaggesi, was vanquished in turn by Sargon of Akkad. The rivalry between Lagash and Umma, however, must not be considered in isolation. Other cities, too, are occasionally named as enemies, and the whole situation resembles the pattern of changing coalitions and short-lived alliances between cities of more recent times. Kish, Umma, and distant Mari on the Middle Euphrates are listed together on one occasion as early as the time of Eannatum. For the most part, these battles were fought by infantry, although mention is also made of war chariots drawn by onagers (wild asses).

Intercity rivalry between Lagash and Umma

The lords of Lagash rarely fail to call themselves by the title of *ensi,* of as yet undetermined derivation; "city ruler," or "prince," would only be approximate translations. Only seldom do they call themselves *lugal,* or "king," the title given the rulers of Umma in their own inscriptions. In all likelihood, these were local titles that were eventually converted, beginning perhaps with the kings of Akkad, into a hierarchy in which the *lugal* took precedence over the *ensi.*

Territorial states. More difficult than describing its external relations is the task of shedding light on the internal structure of a state like Lagash. For the first time, a state consisting of more than a city with its surrounding territory came into being because aggressively minded rulers had managed to extend that territory until it comprised not only Girsu, the capital, and the cities of Lagash and Nina (Zurghul) but also many smaller localities and even a seaport, Guabba. But it is not clear to what extent the conquered regions were also integrated administratively. On one occasion Uru'inimgina uses the formula "from the limits of the Ningirsu [that is, the city god of Girsu] to the sea," having in mind a distance of up to 125 miles. It would be unwise to harbour any exaggerated notion of well-organized states exceeding that size.

The Sumerian temple city

For many years views were conditioned by the slogan of the Sumerian temple city, used to convey the idea of an organism whose ruler, as representative of his god, theoretically owned all land, privately held agricultural land being a rare exception. The image of the temple city had its origin partly in the overinterpretation of a passage in the so-called reform texts of Uru'inimgina, which state that "on the field of the *ensi* [or else, his wife and the crown prince], the city god Ningirsu [or else, the city goddess Baba and the divine couple's son]" had been "reinstated as owners." On the other hand, the statements in the archives of the temple of Baba in Girsu, dating from Lugalanda and Uru'inimgina, were held to be altogether representative. Here is a system of administration, directed by the *ensi*'s spouse or by a *sangu* (head steward of a temple), in which every economic process, including commerce, stands in a direct relationship to the temple: agriculture, vegetable gardening, tree farming, cattle raising and the processing of animal products, fishing, and the payment in merchandise of workers and employees.

The conclusion from this analogy proved to be dangerous because the archives of the temple of Baba provide information about only a portion of the total temple administration and that portion, furthermore, is limited in time. Understandably enough, the private sector, which of course was not controlled by the temple, is scarcely mentioned at all in these archives. The existence of such a sector is nevertheless documented by bills of sale for land purchases of the pre-Sargonic period and from various localities. Written in Sumerian as well as in Akkadian language, they prove the existence of private land ownership or, in the opinion of some scholars, of lands predominantly held as undivided family property. Although a substantial part of the population was forced to work for the temple and drew its pay and board from it, it is not yet known whether it was year-round work.

Conscript of labour

It is probable, if unfortunate, that there will never exist a detailed and numerically accurate picture of the demographic structure of a Sumerian city. It is assumed that in the oldest city the government was in a position to summon sections of the populace for the performance of public works. The construction of monumental buildings

or the excavation of long and deep canals could be carried out only by means of such a levy. The large-scale employment of indentured persons and of slaves are of no concern in this context. Evidence of male slavery is fairly rare before Ur III, and even in Ur III and in the Old Babylonian period slave labour was never an economically relevant factor. It was different with female slaves. According to one document, the temple of Baba employed 188 such women; the temple of the goddess Nanshe, 180, chiefly in grinding flour and in the textile industry; and this continued to be the case in later times. For accuracy's sake it should be added that the terms male slave and female slave are used here in the significance they possessed around 2000 and later, designating persons in bondage who were bought and sold and who could not acquire personal property through their labour. A distinction is made between captured slaves (prisoners of war and kidnapped persons) and others who had been sold.

In one inscription, Entemena of Lagash boasts of having "allowed the sons of Uruk, Larsa, and Bad-tibira to return to their mothers" and of having "restored them into the hands" of the respective city god or goddess. Read in the light of similar but more explicit statements of later date, this laconic formula represents the oldest known evidence of the fact that the ruler occasionally endeavoured to mitigate social injustices by means of a decree. Such decrees might refer to the suspension or complete cancellation of debts or to exemption from public works. Whereas a set of inscriptions of the last ruler from the 1st dynasty of Lagash, Uru'inimgina (possibly the correct reading in lieu of the former "Urukagina"), has long been considered a prime document of social reform in the 3rd millennium, the designation "reform texts" is only partly justified. Reading between the lines, it is possible to discern that tensions had arisen between the "palace"—the ruler's residence with its annex, administrative staff, and landed properties—and the "clergy"; that is, the stewards and priests of the temples. In seeming defiance of his own interests, Uru'inimgina, who in contrast to practically all of his predecessors lists no genealogy and has therefore been suspected of having been a usurper, defends the clergy, whose plight he describes somewhat tearfully. If the foregoing passage about restoring the *ensi*'s fields to the divinity is interpreted carefully, it would follow that the situation of the temple was ameliorated and that palace lands were assigned to the priests. Along with these measures, which resemble the policies of a newcomer forced to lean on a specific party, are found others that do merit the designation of "measures taken toward the alleviation of social injustices"—for instance, the granting of delays in the payment of debts or their outright cancellation and the setting up of prohibitions to keep the economically or socially more powerful from forcing his inferior to sell his house, his ass's foal, and the like. Besides this, there were tariff regulations, such as newly established fees for weddings and burials, as well as the precise regulation of the food rations of garden workers, called blind men.

These conditions, described on the basis of source materials from Girsu, may well have been paralleled elsewhere; but it is equally possible that other archives, yet to be found in other cities of pre-Sargonic southern Mesopotamia, may furnish entirely new historical aspects. At any rate, it is wiser to proceed cautiously, by keeping to analysis and evaluation of the available material rather than to construct generalizations.

This, then, is the horizon of Mesopotamia shortly before the rise of the Akkadian Empire. In Mari, writing was introduced at the latest around the mid-26th century BC; and from that time this city, situated on the Middle Euphrates, forms an important centre of cuneiform civilization, especially in regard to its Semitic component. Reaching out across the Diyālā region and the Persian Gulf, Mesopotamian influences extended to Iran, where Susa is mentioned along with Elam and other, not yet localized, towns. In the west, the Amanus Mountains were known; and under Lugalzaggesi, the "upper sea" in other words, the Mediterranean, is mentioned for the first time. To the east, the inscriptions of Ur-Nanshe of Lagash already name the isle of Tilmun (modern Bahrain), which may have been even then a transshipment point for trade with the Oman coast and the Indus region, the Magan and Meluhha of more recent texts. Trade with Asia Minor and Afghanistan was nothing new to the 3rd millennium, even if these regions are not yet listed by their names. It was the task of the Akkadian dynasty to unite, within these boundaries, a territory that transcended the dimensions of a state of the type represented by Lagash.

Sumer and Akkad in 2350–2000 BC. There are several reasons for taking the year 2350 as a turning point in the history of Mesopotamia. For the first time, an empire arose on Mesopotamian soil. The driving force of that empire was the "Akkadians," so called after the city of Akkad, which Sargon chose for his capital (it has not yet been identified but was presumably located on the Euphrates between Sippar and Kish). The name Akkad becomes synonymous with a population group that stands side-by-side with the Sumerians. The "land of Sumer and Akkad" becomes a name for southern Mesopotamia; "Akkadian" becomes the name of a language; and the arts rise to new heights. But even this turning point was not the first time the Akkadians had emerged in history. Semites—whether Akkadians or a Semitic language group that had settled before them—perhaps had had a part in the urbanization that took place at the end of the 4th millennium. The earliest Akkadian names and words occur in written sources of the 27th century. The names of several Akkadian scribes are found in the archives of Tell Abu Ṣalābīkh near Sippar, synchronous with those of Shuruppak (shortly after 2600). The Sumerian king list places the 1st dynasty of Kish, together with a series of kings bearing Akkadian names, immediately after the Flood. In Mari, Akkadian script was probably used from the very beginning. Thus, the founders of the dynasty of Akkad were presumably members of a people who had been familiar for centuries with Mesopotamian culture in all its forms.

Sargon's reign. According to the Sumerian king list, the first five rulers of Akkad (Sargon, Rimush, Manishtusu, Naram-Sin, and Shar-kali-sharri) ruled for a total of 142 years; Sargon alone ruled for 56. Although these figures cannot be checked, they are probably trustworthy, because the king list for Ur III, even if 250 years later, did transmit dates that proved to be accurate.

As stated in an annotation to his name in the king list, Sargon started out as a cupbearer to King Ur-Zababa of Kish. There is also an Akkadian legend about Sargon, who is supposed to have been exposed after birth, brought up by a gardener, and beloved of the goddess Ishtar. Nevertheless, there are no historical dates about his career. Yet it is feasible to assume that in his case a high court office served as springboard for a dynasty of his own. The original inscriptions of the kings of Akkad that have come down to posterity are brief, more informative for geographical distribution than content. The main source material for Sargon's reign, with its high points and catastrophes, are copies made by Old Babylonian scribes in Nippur, from the very extensive originals that presumably had been kept there. They are in part Akkadian, in part bilingual Sumerian-Akkadian texts. According to these texts, Sargon is said to have fought against the Sumerian cities of southern Babylonia, thrown down city walls, taken prisoner 50 *ensi*s, and have "cleansed his weapons in the sea." He is also supposed to have captured Lugalzaggesi of Uruk, the former ruler of Umma, who had vigorously attacked Uru'inimgina in Lagash, forcing his neck under a yoke and leading him thus to the gate of the god Enlil at Nippur. "Citizens of Akkad" filled the offices of *ensi* from the "nether sea" (the Persian Gulf) upward, which was perhaps a device used by Sargon to further his dynastic aims. Aside from the 34 battles fought in the south, Sargon also tells of conquests in northern Mesopotamia: Mari, Tuttul on the Balīkh, where he venerated the god Dagan (Dagon), Ebla (Tall Mardīkh of Syria), the "cedar forest" (Amanus or Lebanon), and the "silver mountains"; battles in Elam and the foreshore of the Zagros are mentioned. Sargon also relates that ships from Meluhha (Indus region), Magan (possibly the coast of Oman), and Tilmun (Bahrain) made fast in the port of Akkad.

Source materials on Sargon

Impressive as they are at first sight by way of a terse synopsis, these reports have only a limited value because they cannot be arranged chronologically, and it is not known whether Sargon built a large empire. Akkadian tradition itself saw it in this light, however, and a learned treatise of the late 8th or the 7th century lists no fewer than 65 cities and lands belonging to that empire. But even if Magan and Kapturu (Crete) are given as the eastern and western limits of the conquered territories, it is impossible to transpose this to the 3rd millennium.

Sargon's daughter

Sargon appointed one of his daughters priestess of the moon god in Ur. She took the name of Enkheduana and was succeeded in the same office by Enmenana, a daughter of Naram-Sin. Enkheduana must have been a very gifted woman; two Sumerian hymns by her have been preserved, and she is also said to have been instrumental in starting a collection of songs dedicated to the temples of Babylonia.

Sargon died in very old age. The inscriptions, likewise preserved only in copies, of his son Rimush are full of reports about battles fought in Sumer and Iran, just as if there had never been a Sargonic empire. It is not known in detail how rigorously Akkad wished to control the cities to the south and how much freedom had been left to them; but they presumably clung tenaciously to their inherited local autonomy. From a practical point of view, it was probably in any case impossible to organize an empire that would embrace all of Mesopotamia.

Since the reports (that is, copies of inscriptions, as was their later elaboration into legend) left by Manishtusu, Naram-Sin, and Shar-kali-sharri speak time and again of rebellions and victorious battles and since Rimush, Manishtusu, and Shar-kali-sharri are themselves said to have died violent deaths, the problem of what remained of Akkad's greatness obtrudes. Wars and disturbances, the victory of one and the defeat of another, and even regicide constitute only one of the aspects suggested to us by the sources. Whenever they extended beyond the immediate Babylonian neighbourhood, the military campaigns of the Akkadian kings were dictated primarily by trade interests instead of being intended to serve the conquest and safeguarding of an empire. Akkad, or more precisely, the king, needed merchandise, money, and gold in order to finance wars, buildings, and the system of administration that he had instituted.

On the other hand, the original inscriptions that have been found so far of a king like Naram-Sin are scattered over a distance of some 620 miles as the crow flies, following the Tigris downriver: Diyarbakir on the Upper Tigris, Nineveh, Tall Birāk on the upper Khābūr (which had an Akkadian fortress and garrison), Susa in Elam, as well as Marad, Puzurish-Dagan, Adab (Bismāyah), Nippur, Ur, and Girsu in Babylonia. Even if all this was not part of an empire, it surely constituted an impressive sphere of influence.

Divine kingship

Also to be considered are other facts that weigh more heavily than high-sounding reports of victories that cannot be checked. After the first kings of the dynasty had borne the title of king of Kish, Naram-Sin had himself addressed as "king of the four quarters of the earth"; that is, of the universe. As if he were in fact divine, he also had his name written with the cuneiform sign "god," the divine determinative that was customarily used in front of the names of gods; furthermore, he assumed the title of "god of Akkad." It is legitimate to ask whether the concept of deification may be used in the sense of elevation to a rank equal to that of the gods. At the very least it must be acknowledged that in relation to his city and his subjects, the king saw himself in the role played by the local divinity as protector of the city and guarantor of its well-being. In synchronous judicial documents from Nippur, the oath is often taken "by Naram-Sin," with a formula identical with that used in swearing by a divinity. Documents from Girsu contain Akkadian date formulas of the type "in the year in which Naram-Sin laid the foundations of the Enlil temple at Nippur and of the Inanna temple at Zabalam." As evidenced by the dating procedures customary in Ur III and in the Old Babylonian period, the use of such formulas presupposes that the respective city acknowledged as its overlord the ruler whose name is invoked.

Ascendancy of Akkad. Under Akkad, the Akkadian language acquired a literary prestige that made it the equal of Sumerian; under the influence, perhaps, of an Akkadian garrison at Susa, it spread beyond the borders of Mesopotamia. After having employed for several centuries an indigenous script patterned after cuneiform writing, Elam adopted Mesopotamian script during the Akkadian period and with a few exceptions used it even when writing in Elamite rather than Sumerian or Akkadian. The so-called Old Akkadian manner of writing is extraordinarily appealing from the aesthetic point of view; as late as in the Old Babylonian era, it served as a model for monumental inscriptions. This bears out the fact that the plastic and graphic arts, especially sculpture in the round, relief work, and cylinder seals reached a high point of perfection.

Thus the reign of the five kings of Akkad may be considered one of the most productive periods of Mesopotamian history. Although separatist forces opposed all unifying tendencies, Akkad brought about a broadening of political horizons and dimensions. The period of Akkad fascinated historiographers as did few other eras; having contributed its share to the storehouse of legend, it has never disappeared from memory. With a phrase such as "There will come a king of the four quarters of the earth," liver omens (sooth-saying done by analyzing the shape of a sheep's liver) of the Old Babylonian period express the yearning for unity at a time when Babylonia had once again disintegrated into a dozen or more small states.

The end of the dynasty. Of the kings after Shar-kalisharri (*c.* 2217–*c.* 2193), only the names and a few brief inscriptions have survived. Quarrels arose over the succession, and the dynasty went under, although modern scholars know as little about the individual stages of this decline as about the rise of Akkad. Two factors contributed to its downfall: the invasion of the nomadic Amurrus (Amorites), called Martu by the Sumerians, from the northwest, and the infiltration of the Gutians, who came, apparently, from the region between the Tigris and the Zagros Mountains eastward. This argument, however, may be a vicious circle, as these invasions were provoked and facilitated by the very weakness of Akkad. In Ur III, the Martu, in part already sedentary, form one ethnic component along with Sumerians and Akkadians. The Gutians, on the other hand, played only a temporary role, even if the memory of a Gutian dynasty persisted until the end of the 17th century BC. As a matter of fact, the wholly negative opinion that even some modern historians have of the Gutians is based solely on a few stereotyped statements by the Sumerians and Akkadians, especially on the victory inscription of Utu-khegal of Uruk (2116–10). While Old Babylonian sources give the region between the Tigris and the Zagros Mountains as the home of the Gutians, these people probably also lived on the Middle Euphrates during the 3rd millennium. According to the Sumerian king list, the Gutians held the "kingship" in southern Mesopotamia for about 100 years. It has long been recognized that there is no question of a whole century of undivided Gutian rule and that some 50 years of this rule coincided with the final half century of Akkad. From this period there has also been preserved a record of a "Gutian interpreter." As it is altogether doubtful whether the Gutians had made any city of southern Mesopotamia their "capital" instead of controlling Babylonia more or less informally from outside, scholars cautiously refer to "viceroys" of this people. The Gutians have left no material records, and the original inscriptions about them are so slight that no binding statements about them are possible.

The Gutians' obscure role

The Gutians' influence probably did not extend beyond Umma. The neighbouring state of Lagash enjoyed a century of complete independence, between Shar-kalisharri and Ur III, during which time it showed expansionist tendencies and had widely ranging trade connections. Of the *ensi* Gudea (*c.* 2144–*c.* 2124) there are extant writings, exclusively Sumerian in language, which are of inestimable value. He had the time, power, and means to carry out an extensive program of temple construction during his reign; and in a hymn divided into two parts and preserved in two clay cylinders 12 inches (30 centimetres) high he describes explicitly the recon-

struction of Eninnu, the temple of the god Ningirsu. Comprising 1,363 lines, the text is second in length only to Eannatum's Stele of Vultures among the literary works of the Sumerians up to that time. While Gudea forges a link, in his literary style, with his country's pre-Sargonic period, his work also bears the unmistakable stamp of the period of Akkad. Thus, the regions that furnish him building materials reflect the geographical horizon of the empire of Akkad, and the *ensi*'s title "god of his city" recalls the "god of Akkad" (Naram-Sin). The building hymn contains interesting particulars about the work force deployed. "Levies" were organized in various parts of the country, and city of Girsu itself "followed the *ensi* as though it were a single man." Unfortunately, synchronous administrative archives of sufficient length to provide less summarily compiled information about the social structure of Lagash shortly before the 3rd dynasty of Ur are lacking. After the great pre-Sargonic archives of the Baba temple at Girsu, only the various administrative archives of the king of Ur III give a closer look at the functioning of a Mesopotamian state.

The 3rd dynasty of Ur. Ut-khegal of Uruk is given credit for having overthrown Gutian rule by vanquishing their king Tiriqan along with two generals. Utukhegal calls himself lord of the four quarters of the earth in an inscription; but this title, adopted from Akkad, is more likely to signify political aspiration than actual rule. If another inscription has been correctly restored, Utu-khegal was a brother of the Ur-Nammu who founded the 3rd dynasty of Ur ("3rd" because it is the third time that Ur is listed in the Sumerian King List). Under Ur-Nammu and his successors Shulgi, Amar-Su'en (Amar-Sin), Shu-Sin, and Ibbi-Sin, this dynasty lasted for a century (*c.* 2112–2004). Ur-Nammu was at first "governor" of the city of Ur under Utu-khegal. How he became king is not known, but there may well be some parallels between his rise and the career of Ishbi-Erra of Isin or, indeed, that of Sargon. By eliminating the state of Lagash, Ur-Nammu caused the coveted overseas trade (Tilmun, Magan, and Meluhha) to flow through Ur. As evidenced by a new royal title that he was the first to bear—that of "king of Sumer and Akkad"—he had built up a state that comprised at least the southern part of Mesopotamia. Like all great rulers, he built much, including the very impressive ziggurats of Ur and Uruk, which acquired their final monumental dimensions in his reign.

The Codex of Ur-Nammu

Assyriologists have given the name of Codex of Ur-Nammu to a literary monument that is the oldest hitherto-known example of a genre extending through the Code of Lipit-Ishtar in Sumerian to the Code of Hammurabi, written in Akkadian. It is a collection of sentences or verdicts mostly following the pattern of "If A [assumption], it follows that B [legal consequence]." The collection is framed by a prologue and an epilogue. The original was most likely a stele, but all that is known of the Code of Ur-Nammu so far are Old Babylonian copies. The term code as used here is conventional terminology and should not give the impression of any kind of "codified" law; furthermore, the content of the Code of Ur-Nammu is not yet completely known. It deals, among other things, with adultery by a married woman, the defloration of someone else's female slave, divorce, false accusation, the escape of slaves, bodily injury, the granting of security, as well as with legal cases arising from agriculture and irrigation.

Before its catastrophic end, under Ibbi-Sin, the state of Ur III does not seem to have suffered setbacks and rebellions as grievous as those experienced by Akkad. There are no clear indications pointing to inner unrest, although it must be remembered that the first 20 years of Shulgi's reign are still hidden in darkness. But from that point on until the beginning of Ibbi-Sin's reign, or for a period of 50 years at least, the sources give the impression of peace enjoyed by a country that lived undisturbed by encroachments from abroad. Some expeditions were sent into foreign lands, to the region bordering on the Zagros, to what later became Assyria, and to the vicinity of Elam, in order to secure the importation of raw materials, in a fashion reminiscent of Akkad. Force seems to have been employed only as a last resort; and every attempt was made to bring about peaceful conditions on the other side of the border through the dispatch of embassies or the establishment of family bonds, for example, by marrying the king's daughters to foreigners.

Shulgi, too, called himself king of the four quarters of the earth. Although he resided in Ur, another important centre was in Nippur, whence—according to the prevailing ideology—Enlil, the chief god in the Sumerian state pantheon, had bestowed on him the royal dignity. Shulgi and his successors enjoyed divine honours, as Naram-Sin of Akkad had before them; by now, however, the process of deification had taken on clearer outlines in that sacrifices were offered and chapels built to the king and his throne, while the royal determinative turned up in personal names. Along with an Utu-khegal (The Sun God Is Exuberance) there appears a Shulgi-khegal (Shulgi Is Exuberance), and so forth.

Administration. The highest official of the state was the *sukkal-makh,* literally "supreme courier," whose position may be paraphrased with "(state) chancellor." The empire was divided into some 40 provinces ruled by as many *ensi*s, who, despite their far-reaching authority (civil administration and judicial powers), were no longer autonomous, even if only indirectly, although the office was occasionally handed down from father to son. They could not enter into alliances or wage wars on their own. The *ensi*s were appointed by the king and could probably also be transferred by him to other provinces. Each of these provinces was obliged to pay a yearly tribute, the amount of which was negotiated by emissaries. Of special significance in this was a system called *bala,* "cycle" or "rotation," in which the *ensi*s of the southern provinces took part; among other things, they had to keep the state stockyards supplied with sacrificial animals. Although the "province" often corresponded to a former city-state, many others were no doubt newly established. The so-called land-register text of Ur-Nammu describes four such provinces north of Nippur, giving the precise boundaries and ending in each case with the statement: "King Ur-Nammu has confirmed the field of the god XX for the god XX." In some cities, notably in Uruk, Mari, or Dēr (near Badrah, Iraq), the administration was in the hands of a *shakana,* a man whose title is rendered partly by "governor" and partly by "general."

The available histories are practically unanimous in seeing in Ur III a strongly centralized state marked by the king's position as absolute ruler. Nevertheless, some caution is indicated. For one thing, the need to deal as tactfully as possible with the *ensi*s must not be underestimated. A further question arises from the borders between and relative extent of the "public" and the "private" sector; the latter's importance may have been underrated as well. What is meant by "private" sector is a population group with land of its own and with revenues not directly granted by a temple or a "palace," such as by the king's or an *ensi*'s household. The traditional picture is derived from the sources, the state archives of Puzurish-Dagan, a gigantic "stockyard" situated outside the gates of Nippur, which supplied the city's temples with sacrificial animals but inevitably also comprised a major wool and leather industry; other such archives are those of Umma, Girsu, Nippur, and Ur. All these activities were overseen by a finely honed bureaucracy that stressed the use of official channels, efficient administration, and precise accounting. The various administrative organs communicated with one another by means of a smoothly functioning network of messengers. Although almost 23,000 documents referring to the economy of Ur III have so far been published, they are still waiting to be properly evaluated. Nor is there yet a serviceable typology for them; only when that has been drawn up will it be possible to write a book entitled "*The Economic System of Ur III.*" Represented in the main by contracts (loans, leases of temple land, the purchase of slaves, and the like), the "private" sector makes up only a small part of this mass of textual material. Neither can the sites at which discoveries have been made so far be taken as representative; in northern Babylonia, for exam-

The bureaucracy

ple, scarcely any contemporary written documents have yet been recovered.

Ethnic, geographic, and intellectual constituents. From the ethnic point of view, Mesopotamia was as heterogeneous at the end of the 3rd millennium as it had been earlier. The Akkadian element predominated, and the proportion of speakers of Akkadian to speakers of Sumerian continued to change in favour of the former. Some scholars consider it altogether wrong to speak of "Sumerians" and "Akkadians" at this point, since the Sumerians had been living in southern Mesopotamia for at least 1,000 years, without any contact with possibly related peoples, whereas the Semitic population regenerated itself continually from the ranks of nomads who became sedentary. The third group, first mentioned under Sharkali-sharri of Akkad, are the Amorites. In Ur III, some members of this people are already found in the higher echelons of the administration, while the bulk, organized in tribes, still led a nomadic life. Their great days came in the Old Babylonian period. While clearly differing linguistically from Akkadian, the Amorite language, which can be reconstructed to some extent from hundreds of proper names, is fairly closely related to the so-called Canaanite branch of the Semitic languages, of which it may in fact represent an older form. The fact that King Shu-Sin had a regular wall built clear across the land, the "wall that keeps out the Tidnum" (name of a tribe), shows how strong the pressure of the nomads was in the 21st century and what efforts were being made to check their influx. The fourth major ethnic group comprised the Hurrians, who were especially important in northern Mesopotamia and in the vicinity of modern Kirkūk.

It is likely that the geographical horizon of the empire of Ur III did not materially exceed that of the empire of Akkad. No names of localities in the interior of Asia Minor have been found, but there was much coming and going of messengers between Mesopotamia and Iran, far beyond Elam. There is also one mention of Gubla (Byblos) on the Mediterranean coast. Oddly enough, there is no evidence of any relations with Egypt, either in Ur III or in the Old Babylonian period. It is odd if no contacts existed at the end of the 3rd millennium between the two great civilizations of the ancient Near East.

Literature of Ur III

Intellectual life at the time of Ur III must have been very active in the cultivation and transmission of older literature, as well as in new creations. Although its importance as a spoken tongue was diminishing, Sumerian still flourished as a written language, a state of affairs that continued into the Old Babylonian period. As shown by the hymn on the deified king, new literary genres arose in Ur III. If Old Babylonian copies are any indication, the king's correspondence with leading officials was also of a high literary level.

In the long view, the 3rd dynasty of Ur did not survive in historical memory as vigorously as did Akkad. To be sure, Old Babylonian historiography speaks of Ur III as *bala-Shulgi,* the "(reigning) cycle of Shulgi"; but there is nothing that would correspond to the epic poems about Sargon and Naram-Sin. The reason is not clear, but it is conceivable that the later, purely Akkadian population felt a closer identification with Akkad than with a state that to a large extent still made use of the Sumerian language.

Ur III in decline. The decline of Ur III is an event in Mesopotamian history that can be followed in greater detail than other stages of that history. There are some clear sources, such as the royal correspondence, two elegies on the destruction of Ur and Sumer, and an archive from Isin that shows how Ishbi-Erra, as usurper and king of Isin, eliminated his former overlord in Ur. Ibbi-Sin was waging war in Elam when an ambitious rival came forward in the person of Ishbi-Erra from Mari, presumably a general or high official. By emphasizing to the utmost the danger threatening from the Amorites, Ishbi-Erra urged the King to entrust to him the protection of the neighbouring cities of Isin and Nippur. Ishbi-Erra's demand came close to extortion, and his correspondence shows how skillfully he dealt with the Amorites and with individual *ensi*s, some of whom soon went over to his side. Ishbi-Erra also took advantage of the depression that the King suffered because the god Enlil "hated him": a phrase presumably referring to bad omens resulting from the examination of sacrificed animals, on which procedure many rulers based their actions (or, as the case may be, their inaction). Ishbi-Erra fortified Isin and, in the 10th year of Ibbi-Sin's reign, began to employ his own dating formula on documents, an act tantamount to a renunciation of loyalty. Ishbi-Erra, on his part, believed himself to be the favourite of Enlil, the more so as he ruled over Nippur, where the god had his sanctuary. In the end, he claimed suzerainty over all of southern Mesopotamia, including Ur.

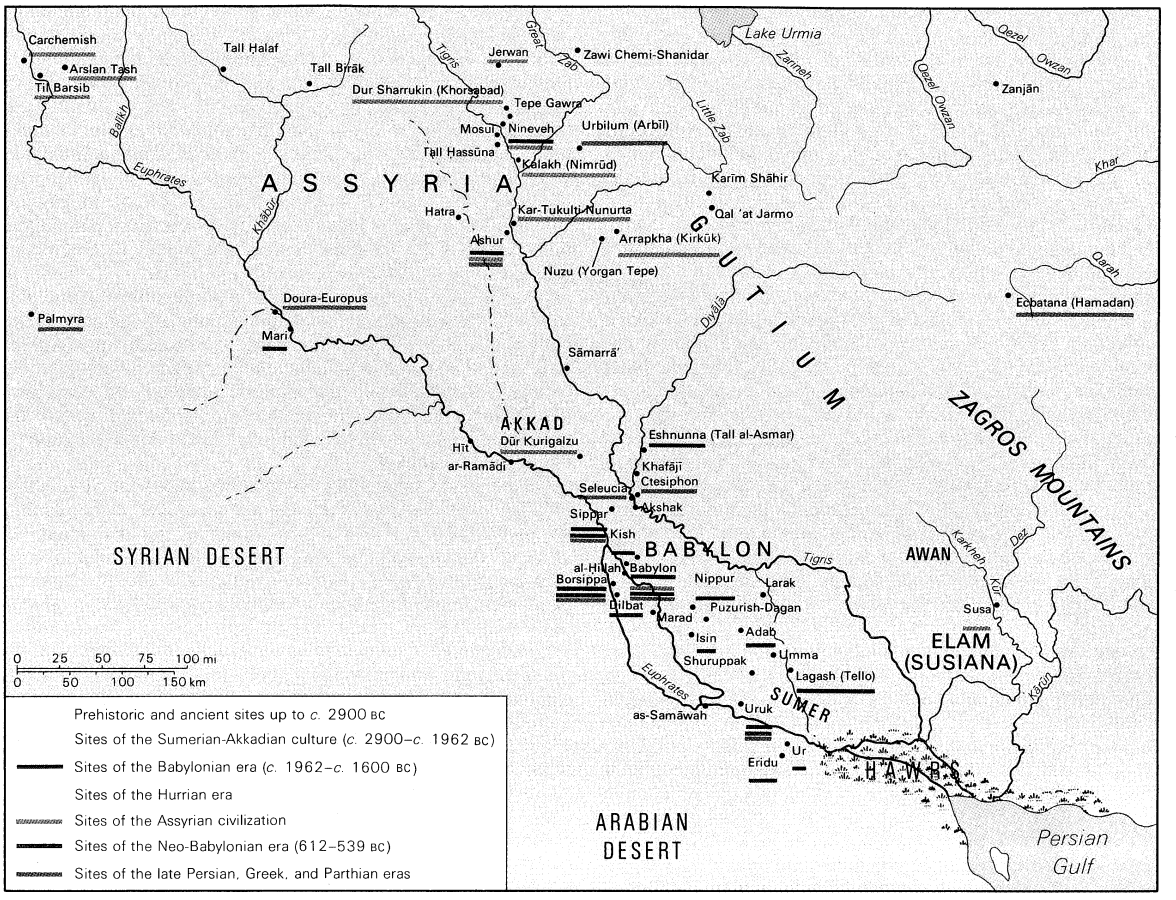
While Ishbi-Erra purposefully strengthened his domains, Ibbi-Sin continued for 14 more years to rule over a decreasing portion of the land. The end of Ur came about through a concatenation of misfortunes: a famine broke out; and Ur was besieged, taken, and destroyed by the invading Elamites and their allies among the Iranian tribes. Ibbi-Sin was led away captive, and no more was heard of him. The elegies record in moving fashion the unhappy end of Ur, the catastrophe that had been brought about by the wrath of Enlil.

Ur's end

THE OLD BABYLONIAN PERIOD

Isin and Larsa. During the collapse of Ur III, Ishbi-Erra established himself in Isin and founded a dynasty there that lasted from 2017 to 1794; his example was followed elsewhere by local rulers, as in Dēr, Eshnunna, Sippar, Kish, and Larsa. In many localities an urge was felt to imitate the model of Ur; Isin probably took over unchanged the administrative system of that state. Ishbi-Erra and his successors had themselves deified, as did one of the rulers of Dēr on the Iranian border. For almost a century Isin predominated within the mosaic of states that were slowly reemerging. Overseas trade revived after Ishbi-Erra had driven out the Elamite garrison from Ur, and under his successor, Shu-ilishu, a statue of the moon god Nanna, the city god of Ur, was recovered from the Elamites who had carried it off. Up to the reign of Lipit-Ishtar (*c.* 1934–*c.* 1924), the rulers of Isin so resembled those of Ur, so far as the King's assessment of himself in the hymns is concerned, that it seems almost arbitrary to postulate a break between Ibbi-Sin and Ishbi-Erra. As a further example of continuity it might be added that the Code of Lipit-Ishtar stands exactly midway chronologically between the Code of Ur-Nammu and the Code of Hammurabi. Yet it is much closer to the former in language and especially in legal philosophy than to Hammurabi's compilation of judgments. For example, the Code of Lipit-Ishtar does not know the lex talionis (or "an eye for an eye and a tooth for a tooth"), the guiding principle of Hammurabi's penal law.

Political fragmentation. It is probable that the definitive separation from Ur III came about through changing components of the population, from "Sumerians and Akkadians" to "Akkadians and Amorites." An Old Babylonian liver omen states that "He of the steppes will enter, and chase out the one in the city." This is indeed an abbreviated formula for an event that took place more than once: the usurpation of the king's throne in the city by the "sheikh" of some Amorite tribe. These usurpations were regularly carried out as part of the respective tribes became settled, although this was not so in the case of Isin because the house of Ishbi-Erra came from Mari and was of Akkadian origin, to judge by the rulers' names. By the same linguistic token the dynasty of Larsa was Amorite. The fifth ruler of the latter dynasty, Gungunum (ruled *c.* 1932–*c.* 1906), conquered Ur and established himself as the equal and rival of Isin; at this stage—the end of the 20th century BC—if not before, Ur had certainly outlived itself. From Gungunum until the temporary unification of Mesopotamia under Hammurabi, the political picture was determined by the disintegration of the balance of power, by incessant vacillation of alliances, by the presumption of the various rulers, by the fear of encroachments by the Amorite nomads, and by increasingly wretched social conditions. The extensive archive of correspondence from the royal palace of Mari (*c.* 1810–1750) is the best source of information about the political and diplomatic game and its



Sites associated with ancient Mesopotamian history.

rules, whether honoured or broken: treaties, the dispatch and reception of embassies, agreements about the integration of allied armies, espionage, "situation reports" from "foreign" courts. Devoid of exaggeration or stylization, these letters, dealing as they do with everyday events, are preferable to the numerous royal inscriptions on buildings even when the latter contain historical allusions.

Literary texts and increasing decentralization. Another indirect but far from negligible source for the political and socioeconomic situation in the 20th–18th centuries BC is the literature of omens. These are long compendiums in which the condition of a sheep's liver or some other divinatory object (for instance, the behaviour of a drop of oil in a beaker filled with water, the appearance of a newborn baby, and the shape of rising clouds of incense) are described at length and commented on with the appropriate prediction: "The king will kill his dignitaries and distribute their houses and property among the temples"; "A powerful man will ascend the throne in a foreign city"; "The land that rose up against its 'shepherd' will continue to be ruled by that 'shepherd' "; "The king will depose his chancellor"; and "They will lock the city gate and there will be a calamity in the city."

The literature of omens as a source

Beginning with Gungunum of Larsa, the texts allow greater insight into the private sector than in any other previous period. There is a considerable increase in the number of private contracts and private correspondences. Especially frequent among the private contracts are those concluded about loans of silver or grain (barley), illustrating the common man's plight, especially when driven to seek out a creditor, the first step on a road that in many instances led to ruin. The rate of interest, murderously high at 20 percent in the case of silver and 33 percent in that of grain, increased further if the deadline for repayment, usually at harvest time, was not kept. Insolvency resulted in imprisonment for debt, slavery by mortagage, and the sale of children and even the debtor's own person. Many private letters contain entreaties for the release of family members from imprisonment at the creditor's hands. Yet considerable fortunes were also made, in "liquid" capital as well as landed property. As these tendencies threatened to end in economic disaster, the kings prescribed as a corrective the liquidation of debts, by way of temporary alleviation at least. The exact wording is known of one such decree from the time of Ammisaduqa of Babylon.

Until the Ur III period, the only archives so far recovered dealt with temples or the palace. But belonging to the Old Babylonian period, along with documents pertaining to civil law, were an increasing number of administrative records of privately managed households, inns, and farms: settlements of accounts, receipts, and notes on various transactions. Here was clearly a regular bourgeoisie, disposing of its own land and of means independent of temple and palace. Trade, too, was now chiefly in private hands; the merchant travelled (or sent his partners) at his own risk, not on behalf of the state. Among the civil-law contracts there was a substantial increase in records of land purchases. Also significant for the economic situation in the Old Babylonian era was a process that might be summarized as "secularization of the temples," even if all the stages of this development cannot be traced. The palace had probably possessed for centuries the authority to dispose of temple property, but, whereas Uru'inimgina of Lagash had still branded the tendency as leading to abuses, the citizen's relationship to the temple now took on individual traits. Revenues from certain priestly offices—benefices, in other words—went to private individuals and were sold and inherited. The process had begun in Ur, where the king bestowed benefices, although the recipients could not own them. The archives of the "canonesses" of the sun god of Sippar furnish a particularly striking example of the fusion of religious service and private economic interest. These women, who lived in a convent called *gagûm,* came from the city's leading families and were not allowed to marry. With their property, consisting of land and silver, they engaged in a lively and remunerative business by granting loans and leasing out fields.

Changing relationship between religious institutions and individuals

The tendency toward decentralization had begun in the Old Babylonian period with Isin. It concluded with the 72-year reign of the house of Kudur-Mabuk in Larsa (*c.* 1834–*c.* 1763). Kudur-mabuk, sheikh of the Amorite tribe of the Jamutbal, despite his Elamite name, helped his son Warad-Sin to secure the throne. This usurpation allowed Larsa, which had passed through a period of internal unrest, to flourish one more time. Under Warad-Sin and in the long reign of his brother Rim-Sin, large portions of southern Babylonia, including Nippur, were once again united in one state of Isin in 1794. Larsa on its part was conquered by Hammurabi in 1763.

Early history of Assyria. In contrast to southern Mesopotamia or the mid-Euphrates region (Mari), written sources in Assyria do not begin until very late, shortly before Ur III. By Assyria—a region that does not lend itself to precise geographical delineation—is understood the territory on the Tigris north of the river's passage through the mountains of the Jabal Ḥamrīn to a point north of Nineveh, as well as the area between Little and Great Zab (a tributary of the Tigris in northeast Iraq) and to the north of the latter. In the north, Assyria was later bordered by the mountain state of Urartu; to the east and southeast, its neighbour was the region around the old Nuzu (near modern Kirkūk, "Arrapachitis" [Arrapkha] of the Greeks). In the early 2nd millennium the main cities of Assyria were Ashur (160 miles north-northwest of modern Baghdad), the capital (synonymous with the city god and national divinity); Nineveh, lying opposite modern Mosul; and Urbilum, the later Arbela (today Arbīl, some 200 miles north of Baghdad).

In Assyria, inscriptions were composed in Akkadian from the beginning. Under Ur III, Ashur was the provincial capital. Assyria as a whole, however, is not likely to have been a permanently secured part of the empire, since two date formulas of Shulgi and Amar-Su'ena mention the destruction of Urbilum. Ideas of the population of Assyria in the 3rd millennium are necessarily very imprecise. It is not known how long Semitic tribes had been settled there. The inhabitants of southern Mesopotamia called Assyria Shubir in Sumerian and Subartu in Akkadian; these names may point to an older Subarean population that was related to the Hurrians. Gasur, the later Nuzu, belonged to the Akkadian language region about the year 2200 but was lost to the Hurrians in the first quarter of the 2nd millennium. The Assyrian dialect of Akkadian found in the beginning of the 2nd millennium differs strongly from the dialect of Babylonia. These two versions of the Akkadian language continue into the 1st millennium.

Language differences

In contrast to the kings of southern Mesopotamia, the Assyrian rulers styled themselves, not king but partly *iššiakum,* in short, the Akkadian for the Sumerian word *ensi,* partly *rubā'um,* or "great one." Unfortunately, the rulers cannot be synchronized precisely with the kings of southern Mesopotamia before the Assyrian Shamshi-Adad I (*c.* 1813–*c.* 1781 BC). For instance, it has not yet been established just when Ilushuma's excursion into Babylon, recorded in an inscription, actually took place. Ilushuma boasts of having freed of taxes the "Akkadians and their children." While he mentions the cities of Nippur and Ur, the other localities listed were situated in the region east of the Tigris. The event itself may have taken place in the reign of Ishme-Dagan of Isin (*c.* 1953–*c.* 1935 BC), although how far Ilushuma's words correspond to the truth cannot be checked; in the Babylonian texts, at any rate, no reference is made to Assyrian intervention. The whole problem of dating is aggravated by the fact that the Assyrians did not, unlike the Babylonians, use date formulas that often contain interesting historical details; instead, every year was designated by the name of a high official (eponymic dating). The conscious cultivation of an old tradition is mirrored in the fact that two rulers of 19th-century Assyria called themselves Sargon and Naram-Sin, after famous models in the Akkadian dynasty.

Aside from the generally scarce reports on projected construction, there is at present no information about the city of Ashur and its surroundings. There is, however, unexpectedly rewarding source material from the Assyrian trading colonies in Asia Minor. The texts come mainly from Kanesh (near modern Kayseri, in southeast Turkey) and from Hattusas (also southeast Turkey), the later Hittite capital. In the 19th century BC three generations of Assyrian merchants engaged in a lively commodity trade (especially in textiles and metal) between the homeland and Anatolia, also taking part profitably in internal Anatolian trade. Like their contemporaries in southern Mesopotamia, they did business privately and at their own risk, living peacefully and occasionally intermarrying with the "Anatolians." As long as they paid taxes to the local rulers, the Assyrians were given a free hand.

It is obvious that these forays by Assyrian merchants led to some transplanting of Mesopotamian culture into Asia Minor. Thus the Anatolians adopted cuneiform writing and used the Assyrian language as a means of communication. While this influence doubtless already affected the first Hittites arriving in Asia Minor, a direct line from the period of these trading colonies to the Hittite Empire cannot yet be traced.

From *c.* 1813 to *c.* 1781 Assyria was ruled by Shamshi-Adad I, a contemporary of Hammurabi and a personality in no way inferior to him. Shamshi-Adad's father—an Amorite, to judge by the name—had ruled near Mari. The son, not being of Assyrian origin, ascended the throne of Assyria as a foreigner and on a detour, as it were, after having spent some time as an exile in Babylonia. He had his two sons rule as viceroys, in Ekallatum on the Euphrates and in Mari, respectively, until the older of the two, Ishme-Dagan, succeeded his father on the throne. Through the archive of correspondence in the palace at Mari, scholars are particularly well informed about Shamshi-Adad's reign and many aspects of his personality. Shamshi-Adad's state had a common border for some time with the Babylonia of Hammurabi; in the west, it reached to the Mediterranean. Soon after Shamshi-Adad's death, Mari broke away, regaining its independence under an Amorite dynasty that had been living there for generations; in the end, Hammurabi conquered and destroyed Mari. After Ishme-Dagan's death, Assyrian history is lost sight of for over 100 years.

The Old Babylonian Empire. *Political fortunes.* Hammurabi (*c.* 1792–50 BC) is surely the most impressive and by now the most popular figure of the ancient Near East of the first half of the 2nd millennium BC. He owes his posthumous reputation to the great stele into which the "Code of Hammurabi" has been carved and indirectly also to the fact that his dynasty has made the name of Babylon famous for all time. In much the same way in which pre-Sargonic Kish exemplified the non-Sumerian area north of Sumer and Akkad lent its name to a country and a language, Babylon became the symbol of the whole country that the Greeks called Babylonia. This term has already been used anachronistically as a geographical concept in reference to the period before Hammurabi. Originally the city's name was probably Babilla, which was reinterpreted in popular etymology as Bab-ilu (Gate of the God).

The 1st dynasty of Babylon rose from insignificant beginnings. The history of the erstwhile province of Ur is traceable from about 1894 onward, when the Amorite Samu'abum came to power there. What is known of these events fits altogether into the modest proportions of the period when Mesopotamia was a mosaic of small states. Hammurabi played skillfully on the instrument of coalitions and became more powerful than his predecessors had been. But it was only in his 30th year, after his conquest of Larsa, that he gave concrete expression to the idea of ruling all of southern Mesopotamia by "strengthening the fundament of Sumer and Akkad," in the words of that year's dating formula. In the prologue to the Code of Hammurabi the King lists the following cities as belonging to his dominions: Eridu near Ur, Ur, Lagash and Girsu, Zabalam, Larsa, Uruk, Adab, Isin, Nippur, Keshi, Dilbat, Borsippa, Babylon itself, Kish, Malgium, Mashkan-shapir, Kutha, Sippar, Eshnunna in the Diyālā region, Mari, Tuttul on the lower Balīkh (a tributary of the Euphrates), and finally Ashur and Nineveh. This was on a scale reminiscent of Akkad or Ur III. Yet Ashur and Nineveh cannot have formed part

of this empire for long because at the end of Hammurabi's reign mention is made again of wars against Subartu—that is, Assyria.

Decline of Babylonia under Samsuiluna

Under Hammurabi's son Samsuiluna (*c.* 1749–*c.* 1712 BC) the Babylonian Empire greatly shrank in size. Following what had almost become a tradition, the south rose up in revolt; Larsa regained its autonomy for some time, and the walls of Ur, Uruk, and Larsa were levelled to the ground. Eshnunna, which evidently had also seceded, was vanquished in *c.* 1730. Later chronicles mention the existence of a state in the Sealand, with its own dynasty; by "sealand" is understood the marshlands of southern Babylonia. Knowledge of this new dynasty is unfortunately very vague, only one of its kings being documented in contemporary texts. In *c.* 1741 Samsuiluna mentions the Kassites for the first time; in *c.* 1726 he constructed a stronghold, "Fort Samsuiluna," as a bulwark against them on the Diyālā near its confluence with the Tigris.

Like the Gutians before them, the Kassites were at first prevented from entering Babylonia and pushed into the mid-Euphrates region; there, in the kingdom of Khana (centred on Mari and Terqa, both below the junction with the Khābūr), a king appears with the Kassitic name of Kashtiliashu, who ruled toward the end of the Babylonian dynasty. From Khana the Kassites moved south in small groups, probably as harvest workers. After the Hittite invasion under Mursilis I, who is said to have dethroned the last king of Babylon, Samsuditana, in 1595, the Kassites assumed the royal power in Babylonia. So far, the contemporary sources do not mention this epoch, and the question arises as to how the Kassite rulers named in king lists mesh with the end of the 2nd millennium BC.

The Code of Hammurabi

Babylonian law. The Code of Hammurabi is the most frequently cited cuneiform document in specialized literature. Its first scholarly examination in 1902 led to the development of a special branch of comparative jurisprudence, the study of cuneiform law. Following the division made by the first editor, Jean-Vincent Scheil, the Code of Hammurabi contains 280 judgments, or "paragraphs," on civil and criminal law, dealing in the main with cases from everyday life in such a manner that it becomes obvious that the "lawgiver" or compiler had no intention of covering all possible contingencies. In broad outline, the themes treated in the Code of Hammurabi are: libel; corrupt administration of justice; theft, receiving stolen goods, robbery, looting, and burglary; murder, manslaughter, and bodily injury; abduction; judicature of tax lessees; liability for negligent damage to fields and crop damage caused by grazing cattle; illegal felling of palm trees; legal problems of trade enterprises, in particular, the relationship between the merchant and his employee travelling overland, and embezzlement of merchandise; trust monies; the proportion of interest to loan money; the legal position of the female publican; slavery and ransom, slavery for debt, runaway slaves, the sale and vindication of slaves, and disclaimer of slave status; the rent of persons, animals, and ships and their respective tariffs, offenses of hired labourers, and the vicious bull; family law: the price of a bride, dowry, the married woman's property, wife and concubine, and the legal position of the respective issue, divorce, adoption, the wet nurse's contract, and inheritance; and the legal position of certain priestesses.

A similar if much shorter compendium of judgments, probably antedating that of Hammurabi by a generation or two, has been discovered in Eshnunna.

Hammurabi, who called his own work *dīnāt mīšarim,* or "verdicts of the just order," states in the epilogue that it was intended as legal aid for persons in search of advice. It is nevertheless doubtful whether these judgments were meant to have binding force in the sense of modern statutes. The Code of Hammurabi differs in many respects from the Code of Lipit-Ishtar, which was written in Sumerian. Its most striking feature lies in the extraordinary severity of its penalties and in the principle of the lex talionis. The same attitude is reflected in various Old Babylonian treaties in which defaulters are threatened with bodily punishment. It is often said, and perhaps rightly so, that this severity, which so contrasts with Sumerian judicial tradition, can be traced back to the Amorite influence.

There is yet another way in which the Code of Hammurabi has given rise to much discussion. Many of its "paragraphs" vary according to whether the case concerns an *awīlum,* a *muškēnum,* or a *wardum.* A threefold division of the populace had been postulated on the basis of these distinctions. The *wardum* is the least problematical of them all: he is the slave—that is, a person in bondage who could be bought and sold, unless he was able to regain his freedom under certain conditions as a debtor-slave. The *muškēnum* were, under King Hammurabi at least, persons employed by the palace who could be given land in usufruct without receiving it as property. *Awīlum* were the citizens who owned land in their own right and depended neither on the palace nor the temple. As the scholar I.M. Diakonoff points out, the distinction cannot have been very sharply drawn because the classes *awīlum* and *muškēnum* are not mutually exclusive: a man in high palace office could fairly easily purchase land as private property, whereas the free citizen who got into debt as a result of a bad harvest or some other misfortune had one foot in the slave class. Still unanswered is the question as to which segment of the population could be conscripted to do public works, a term that included the levy in case of war.

Economic legislation

Ammisaduqa (*c.* 1646–*c.* 1626 BC) comes a century and a half after Hammurabi. His edict, already referred to, lists, among others, the following social and economic factors: private debts in silver and grain, if arising out of loans, were cancelled; also cancelled were back taxes that certain officials owed the palace and that had to be collected from the people; the "female publican" had to renounce the collection of outstanding debts in beer and barley and was, in turn, excused from paying amounts of silver and barley to the king; taxes on leased property were reduced; debt slaves who had formerly been free (as against slaves made over from debtor to creditor) were ransomed; and high officials were forbidden on pain of death to press those who held property in fee into harvest work by prepayment of wages. The phrase "because the king gave the land a just order" serves as a rationale for many of these instances. In contrast to the "codes," about whose binding force there is much doubt, edicts such as those of Ammisaduqa had legal validity since there are references to the edicts of other kings in numerous legal documents of the Old Babylonian period.

Babylonian literature. The literature and the literary languages of Babylonia during the three centuries following Ur III deserve attention. When commenting on literary and historical texts such as the inscriptions of the kings of Akkad, it was pointed out that these were not originals but copies of Old Babylonian vintage. So far, such copies are the main source for Sumerian literature. But, while the Old Babylonian period created much literature itself (royal hymns of the kings of Isin, Larsa, and Babylon and elegies), it was above all a time of intensive cultivation of traditional literature. The great Sumerian poems, whose origins or first written version, respectively, can now be traced back to about 2600, were copied again and again. After 2000, when Sumerian as a spoken language rapidly receded to isolated regions and eventually disappeared altogether, texts began to be translated, line by line, into Akkadian until there came to be bilingual versions. An important part of this, especially in the instructional program in schools, were the so-called lexicographic texts. Sumerian word lists are almost as old as cuneiform writing itself because they formed the perfect material for those learning to write. In the Old Babylonian period, the individual lexical entries were translated and often annotated with phonetic signs. This led to the creation of "dictionaries," the value of which to the modern philologist cannot be exaggerated. Since Sumerian had to be taught much more than before, regular "grammatical treatises" also came into being: so far as it was possible, in view of the radically different structure of the two languages, Sumerian pronouns, verb forms, and the like were translated into Akkadian, including entire "paradigms" of individual verbs.

In belles-lettres, Sumerian still predominates, although there is no lack of Akkadian masterpieces, including the

oldest Akkadian version of the *Epic of Gilgamesh.* The very high prestige still enjoyed by Sumerian should not be underestimated, and it continued to be used for inscriptions on buildings and the yearly dating formulas. Aside from being the language of practical affairs (letters and contracts), there was a high incidence of Akkadian in soothsaying and divinatory literature. To be sure, the Sumerians also practiced foretelling the future from the examination of animal entrails, but as far as is known they did not write down the results. In Akkadian, on the other hand, there are extensive and "scientifically" arranged compendiums of omens based on the liver (as well as other omens), reflecting the importance that the divination of the future had in religion, in politics, and in all aspects of daily life.

Judging by its continually more refined juridical thought, its ability to master in writing ever more complicated administrative procedures, its advanced knowledge of mathematics, and the fact that it marks the beginning of the study of astronomy, the Old Babylonian period appears to have been a time of exceedingly active intellectual endeavour—despite, if not because of, its lack of political cohesiveness.

The Hurrians. The Hurrians enter the orbit of ancient Near Eastern civilization toward the end of the 3rd millennium BC. They arrived in Mesopotamia from the north or the east, but it is not known how long they had lived in the peripheral regions. There is a brief inscription in Hurrian language from the end of the period of Akkad, while that of King Arishen (or Atalshen) of Urkish and Nawar is written in Akkadian. The language of the Hurrians must have belonged to a widespread group of ancient Near Eastern languages. The relationship between Hurrian and Subarean has already been mentioned, and the language of the Urarteans, who played an important role from the end of the 2nd millennium to the 8th century BC, is likewise closely related to Hurrian. Soviet scholars claim to be able to show that the "Veinakh" group of northeastern Caucasian languages is an offspring of the Hurrian-Urartean group.

It is not known whether the migrations of the Hurrians ever took the form of aggressive invasion: 18th-century-BC texts from Mari speak of battles with the Hurrian tribe of Turukku south of Lake Urmia (some 150 miles from the Caspian Sea's southwest corner); but these were mountain campaigns, not the warding off of an offensive. Proper names in cuneiform texts, increasing in the period of Ur III, constitute the chief criterion for the presence of Hurrians. Nevertheless, there is no clear indication that the Hurrians had already advanced west of the Tigris at that time. An entirely different picture results from the 18th-century palace archives of Mari and from texts originating near the upper Khābūr. Northern Mesopotamia, west of the Tigris, and Syria now appear settled by a population that is mainly Amorite and Hurrian; and the latter have already reached the Mediterranean littoral, as shown by texts from Alalakh on the Orontes. In Mari, literary texts in Hurrian have also been found, indicating that Hurrian had by now become a full-fledged written language as well.

The high point of the Hurrian period was not reached until about the middle of the 2nd millennium. In the 15th century, Alalakh was heavily Hurrianized; and in the empire of Mitanni the Hurrians represented the leading and perhaps the most numerous population group. (D.O.E.)

THE KASSITES, THE MITANNI, AND THE RISE OF ASSYRIA

About 150 years after the death of Hammurabi, his dynasty was destroyed by an invasion of new peoples. Because there are hardly any written records of this era, the time from about 1600 BC to about 1450 BC (in some areas until 1400 BC) is called the dark ages. The remaining Semitic states, such as the state of Ashur, became minor states within the sphere of influence of the new states of the Kassites and the Hurrians-Mitanni. But the languages of the older cultures, the Akkadian and Sumerian, continued or were soon reestablished. The cuneiform script persisted as the only type of writing in the entire area. Cultural continuity was not broken off, either, particularly in Babylonia. A matter of importance was the emergence of new Semitic leading classes from the ranks of the priesthood and the scribes. These gained increasing power.

The people and the language of the Kassites

The Kassites in Babylonia. The Kassites had settled by 1800 BC in what is now western Iran in the region of Hamadan-Kermanshah. The first to feel their forward thrust was Samsuiluna, who had to repel groups of Kassite invaders. Increasing numbers of Kassites gradually reached Babylonia and southern Mesopotamia. There they founded principalities, of which little is known. No inscription or document in the Kassite language has been preserved. Some 300 Kassite words have been found in Babylonian documents. Nor is much known about the social structure of the Kassites or their culture. There seems to have been no hereditary kingdom. Their religion was polytheistic; the names of 30 gods are known.

The beginning of Kassite rule in Babylonia cannot be dated exactly. A king called Agum II ruled over a state that stretched from western Iran to the middle part of the Euphrates Valley; 24 years after the Hittites had carried off the statue of the Babylonian god Marduk, he regained possession of the statue, brought it back, and renewed the cult, making the god Marduk the equal of the Kassitic equivalent god, Shuqamuna. Meanwhile, native princes continued to reign in southern Babylonia. It was Ulamburiash who finally annexed this area around 1450 and began negotiations with Egypt in Syria. Karaindash built a temple with bas-relief tile ornaments in Uruk (Erech) around 1420. A new capital west of Baghdad, Dur Kurigalzu, was named after Kurigalzu I (*c.* 1415–*c.* 1390). His successors Kadashman-Enlil I (*c.* 1390–*c.* 1375) and Burnaburiash II (*c.* 1375–*c.* 1347) were in correspondence with the Egyptian rulers Amenhotep III and Akhenaton (Amenhotep IV). They were interested in trading their lapis lazuli and other items for gold as well as in planning political marriages. Kurigalzu II (*c.* 1345–*c.* 1324) fought against the Assyrians but was defeated by them. His successors sought to ally themselves with the Hittites in order to stop the expansion of the Assyrians. During the reign of Kashtiliash IV (*c.* 1242–*c.* 1235), Babylonia waged war on two fronts at the same time—against Elam and Assyria—ending in the catastrophic invasion and destruction of Babylon by Tukulti-Ninurta I. Not until the time of the kings Adad-shum-nasir (*c.* 1218–*c.* 1189) and Melishipak (*c.* 1188–*c.* 1172) was Babylon again able to experience a period of prosperity and peace. Their successors were again forced to fight, facing the conqueror, King Shutruk-Nahhunte of Elam (*c.* 1185–*c.* 1155). Cruel and fierce, the Elamites finally destroyed the dynasty of the Kassites during these wars (about 1155). Some poetical works lament this catastrophe.

Babylonia under the Kassites

Letters and documents of the time after 1380 show that many things had changed after the Kassites took power. The Kassite upper class, always a small minority, had been largely babylonized. Babylonian names were to be found even among the royalty, and they predominated among the civil servants and the officers. The new feudal character of the social structure showed the influence of the Kassites. Babylonian town life had revived on the basis of commerce and handicrafts. The Kassitic nobility, however, maintained the upper hand in the rural areas, their wealthiest representatives holding very large landed estates. Many of these holdings came from donations of the king to deserving officers and civil servants, considerable privileges being connected with such grants. From the time of Kurigalzu II these were registered on stone tablets or, more frequently, on boundary stones. After 1200 the number of these increased substantially, because the kings needed a steadily growing retinue of loyal followers. The boundary stones had pictures in bas-relief, very often a multitude of religious symbols, and frequently detailed inscriptions giving the borders of the particular estate; sometimes the deserts of the recipient were listed and his privileges recorded; finally, trespassers were threatened with the most terrifying curses. Agriculture and cattle husbandry were the main pursuits on these estates, and horses were raised for the light war chariots of the cavalry. There was an export trade in horses and vehicles, in exchange for raw material.

The decline of the Babylonian culture at the end of the

Old Babylonian period continued for some time under the Kassites. Not until approximately 1420 did the Kassites develop a distinctive style in architecture and sculpture. Kurigalzu I played an important part, especially in Ur, as a patron of the building arts. Poetry and scientific literature developed only gradually after 1400. The existence of earlier work is clear from poetry, philological lists, and collections of omens and signs that were in existence by the 14th century or before and have been discovered in the Hittite capital of Hattusas, in the Syrian capital of Ugarit, and even as far away as Palestine. Somewhat later, new writings appear—medical diagnoses and recipes, more Sumero–Akkadian word lists, and collections of astrological and other omens and signs with their respective interpretations. Most of these works are known today only from copies of more recent date. The most important of them all is the Babylonian epic of the creation of the world, *Enuma elish.* Composed by an unknown poet, in all probability in the 14th century, it tells the story of the god Marduk. He began as the god of Babylon and was elevated to be king over all other gods after having successfully accomplished the destruction of the powers of chaos. For almost 1,000 years this epic was recited during the new year's festival in the spring as part of the Marduk cult in Babylon. The literature of this time contains hardly any Kassitic words. Many scholars believe that the essential groundwork for the development of the subsequent Babylonian culture was laid during the later epoch of the Kassite era.

Kingdom of the Hurrians and the Mitanni. The weakening of the Semitic states in Mesopotamia after 1600 enabled the Hurrians to penetrate deeper into this region, where they founded numerous small states in the eastern parts of Asia Minor, Mesopotamia, and Syria. The Hurrians came from northwestern Iran, but their early history is almost unknown. After 1500, isolated dynasties appeared with Indo-Aryan names, but the significance of this is disputed. The fact that Old Indian technical terms appear in later records about horse breeding and that the names of Indian gods are used in some compacts of state led many scholars to assume that numerous groups of Aryans, closely related to the Indians, pushed into Asia Minor from the northeast. They were also credited with the introduction of the light war chariot with spoke wheels. This conclusion, however, is by no means established fact. So far it has not been possible to appraise the quantity and the political and cultural influence of the Aryans in Asia Minor relative to that of the Hurrians.

Empire of the Mitanni

Shortly after 1500 the kingdom of the Mitanni arose near the sources of the Khābūr River in Mesopotamia. Since no record or inscription of their kings has been unearthed, there is hardly anything known about the development and history of the Mitanni before King Tushratta. The Mitanni were known to the Egyptians under the name of Naharina, and Thutmose III fought frequently against them after 1468 BC. By 1440 the domain of the Mitanni king Saustatar (Saushatar) stretched from the Mediterranean all the way to the Zagros Mountains, in western Iran, including Alalakh, in northern Syria, as well as Nuzu and Arrapkha. The northern boundary dividing the Mitanni from the Hittites and the Hurrians was in all probability not fixed even under his successors Artatama and Shuttarna II, who married their daughters to the pharaohs Thutmose IV (1425–1417) and Amenhotep III (1417–1379). Tushratta (*c.* 1390–*c.* 1340), the son of Shuttarna, was able to maintain the kingdom he had inherited from his predecessors for many years. In his sometimes very long letters to Amenhotep III and Akhenaton he writes about commerce, his desire for gold, and marriage. Weakened by internal strife, the Mitanni kingdom eventually became a pawn between the rising kingdoms of the Hittites and the Assyrians.

The kingdom of the Mitanni was a feudal state led by a warrior nobility of Aryan or Hurrian origin. Frequently horses were bred on their large landed estates. Documents and contract agreements in Syria often mention a chariot-warrior caste that also constituted the social upper class in the cities. The aristocratic families usually received their landed property as an inalienable fief. Consequently, no documents on the selling of landed property are to be found in the great archives of Akkadian documents and letters discovered in Nuzu, near Kirkūk. The prohibition against selling landed property was often dodged, however, with a stratagem: the previous owner "adopted" a willing buyer against an appropriate sum of money. The wealthy lord Tehiptilla was "adopted" almost 200 times, acquiring tremendous holdings of landed property in this way without interference by the governmental authorities. He had gained his wealth through trade and commerce and in agriculture. For a long time, a prince was in charge of the royal governmental administration in the district capital. Society was highly structured in classes, ranks, and professions. The judiciary, patterned after the Babylonian model, was well organized; the documents place heavy emphasis on correct procedure.

Temple monuments of modest dimensions have been unearthed; in all probability, specific local traditions played a deciding factor in their design. The dead were probably buried outside the settlement. There is evidence that members of the royal family were cremated. Small artifacts, particularly seals, show a peculiar continuation of Babylonian and Assyrian traditions in their preference for the naturalistic representation of figures. There were painted ceramics with finely drawn decorations (white on a dark background). The strong position of the royal house was evident in the large palaces, existing even in district capitals. The palaces were decorated with frescoes. Because only a few settlements of the Mitanni have been unearthed in Mesopotamia, knowledge of the Mitanni arts and culture is as yet insufficient.

Assyria under Mitanni rule

The rise of Assyria. Very little can be said about northern Assyria during the 2nd millennium BC. Information on the old capital, Ashur, located in the south of the country, is somewhat more plentiful. The old lists of kings suggest that the same dynasty ruled continuously over Ashur from about 1670 on. All the names of the kings are given. But little else is known about Ashur before 1420. Almost all the princes had Akkadian names, and it can be assumed that their sphere of influence was rather small. Although Assyria belonged to the kingdom of the Mitanni for a long time, it seems that Ashur retained a certain autonomy. Located close to the boundary with Babylonia, it played that empire off against the Mitanni whenever possible. Puzur-Ashur III concluded a border treaty with Babylonia in 1490 as did Ashur-belnisheshu in 1415. Ashur-nadin-akhe II (*c.* 1402–*c.* 1393) was even able to obtain support from Egypt, which sent him a consignment of gold.

Ashur-uballit I (*c.* 1365–*c.* 1330) was at first subject to King Tushratta of the Mitanni. After 1350, however, he attacked Tushratta, presumably together with Suppiluliumas of the Hittites. Taking away from the Mitanni parts of northeastern Mesopotamia, Ashur-uballit I now called himself "Great King" and socialized with the King of Egypt on equal terms—arousing the indignation of the King of Babylonia. Ashur-uballit was the first to name Assyria the Land of Ashur, because the old name, Subartu, was often used in a derogatory sense in Babylonia. He ordered his short inscriptions to be partly written in the Babylonian language rather than the Assyrian, since this was considered refined. Marrying his daughter to a Babylonian, he intervened there energetically when Kassite nobles murdered his grandchild. Future generations came to consider him rightfully as the real founder of the Assyrian Empire. His son Enlil-nirari (*c.* 1328–*c.* 1320) also fought against Babylonia. Arik-den-ili (*c.* 1319–*c.* 1308) turned westward where he encountered Semitic tribes of the so-called Akhlamu-group.

Still greater successes were achieved by Adad-nirari I (*c.* 1307–*c.* 1275). Defeating the Kassite king Nazimaruttash, he forced him to retreat. After that he defeated the kings of the Mitanni, at first Shattuara I, then Wasashatta. This enabled him for a time to incorporate all of Mesopotamia into his empire as a province, although in later struggles he lost large parts to the Hittites. In the east, he was satisfied with the defense of his lands against the mountain people.

Adad-nirari's inscriptions were more elaborate than those

of his predecessors and were written in the Babylonian language. In them he declares that he feels called to these wars by the gods, a statement that was to be repeated by other kings after him. Assuming the old title of great king, he called himself "King of All." He enlarged the temple and the palace in Ashur and also developed the fortifications there, particularly at the banks of the Tigris River. He worked on large building projects in the provinces.

His son Shalmaneser I (*c.* 1274–*c.* 1245) attacked Uruatru (later called Urartu) in southern Armenia, which had allegedly broken away. But Shattuara II of Hanigalbat put him into a difficult situation, cutting his forces off from their water supplies. With courage born of despair, the Assyrians fought themselves free. They then set about reducing what was left of the Mitanni kingdom into an Assyrian province. The king claimed to have blinded 14,400 enemies in one eye—psychological warfare of a similar kind was used more and more as time went by. The Hittites tried in vain to save Hanigalbat. Together with the Babylonians they fought a commercial war against Ashur for many years. Like his father, Shalmaneser was a great builder. At the juncture of the Tigris and the Great Zab rivers he founded a strategically situated second capital, Kalakh (biblical Calah; modern Nimrūd).

His son was Tukulti-ninurta (*c.* 1244–*c.* 1208), the Ninos of Greek legends. Gifted but extravagant, he made his nation a great power. He carried off thousands of Hittites from eastern Anatolia. He fought particularly hard against Babylonia, deporting Kashtiliash IV to Assyria. When the Babylonians rebelled again, he plundered the temples in Babylon, an act regarded as a sacrilege, even in Assyria. The relationship between the king and his capital deteriorated steadily. For this reason the king began to build a new city, Kar-Tukulti-Ninurta, on the other side of the Tigris River. Ultimately, even his sons rebelled against him and laid siege to him in his city; in the end he was murdered. His victorious wars against Babylonia were glorified in an epic poem, but his empire broke up soon after his death. Assyrian power declined for a time, while that of Babylonia rose.

Assyria had suffered under the oppression of both the Hurrians and the Mitannis. Its struggle for liberation and the bitter wars that followed had much to do with its development into a military power. In his capital of Ashur, the king depended on the citizen class and the priesthood, as well as on the landed nobility that furnished him with the war-chariot troops.

Organization of the Assyrian state

Documents and letters show the important role agriculture played in the development of the state. Assyria was less dependent on artificial irrigation than was Babylonia. The breeding of horses was carried on intensively; remnants of elaborate directions for their training are extant. Trade and commerce were also of notable significance: metals were imported from Asia Minor or Armenia, tin from northwestern Iran, and lumber from the west. The opening up of new trade routes was oftentimes a cause and the purpose of war.

Assyrian architecture, derived from a combination of Mitannian and Babylonian influences, developed early quite an individual style. The palaces often had colourful wall decorations. The art of seal cutting, taken largely from the Mitanni, continued creatively on its own. The schools for scribes, where all the civil servants were trained, taught both the Babylonian and the Assyrian dialects of the Akkadian tongue. Babylonian works of literature were assimilated into Assyrian, often reworked into a partially different form. The Hurrian tradition remained strong in the military and political sphere while at the same time influencing the language.

ASSYRIA AND BABYLONIA AT THE END OF THE 2ND MILLENNIUM

Babylonia under the 2nd dynasty of Isin (c. 1156–c. 1025). In a series of heavy wars about which not much is known, Marduk-kabit-akheshu (*c.* 1156–*c.* 1139) established what came to be known as the 2nd dynasty of Isin. His successors were often forced to continue the fighting. The most famous king of the dynasty was Nebuchadrezzar I (*c.* 1124–*c.* 1103). He fought mainly against Elam, which had conquered most of Babylonia. His first attack miscarried because of an epidemic among his troops, but in a later campaign he conquered Susa, the capital of Elam, and returned the previously stolen statue of the god Marduk to its proper place. Soon thereafter the king of Elam was assassinated, and his kingdom fell apart into small states. This enabled Nebuchadrezzar to turn west, using the later years of peace to complete extensive building projects. After him, his son became king, succeeded by his brother Marduk-nadin-ahhe (*c.* 1098–*c.* 1081). At first successful in his wars against Assyria, he later experienced heavy defeat. A famine of catastrophic proportions triggered an attack from Aramaean tribes, the ultimate blow. His successors made peace with Assyria, but the country suffered more and more from repeated attacks by Aramaeans and other Semitic nomads. Even though some of the kings still had grand titles, they were unable to stem the progressive disintegration of their empire. There followed the era known as the 2nd dynasty of the Sealand (*c.* 1024–*c.* 1004), which included three usurpers. The first of these had the Kassitic name of Simbar-shikhu (*c.* 1024–*c.* 1007).

Toward the end of its reign, the dynasty of the Kassites became completely babylonized. The changeover to the dynasty of Isin, actually a succession of kings from different families, brought no essential transformation of the social structure. The feudal order remained. New landed estates came into existence in many places through grants to deserving officers; many boundary stones (*kudurru*) have been found that describe them. The cities of Babylonia retained much of their former autonomy. The border provinces, however, were administered by royally appointed governors with civil and military functions.

This was a period of creativity in the literary arts. The successes of Nebuchadrezzar I provided new material. One heroic epic, modelled upon older epics, celebrates his deeds, but unfortunately little of this is extant. Other material came from ancient myths. The best known poet is Sinleqeunnini (*c.* 1100–?) of Uruk (Erech), the poet of the later version of the *Epic of Gilgamesh* known as the Twelve-Tablet Poem. This version is distinguished by its greater stress on the human qualities of Gilgamesh and Enkidu, making it one of the great works of world literature. Another poet working at about the same time was the author of verses meditating on the workings of divine justice, a matter that had acquired vital importance in the contemporary religion of Babylon. This poem elaborates on the great suffering of a high official and his subsequent salvation by the god Marduk. Other religious poems have been preserved, but their time of origin is not certain. The later Babylonians with good reason regarded the time of Nebuchadrezzar I as one of the great eras of their history.

Assyria between 1200 and 1000. After a period of decline following Tukulti-Ninurta I, Assyria was consolidated and stabilized under Ashur-dan I (*c.* 1179–*c.* 1134) and Ashur-resh-ishi I (*c.* 1133–*c.* 1116). Several times forced to fight against Babylonia, the latter was even able to defend himself against an attack by Nebuchadrezzar I. According to the inscriptions, most of his building efforts were in Nineveh, rather than in the old capital of Ashur.

The reign of Tiglath-pileser I

His son Tiglath-pileser I (Tukulti-apal-Esharra, *c.* 1115–*c.* 1077) raised the power of Assyria to new heights. First he turned against a large army of the Mushki that had entered into southern Armenia from Asia Minor, defeating them decisively. After this he forced the small Hurrian states of southern Armenia to pay him tribute. Trained in mountain warfare themselves and helped by capable pioneers, the Assyrians were now able to advance far into the distant valleys of the interior. Their main enemies were the Aramaeans, the Semitic camel Bedouins whose many small states often combined against the Assyrians. Tiglath-pileser I also went to Syria and even reached the Mediterranean, where he took a sea voyage. After 1100 these campaigns led to conflicts with Babylonia. Tiglath-pileser conquered northern Babylonia and plundered Babylon, without decisively defeating Marduk-nadin-ahhe. In his own country the King paid particular attention to agriculture and fruit growing, improved the

administrational system, and developed more thorough methods of training scribes.

Three of his sons reigned after Tiglath-pileser, including Ashur-bel-kala (*c.* 1074–*c.* 1057). Like his father, he fought in southern Armenia and with Babylonia as his ally, against the Aramaeans. Disintegration of the empire could not be arrested, however. The grandson of Tiglath-pileser, Ashurnasirpal I (*c.* 1050–*c.* 1032), was sickly and unable to do more than defend Assyria proper against his enemies. Fragments of three of his prayers to the goddess Ishtar are preserved; among them is a penitential prayer in which he wonders about the cause of so much adversity. Referring to his many good deeds but admitting his guilt at the same time, he asks for forgiveness and health. According to the king, part of his guilt lay in neglecting to teach his subjects the fear of god in sufficient measure. After him, little is known for 100 years.

State, law, and justice under Tiglath-pileser

State and society during the time of Tiglath-pileser were not essentially different from those of the 13th century. Collections of laws, drafts, and edicts of the court exist that go back as far as the 14th century BC. Presumably, most of these remained in effect. One tablet defining the marriage laws shows that the social position of women in Assyria was lower than in Babylonia or Israel or among the Hittites. A man was allowed to send away his wife at his own pleasure with or without divorce money. In case of adultery, he was permitted to kill or maim her. Outside of her house the woman was forced to observe many restrictions, such as the wearing of a veil. It is not clear whether these regulations carried the weight of law, but they seem to have represented a reaction against practices that were more favourable to women. Two somewhat older marriage contracts, for example, grant equal rights to both partners, even in divorce. The women of the king's harem were subject to severe punishment, including beating, maiming, and death, along with those who guarded and looked after them. The penal laws of the time were generally more severe in Assyria than in other countries of the East. The death penalty was not uncommon. In less serious cases the penalty was forced labour after flogging. In certain cases there was trial by ordeal. One tablet treats the subject of landed property rights. Offenses against the established boundary lines called for extremely severe punishment. A creditor was allowed to force his debtor to work for him, but he could not sell him.

The greater part of Assyrian literature was either taken over from Babylonia or written by the Assyrians in the Babylonian dialect, who modelled their works upon Babylonian originals. The Assyrian dialect was used in legal documents, court and temple rituals, and collections of recipes—as, for example, in directions for manufacturing perfumes. A new art form was the picture tale: a continuing series of pictures carved on square steles of stone. The pictures, showing war or hunting scenes, begin at the top of the stele and run down around it, with inscriptions under the pictures explaining them. These and the finely cut seals show that the fine arts of Assyria were beginning to surpass those of Babylonia. Architecture and other forms of the monumental arts also began a further development, such as the double temple with its two towers (ziggurat). Colourful, enamelled tiles were used to decorate the façades.

ASSYRIA AND BABYLONIA C. 1000–C. 750 BC

Assyria and Babylonia until Ashurnasirpal II. The most important factor in the history of Mesopotamia in the 10th century was the continuing threat from the Aramaean seminomads. Again and again, the kings of both Babylonia and Assyria were forced to repel their invasions. Even though the Aramaeans were not able to gain a foothold in the main cities, there are evidences of them in many rural areas. Ashur-dan II (934–912) succeeded in suppressing the Aramaeans and the mountain people, in this way stabilizing the Assyrian boundaries. He was the first to use the Assyrian dialect in his written records.

The reign of Adad-nirari II

Adad-nirari II (*c.* 911–891) left very detailed accounts of his wars and his efforts to improve agriculture. He led six campaigns against Aramaean intruders from northern Arabia. In two campaigns against Babylonia he forced Shamash-mudammiq (about 930–904) to surrender extensive territories. Shamash-mudammiq was murdered, and a treaty with his successor, Nabu-shum-ukin (904–888) secured peace for many years. Tukulti-Ninurta II (890–884), the son of Adad-nirari II, reigned only for a short time. He preferred Nineveh to Ashur. His wars led him to southern Armenia. He was portrayed on steles in blue-yellow enamel in late Hittite style, showing him under the winged sun—a theme adopted from Egyptian art. His son Ashurnasirpal II (883–859) continued the policy of conquest and expansion. He left a very detailed account of his campaigns, which were impressive in their cruelty. Defeated enemies were impaled, flayed, or beheaded in great numbers. But mass deportations were found to serve the interests of the growing empire better than terror. Through the systematic exchange of native populations, conquered regions were denationalized. The result was a servile, submissive, mixed population in which the Aramaean element became the preponderant majority. This provided the labour force for the various public works in the metropolitan centres of the Assyrian Empire. Ashurnasirpal II rebuilt Kalakh, founded by Shalmaneser I, and made it his capital. Ashur remained the centre of the worship of the god Ashur—in whose name all the wars of conquest were fought. A third capital was Nineveh.

The reign of Ashurnasirpal II

Ashurnasirpal II was the first to use cavalry units to any large extent in addition to infantry and war-chariot troops. He was also the first to employ heavy, mobile battering rams and wall breakers in his sieges. Following after the conquering troops came officials from all branches of the civil service, because the king wanted to lose no time in incorporating the new lands into his empire. The supremacy of Assyria over its neighbouring states owed much to the proficiency of the government service under the leadership of the minister Gabbilani'-eresh. The campaigns of Ashurnasirpal II lead him mainly to southern Armenia and Mesopotamia. After a series of heavy wars, he incorporated Mesopotamia as far as the Euphrates River. A campaign to Syria encountered little resistance. There was no great war against Babylonia. Ashurnasirpal, like other Assyrian kings, may have been moved by religion not to destroy Babylonia, which had almost the same gods as Assyria. Both empires must have profited from mutual trade and cultural exchange. The Babylonians, under the energetic Nabu-apal-iddin (*c.* 887–855) attacked the Aramaeans in southern Mesopotamia and occupied the valley of the Euphrates River to about the mouth of the Khābūr River.

Ashurnasirpal, so brutal in his wars, was able to inspire architects, structural engineers, and artists and sculptors to heights never before achieved. He built and enlarged temples and palaces in several cities. His most impressive monument was his own palace in Kalakh, covering a space of 25,000 square metres (269,000 square feet). Hundreds of large limestone slabs were used in murals in the state rooms and living quarters. Most of the scenes were done in relief, but painted murals have also been found. Most of them depict mythological themes and symbolic fertility rites, with the king participating. Brutal war pictures were aimed to discourage enemies. The chief god of Kalakh was Ninurta, god of war and the hunt. The tower of the temple dedicated to Ninurta served also as an astronomical observatory. Kalakh soon became the cultural centre of the empire. Ashurnasirpal claimed to have entertained 69,574 guests at the opening ceremonies of his palace.

Shalmaneser III and Shamshi-Adad V of Assyria. The son and successor of Ashurnasirpal was Shalmaneser III (858–824). His father's equal in both brutality and energy, he was less realistic in his undertakings. His inscriptions, in a peculiar blend of Assyrian and Babylonian, record his considerable achievements but are not always able to conceal his failures. His campaigns were directed mostly against Syria. While he was able to conquer northern Syria and make it a province, in the south he could only weaken the strong state of Damascus and was unable, even after several wars, to eliminate it. In 841 he laid unsuccessful siege to Damascus. Also in 841 King Jehu of Israel was

Conquest of northern Syria

forced to pay tribute. In his invasion of Cilicia, Shalmaneser had only partial success. The same was true of the kingdom of Urartu in Armenia, from which, however, the troops returned with immense quantities of lumber and building stones. The king and, in later years, the general Daian-Ashur, went several times to western Iran, where they found such states as Mannai in northwestern Iran and, further away in the southeast, the Persians. They also encountered the Medes during these wars. Horse tribute was collected.

In Babylonia, Marduk-zakir-shumi I ascended the throne around the year 855. His brother Marduk-bel-usate rebelled against him, and in 851 the king was forced to ask Shalmaneser for help. Shalmaneser was only too happy to oblige; when the usurper had been finally eliminated (850), Shalmaneser went to southern Babylonia, which at that time was almost completely dominated by Aramaeans. There he encountered, among others, the Chaldeans, mentioned for the first time in 878 BC, who were to play a leading role in the history of later times; Shalmaneser made them tributaries.

During his long reign he built temples, palaces, and fortifications in Assyria as well as in the other capitals of his provinces. His artists created many statues and steles. Among the best known is the Black Obelisk, which includes a picture of Jehu of Israel paying tribute. The bronze doors of Imgur-Enlil in Assyria portray the course of his campaigns and other undertakings in rows of pictures, often very lifelike. Hundreds of delicately carved ivories were carried away from Phoenicia, and many of the artists along with them; these later made Kalakh a centre for the art of ivory sculpture.

In the last four years of the reign of Shalmaneser, the crown prince Ashur-danin-apal led a rebellion. The old King appointed his younger son Shamshi-Adad as the new crown prince. Forced to flee to Babylonia, Shamshi-Adad V (823–811) finally managed to regain the kingship with the help of Marduk-zākir-shumi I under humiliating conditions. As king he campaigned with varying success in southern Armenia and Azerbaijan, later turning against Babylonia. He won several battles against the Babylonian kings Marduk-balassu-iqbi and Baba-ahaiddina (about 818–12) and pushed through to Chaldea. Babylonia remained independent, however.

Adad-nirari III and his successors. Shamshi-Adad V died while Adad-nirari III (810–783) was still a minor. His Babylonian mother, Sammu-ramat, took over the regency, governing with great energy until 806. The Greeks, who called her Semiramis, credited her with legendary accomplishments, but historically little is known about her. Adad-nirari later led several campaigns against the Medes and also against Syria and Palestine. In 804 he reached Gaza, but Damascus proved invincible. He also fought in Babylonia, helping to restore order in the north.

Shalmaneser IV (*c.* 783–773) fought against Urartu, then at the height of its power under King Argishti (about 780–755). He successfully defended eastern Mesopotamia against attacks from Armenia. On the other hand, he lost most of Syria after a campaign against Damascus in 773. The reign of Ashur-dan III (772–755) was shadowed by rebellions and by epidemics of plague. Of Ashur-hirari V (754–746) little is known.

In Assyria, the feudal structure of society remained largely unchanged. Many of the conquered lands were combined to form large provinces. The governors of these provinces sometimes acquired considerable independence, particularly under the weaker monarchs after Adad-nirari III. Some of them even composed their own inscriptions. The influx of displaced peoples into the cities of Assyria created large metropolitan centres. The spoils of war, together with an expanding trade, favoured the development of a well-to-do commercial class. The dense population of the cities gave rise to social tensions that only the strong kings were able to contain. A number of the former capitals of the conquered lands remained important as capitals of provinces. There was much new building. A standing occupational force was needed in the provinces, and these troops grew steadily in proportion to the total military forces. The civil service also expanded, the largest administrative body being the royal court, with thousands of functionaries and craftsmen in the several residential cities.

The cultural decline around the year 1000 was overcome during the reigns of Ashurnasirpal II and Shalmaneser III. The arts in particular experienced a tremendous resurgence. Literary works continued to be written in Assyrian and were seldom of great importance. The literature that had been taken over from Babylonia was further developed with new writings, although one can rarely distinguish between works written in Assyria and works written in Babylonia. In religion, the official cults of Ashur and Ninurta continued, while the religion of the common people went its separate way.

In Babylonia, not much was left of the feudal structure; the large landed estates almost everywhere fell prey to the inroads of the Aramaeans, who were at first half nomadic. The leaders of their tribes and clans slowly replaced the former landlords. Agriculture on a large scale was no longer possible except on the outskirts of metropolitan areas. The predominance of the Babylonian schools for scribes may have prevented the emergence of an Aramaean literature. In any case, the Aramaeans seem to have been absorbed into the Babylonian culture. The religious cults in the cities remained essentially the same. The Babylonian Empire was slowly reduced to poverty, except perhaps in some of the cities. In 764, after an epidemic, the myth of pest god Erra was written by the poet Kabt-ilani-Marduk.

THE NEO-ASSYRIAN EMPIRE (746–609)

Tiglath-pileser III and Shalmaneser V. The decline of Assyrian power after 780 was notable; Syria and considerable lands in the north were lost. A military coup deposed King Ashur-nirari V and raised a general to the throne. Under the name of Tiglath-pileser III (746–727) he brought the empire to its greatest expanse. He reduced the size of the provinces in order to break the independence of the governors. He also invalidated the tax privileges of cities such as Ashur and Haran, in order to distribute the tax load more evenly over the entire realm. Military equipment was improved substantially. In 746 he went to Babylonia to aid Nabu-nasir (747–734) in his fight against Aramaean tribes. Tiglath-pileser beat the Aramaeans and then made visits to the large cities of Babylonia. Here he tried to get the support of the priesthood by patronizing their building projects. Babylonia retained its independence.

Assyria at its peak

His next undertaking was to throw back Urartu. His campaigns in Azerbaijan were designed to drive a wedge between Urartu and the Medes. In 743 he went to Syria, defeating there an army of Urartu. The Syrian city of Arpad, which had formed an alliance with Urartu, did not surrender so easily. It took Tiglath-pileser three years of siege to conquer Arpad, whereupon he massacred the inhabitants and destroyed the city. In 738 a new coalition formed against Assyria under the leadership of Samal of northern Syria. It was defeated, and all the princes from Damascus to eastern Anatolia were forced to pay tribute. Another campaign in 735, this time directed against Urartu itself, was only partly successful. In 734 Tiglath-pileser invaded Palestine, going as far as the Egyptian border. Damascus and Israel tried to organize resistance against him, seeking to bring Judah into their alliance. But Ahaz of Judah asked Tiglath-pileser for help. In 733 Tiglath-pileser devastated Israel and forced it to surrender large territories. In 732 he advanced upon Damascus, first devastating the gardens outside the city and then conquering the capital and killing the king, whom he replaced with a governor. The queen of southern Arabia, Samsia, was now obliged to pay tribute, being permitted in return to use the harbour of the city of Gaza, which was in Assyrian hands.

The death of King Nabonassar of Babylonia caused a chaotic situation to develop there, and the Aramaean Ukin-zer crowned himself king. In 731 Tiglath-pileser fought and beat him and his allies, but he did not capture Ukin-zer until 729. This time he did not appoint a new king for Babylonia but assumed the crown himself under the name Pulu (Pul in the Old Testament). In his

old age he abstained from further campaigning, devoting himself to the improvement of his capital Kalakh. He rebuilt the palace of Shalmaneser III, filled it with treasures from his wars, and decorated the walls with bas-reliefs. The latter were almost all of warlike character, as if designed to intimidate the onlooker with their presentation of gruesome executions. These pictorial narratives on slabs have also been found in several provincial capitals of Assyria.

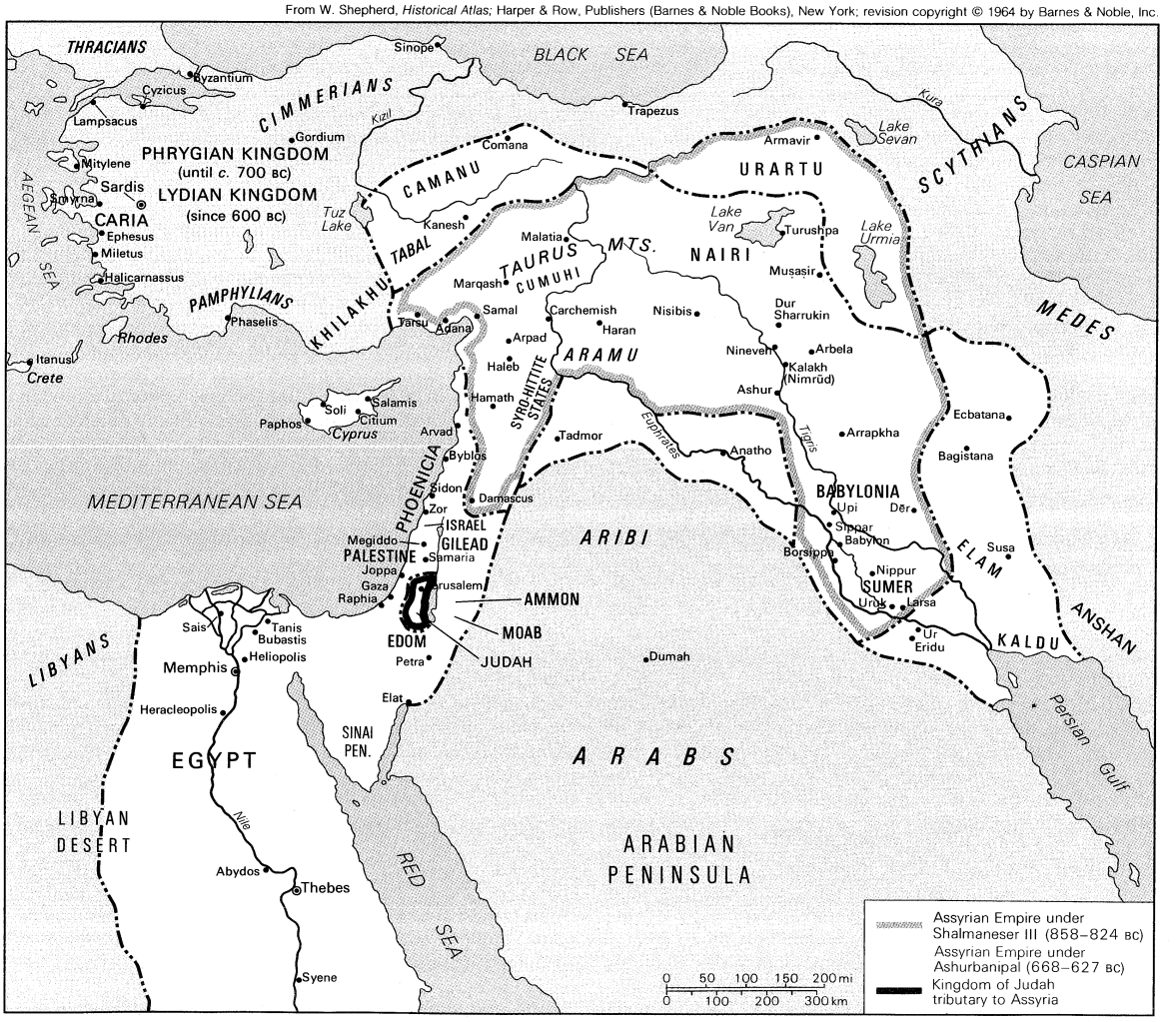
Tiglath-pileser was succeeded by his son Shalmaneser V (726–722), who continued the policy of his father. As king of Babylonia he called himself Ululai. Almost nothing is known about his enterprises, since his successor destroyed all his inscriptions. The Old Testament relates that he marched against Hoshea of Israel in 724 after Hoshea had rebelled. He was probably assassinated during a three-year siege of Samaria. His successor maintained that the god Ashur had withdrawn his support of Shalmaneser V for acts of disrespect.

Sargon II (721–705) and Marduk-apal-iddina of Babylonia. It was probably a younger brother of Shalmaneser who ascended the throne of Assyria. Assuming the old name of Sharru-kin (Sargon in the Bible), meaning Legitimate King, he assured himself of the support of the priesthood and the merchant class by restoring privileges they had lost, particularly the tax exemptions of the great temples. The change of sovereign in Assyria triggered another crisis in Babylonia. An Aramaean prince from the south, Marduk-apal-iddina II (the biblical Merodach-Baladan), seized power in Babylon in 721 and was able to retain it until 710 with the help of Humbanigash I of Elam. A first attempt by Sargon to recover Babylonia miscarried when Elam defeated him in 721. During the same year the three-year siege of Samaria was brought to a close. The Samarian upper class was deported, and Israel became an Assyrian province. Samaria was repopulated with Syrians and Babylonians. Judah remained independent by paying tribute. In 720 Sargon squelched a rebellion in Syria that had been supported by Egypt. Then he defeated both Hanunu of Gaza and an Egyptian army near the Egyptian border. In 717 and 716 he campaigned in northern Syria, making the hitherto independent state of Carchemish one of his provinces. He also went to Cilicia in an effort to prevent further encroachments of the Phrygians under King Midas.

Sargon II's military campaigns

In order to protect his ally, the state of Mannai, in Azerbaijan, Sargon embarked on a campaign in Iran in 719 and incorporated parts of Media as provinces of his empire. But in 716 another war became necessary. At the same time, he was busy preparing a major attack against Urartu. Under the leadership of the crown prince Sennacherib, armies of agents infiltrated Urartu, which was also threatened from the north by the Cimmerians. Many of their messages and reports have been preserved. The longest inscription ever composed by the Assyrians about a year's enterprise (430 very long lines) is dedicated to this Urartu campaign of 714. Phrased in the style of a first report to the god Ashur, it is interlaced with stirring descriptions of natural scenery. The strong points of Urartu must have been well fortified. Sargon tried to avoid them by going through the province of Mannai and attacking the Median principalities on the eastern side of Lake Urmia. In the meantime, hoping to surprise the Assyrian troops, Rusas of Urartu had closed the narrow pass lying between Lake Urmia and Sahand Mount. Sargon, anticipating this, led a small band of cavalry in a surprise charge that developed into a great victory for the Assyrians. Rusas fled and died. The Assyrians pushed forward, destroying all the

From W. Shepherd, *Historical Atlas*; Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.



The Assyrian Empire, 858–627 BC.

cities, fortifications, and even irrigation works of Urartu. They did not conquer Mannai itself but took possession of the mountain city of Muṣaṣir. The spoils were immense. The following years saw only small campaigns in Media, eastern Asia Minor, and against Ashdod, in Palestine. King Midas of Phrygia and some cities on Cyprus were quite ready to pay tribute.

Sargon was now free to settle accounts with Marduk-apal-iddina of Babylonia. Abandoned by his ally Shutruk-Nahhunte II of Elam, Marduk-apal-iddina found it best to flee: first to his native land on the Persian Gulf and later to Elam. Because the Aramaean prince had made himself very unpopular with his subjects, Sargon was hailed as the liberator of Babylonia. He complied with the wishes of the priesthood while at the same time putting down the Aramaean nobility. The modest title of governor of Babylonia was enough for him.

The founding of Dur Sharrukin

At first Sargon resided in Kalakh, but he then decided to found an entirely new capital north of Nineveh. He called the city Dur Sharrukin—Sargonsburg (now Khorsabad). He erected his palace on a high terrace in the northeastern part of the city. The temples to the main gods, smaller in size, were built within the palatial rectangle, which was surrounded by a special wall. This arrangement enabled Sargon to supervise the priests better than had been possible in the old, large temple complexes. One consequence of this design was that the figure of the king pushed the gods somewhat into the background, thereby gaining in importance. Wanting his palace to match the vastness of his empire, the King planned it in monumental dimensions. Stone reliefs of two winged bulls with human heads flanked the entrance—much larger than anything comparable built before. The walls were decorated with long rows of bas-reliefs showing scenes of war and festive processions. A comparison with a well-executed stele of the Babylonian king Marduk-apal-iddina shows that the fine arts of Assyria had far surpassed those of Babylonia. Sargon never completed his capital, though from 713 to 705 BC tens of thousands of labourers and hundreds of artisans worked upon the great city. Yet, with the exception of some magnificent buildings for public officials, only a few durable edifices were completed in the residential section. In 705, in a campaign in northwestern Iran, the King was ambushed and killed. His corpse remained unburied, to be devoured by birds of prey. Sargon's son Sennacherib, who had quarrelled with his father, was inclined to believe with the priests that the death of the King was a punishment from the gods.

Sennacherib. Sennacherib (Assyrian Sin-akhkhe-eriba; 704–681) was well prepared for his position as sovereign. With him Assyria acquired an exceptionally clever and gifted, though often extravagant, ruler. His father, interestingly enough, is not mentioned in any of his many inscriptions. He left the new city of Dur Sharrukin at once and resided in Ashur for a few years, until in 701 he made Nineveh his capital.

Sennacherib had considerable difficulty with Babylonia. In 703 Marduk-apal-iddina again crowned himself king with the aid of Elam, proceeding at once to ally himself with other enemies of Assyria. After nine months he was forced to withdraw when Sennacherib defeated a coalition army consisting of Babylonians, Aramaeans, and Elamites. The new puppet king of Babylonia was Bel-ibni (702–700), who had been raised in Assyria.

The siege of Jerusalem

In 702 Sennacherib undertook a predatory raid into western Iran, then moved against Syria and Palestine in 701 to smash a coalition of Phoenicians, Palestinians, and Egyptians. Sidon immediately expelled its ruler, who was hostile to Assyria. The other allies surrendered or were beaten. One Egyptian army was defeated at Eltekeh. Only Hezekiah of Judah did not surrender, fearing grave judgment and retribution. Failing in their efforts to incite the people of Jerusalem against their king, the Assyrians laid siege to the city. The Old Testament relates these events in great detail. The most likely reconstruction of events is that, when an epidemic broke out among the Assyrian troops during the siege of Jerusalem, Sennacherib found it advisable to return to Assyria. In the course of time several generations of storytellers in Judah swelled the number of Assyrian dead to 185,000. Hezekiah nevertheless paid tribute to Nineveh.

Bel-ibni of Babylonia seceded from the union with Assyria in 700. Sennacherib moved quickly, defeating Bel-ibni and replacing him with Sennacherib's oldest son, Ashurnadin-shum. The next few years were relatively peaceful. The King used this time to prepare a decisive attack against Elam, which time and again had supported Babylonian rebellions. The overland route to Elam had been cut off and fortified by the Elamites. Sennacherib had ships built in Syria and at Nineveh. The ships from Syria were moved on rollers from the Euphrates to the Tigris. The fleet sailed downstream and was quite successful in the lagoons of the Persian Gulf and along the southern coastline of Elam. The Elamites launched a counteroffensive by land, occupying Babylonia and putting a man of their choice on the throne. Not until 693 were the Assyrians again able to fight their way through to the north. Finally, in 689, Sennacherib had his revenge. Babylon was conquered and completely destroyed, the temples plundered and levelled with the ground. The waters of the Arakhtu Canal were diverted over the ruins, and the inner city remained almost totally uninhabited for eight years. Even many Assyrians were indignant at this, believing that the Babylonian god Marduk must be grievously offended at the destruction of his temple and the carrying off of his image. Marduk was also an Assyrian deity, to whom many Assyrians turned in time of need. A political-theological propaganda campaign was launched to explain to the people that what had taken place was in accord with the wish of most of the gods. A story was written in which Marduk, because of a transgression, was captured and brought before a tribunal. Only a part of the commentary to this botched piece of literature is extant. Even the great poem of the creation of the world, the *Enuma elish,* was altered: the god Marduk was replaced by the god Ashur. Sennacherib's boundless energies brought no gain to his empire, however, and probably weakened it. The tenacity of the King can be seen in his building projects: for example, when Nineveh needed water for irrigation, Sennacherib had his engineers divert the waters of a tributary of the Great Zab River. The canal had to cross a valley at Jerwan. An aqueduct was constructed, consisting of about 2,000,000 blocks of limestone, with five huge, pointed archways over the brook in the valley. The bed of the canal on the aqueduct was sealed with cement containing magnesium. Parts of this aqueduct are still standing today. The King wrote of these and other technological accomplishments in minute detail, with illustrations.

The destruction of Babylon

Sennacherib built a huge palace in Nineveh, adorned with reliefs, some of them depicting the transport of colossal bull statues by water and by land. Many of the rooms were decorated with pictorial narratives in bas-relief telling of war and of building activities. Considerable advances can be noted in artistic execution, particularly in the portrayal of landscapes and animals. Outstanding are the depictions of the battles in the lagoons, the life in the military camps, and the deportations.

In 681 BC there was a rebellion. Sennacherib was assassinated by one or two of his sons in the temple of the god Ninurta. This god, along with the god Marduk, had been badly treated by Sennacherib, and the event was widely regarded as punishment of divine origin.

Esarhaddon. Ignoring the claims of his older brothers, an imperial council appointed Esarhaddon (Ashur-akhiddina, 680–669) as the successor. The choice is all the more difficult to explain in that Esarhaddon, quite unlike his father, was friendly toward the Babylonians. It can be assumed that his energetic and designing mother, Nakija, who hailed from Syria or Judah, used all her influence in his behalf to override the national party of Assyria. He may very well have been a partner in plotting the murder of his father; at any rate, he was able to procure the loyalty of his father's army. His brothers had to flee to Urartu. In his inscriptions, Esarhaddon always mentions both his father and grandfather.

Esarhaddon's policies

Defining the destruction of Babylon explicitly as punishment by the god Marduk, the new king soon ordered the

reconstruction of the city. He referred to himself only as governor of Babylonia and by his policies obtained the support of the cities of Babylonia. At the beginning of his reign the Aramaean tribes were still allied with Elam against him, but Urtaku of Elam (675–664) signed a peace treaty and freed him for campaigning elsewhere. In 679 he stationed a garrison at the Egyptian border, because Egypt, under the Ethiopian king Taharqa, was intervening in Syria. He put down with great severity a rebellion of the combined forces of Sidon, Tyre, and other Syrian cities. The time was ripe to attack Egypt, which was suffering under the rule of the Ethiopians and by no means a united country. Esarhaddon's first attempt in 674/673 miscarried. In 671 BC, however, his forces took Memphis, the Egyptian capital. Assyrian consultants were assigned to assist the princes of the 22 provinces, their main duty being the collection of tribute.

Occasional threats came from the mountainous border regions of eastern Anatolia and Iran. Pushed forward by the Scythians, the Cimmerians in northern Iran and Transcaucasia tried to get a foothold in Syria and western Iran. Esarhaddon allied himself with the Scythian king Partatua by giving him one of his daughters in marriage. In so doing he checked the movement of the Cimmerians. Nevertheless, the apprehensions of Esarhaddon can be seen in his many offerings, supplications, and requests to the sun-god. These were concerned less with his own enterprises than with the plans of enemies and vassals and the reliability of civil servants. The priestesses of Ishtar had to reassure the King constantly by calling out to him, "Do not be afraid." Previous kings had never needed this kind of encouragement.

At home the King was faced with serious difficulties from factions in the court. His oldest son had died early. The national party suspected his second son Shamashshum-ukin of being too friendly with the Babylonians; he may also have been considered unequal to the task of kingship. The King's third son, Ashurbanipal, was given the succession in 672, Shamash-shum-ukin remaining crown prince of Babylonia. This arrangement caused much dissension, and some farseeing civil servants warned of disastrous effects. Nevertheless, the Assyrian nobles, priests, and city leaders were sworn to just such an adjustment of the royal line; even the vassal princes had to take an oath of allegiance to Ashurbanipal.

Another matter of deep concern for Esarhaddon was his failing health. He regarded eclipses of the moon as particularly alarming omens, and in order to prevent a fatal illness from striking him at such times he had substitute kings chosen who ruled during the three such critical periods that occurred during his 12-year reign. The replacement kings died or were put to death after their brief term of office. During his off terms Esarhaddon called himself "Mister Peasant." This practice implied that the gods could not distinguish between the real king and a false one—quite contrary to the usual assumptions of the religion.

Esarhaddon enlarged and improved the temples in both Assyria and Babylonia. He also constructed a palace in Kalakh, using many of the picture slabs of Tiglath-pileser III. The works that remain are not on the level of those of either his predecessors or of Ashurbanipal. He died while on an expedition to put down a revolt in Egypt.

Ashurbanipal (668–627) and Shamash-shum-ukin (668–648). Although the death of his father occurred far from home, Ashurbanipal assumed his kingship as planned. He may have owed his fortunes to the intercession of his grandmother Nakija, who had recognized his superior capacities. He tells of his diversified education by the priests and his training in armour making as well as in other military arts. He may have been the only king in Assyria with a scholarly background. As crown prince he had also studied the administration of the vast empire. The record notes that the gods granted him a record harvest during the first year of his reign. There were also good crops in subsequent years. During these first years he was also successful in foreign policy, and his relation to his brother in Babylonia was good.

Ashurbanipal's education and early years

In 668 he put down a rebellion in Egypt and drove out King Taharqa. But in 664 the nephew of Taharqa, Tanutamon, gathered forces for a new rebellion. Ashurbanipal went to Egypt, pursuing the Ethiopian prince far into the south. His smashing victory moved Tyre and other parts of the empire to resume regular payments of tribute. Ashurbanipal installed Psamtik as prince over the Egyptian region of Sais. In 656 Psamtik dislodged the Assyrian garrisons with the aid of Carian and Ionian mercenaries, making Egypt again independent. Ashurbanipal did not attempt to reconquer it again. A former ally of Assyria, Gyges of Lydia had now aided Psamtik in his rebellion. In return, Assyria did not help Gyges when he was attacked by the Cimmerians. Gyges lost his throne and his life. His son Ardys decided that the payment of tribute to Assyria was a lesser evil than conquest by the Cimmerians.

Graver difficulties loomed in southern Babylonia, which was attacked by Elam in 664. Another attack came in 653, whereupon Ashurbanipal sent a large army that decisively defeated the Elamites. Their king was killed, and some of the Elamite states were encouraged to secede. Elam was no longer strong enough to assume an active part on the international scene. This victory had serious consequences for Babylonia. Shamash-shum-ukin had grown weary of being patronized by his domineering brother. He formed a secret alliance in 656 with the Iranians, Elamites, Aramaeans, Arabs, and Egyptians, directed against Ashurbanipal. The withdrawal of defeated Elam from this alliance was probably the reason for a premature attack by Shamash-shum-ukin at the end of the year 652, without waiting for the promised assistance from Egypt. Ashurbanipal, taken by surprise, soon pulled his troops together. The Babylonian army was defeated and Shamash-shum-ukin was surrounded in his fortified city of Babylon. His allies were not able to stand on their own against the Assyrians. Reinforcements of Arabian camel troops were also defeated. The city of Babylon was under siege for three years. It fell in 648 amid scenes of horrible carnage, Shamash-shum-ukin dying in his burning palace.

Defeat of Shamash-shum-ukin

After 648 the Assyrians made a few punitive attacks on the Arabs, breaking the forward thrust of the Arab tribes for a long time to come. The main objective of the Assyrians, however, was a final settlement of their relationship with Elam. The refusal of Elam in 647 to extradite an Aramaean prince was used as pretext for a new attack that drove deep into its territory. The assault on the solidly fortified capital of Susa followed, probably in 646. The Assyrians destroyed the city, including its temples and palaces. Vast spoils were taken. As usual, the upper classes of the land were exiled to Assyria and other parts of the empire, and Elam became an Assyrian province. Assyria had now extended its domain to southwestern Iran. Cyrus I of Persia sent tribute and hostages to Nineveh, hoping perhaps to secure protection for his borders with Media. Little is known about the last years of Ashurbanipal's reign.

Ashurbanipal left more inscriptions than any of his predecessors. His campaigns were not always recorded in chronological order but clustered in groups according to their purpose. The accounts were highly subjective. One of his more remarkable accomplishments was the founding of the great library in Nineveh, which is today one of the most important sources for the study of that era. The King himself supervised its construction. Important works were kept in more than one copy, some intended for the King's personal use. The work of arranging and cataloging drew upon the experience of centuries in the management of collections in huge temple archives such as the one in Ashur. In his inscriptions Ashurbanipal tells of becoming an enthusiastic hunter of big game, acquiring a taste for it during a fight with marauding lions. In his palace at Nineveh the long rows of hunting scenes show what a masterful artist can accomplish in bas-relief; with these reliefs Assyrian art reached its peak. In the series picturing his wars, particularly the wars fought in Elam, the scenes are overloaded with human figures. Those portraying the battles with the Arabian camel troops are magnificent in execution.

One reason for the durability of the Assyrian Empire was

State and culture in the Neo-Assyrian Empire

the practice of deporting large numbers of people from conquered areas and resettling others in their place. This kept many of the conquered nationalities from regaining their power. Equally important was the installation in conquered areas of a highly developed civil service under the leadership of trained officers. The highest ranking civil servant carried the title of *tartan,* a Hurrian word. The *tartan*s also represented the king during his absence. In descending rank were the palace overseer, the main cupbearer, the palace administrator, and the governor of Assyria. The generals often held high official positions, particularly in the provinces. The civil service numbered about 100,000, many of them former inhabitants of subjugated provinces. Prisoners became slaves but were later often freed.

No laws are known for the empire, although documents point to the existence of rules and standards for justice. Those who broke contracts were subject to severe penalties, even in cases of minor importance: the sacrifice of a son or the eating of a pound of wool and drinking of a great deal of water afterward, which led to a painful death. The position of women was inferior, except for the queen and some priestesses.

Good estimates of the economic situation of the period do not exist. The arts and crafts experience a golden age in Assyria; many qualified artists and artisans went there, bringing renewed creativity and new ideas. Finds of Aramaic texts from the Assyrian era are notably few because neither papyrus nor parchment preserved well in the climate; evidence of literary efforts in other languages is lacking, except of course in the border areas of Syria-Palestine and Judah.

Babylonia lagged behind Assyria during the period. Many small wars severely hampered the development of the country, undermining agricultural production and commerce. The temples fared best because many Assyrians supported them, but they left no document—a fact that can hardly be regarded as accidental. The continuing existence of scholarly study can be inferred from extant letters and the continuation of old traditions. The pictorial arts, however, were neglected, and the Babylonian artists may have found better jobs in Assyria. This period saw the first use of ancestral names as family appelations.

The era apparently was not fruitful for literary accomplishments either in Babylonia or Assyria. The large inscriptions of the kings were the most important achievements in Assyria; there were also hymns and prayers, executed mostly in the traditional style, and unusual oracle texts in the Assyrian language. Both countries saw the beginnings of a historical literature, going far beyond the ordinary chronicals.

Decline of the Assyrian Empire. Few historical sources remain for the last 30 years of the Assyrian Empire. There are no extant inscriptions of Ashurbanipal after 640 BC, and the few surviving inscriptions of his successors contain only vague allusions to political matters. In Babylonia the silence is almost total until 625 BC, when the chronicles resume. The rapid downfall of the Assyrian Empire was formerly attributed to military defeat, although it was never clear how the Medes and the Babylonians alone could have accomplished this. More recent work has established that after 635 a civil war occurred, weakening the empire, so that it could no longer stand up against a foreign enemy. Ashurbanipal had twin sons. Ashur-etel-ilani was appointed successor to the throne, but his twin brother Sin-shar-ishkun did not recognize him. The fight between them and their supporters forced the old king to withdraw to Haran, in 632 at the latest, perhaps ruling from there over the western part of the empire until his death in 627. Ashur-etel-ilani governed in Assyria from about 633 on, but a general, Sin-shum-lisher, soon rebelled against him and proclaimed himself counterking. Some years later (629?) Sin-sharishkun finally succeeded in obtaining the kingship. In Babylonian documents dates can be found for all three kings. To add to the confusion, until 626 there are also dates of Ashurbanipal and a king named Kandalanu. In 626 the Chaldean Nabopolassar revolted from Uruk and occupied Babylon. There were several changes in government. King Ashur-etel-ilani was forced to withdraw to the west, where he died sometime after 625.

Internal struggles in the last years of the empire

Around the year 626 the Scythians laid waste to Syria and Palestine. In 625 the Medes became united under Cyaxares and began to conquer the Iranian provinces of Assyria. One chronicle relates of wars between Sin-sharishkun and Nabopolassar in Babylonia in 625–623. It was not long until the Assyrians were driven out of Babylonia. In 616 the Medes struck against Nineveh, but, according to the Greek historian Herodotus, were driven back by the Scythians. In 615, however, the Medes conquered Arrapkha (Kirkūk) and in 614 the old capital of Ashur, looting and destroying the city. Now Cyaxares and Nabopolassar made an alliance for the purpose of dividing Assyria. In 612 Kalakh and Nineveh succumbed to the superior strength of the allies. The revenge taken on the Assyrians was terrible: 200 years later Xenophon found the country still sparsely populated.

Sin-shar-ishkun, king of Assyria, found death in his burning palace. The commander of the Assyrian army in the west crowned himself king in the city of Haran, assuming the name of the founder of the empire, Ashuruballit II (611–609 BC). Ashuruballit had to face both the Babylonians and the Medes. They conquered Haran in 610, without, however, destroying the city completely. In 609 the remaining Assyrian troops had to capitulate. With this event, Assyria disappeared from history. The great empires that succeeded it learned a great deal from the hated Assyrians, both in the arts and in the organization of their states.

THE NEO-BABYLONIAN EMPIRE (626–539)

The Chaldean kings

The Chaldeans, who inhabited the coastal area near the Persian Gulf, had never been entirely pacified by the Assyrians. About 630 Nabopolassar became king of the Chaldeans. In 626 he forced the Assyrians out of Uruk and crowned himself king of Babylonia. He took part in the wars aimed at the destruction of Assyria. At the same time, he began to restore the dilapidated network of canals in the cities of Babylonia, particularly those in Babylon itself. He fought against the Assyrian Ashuruballit II and then against Egypt, his successes alternating with misfortunes. In 605 Nabopolassar died in Babylon.

Campaigns in Syria and Palestine

Nebuchadrezzar (Nabu-kudurri-usur) II. Nabopolassar had named his oldest son after the famous king of the second dynasty of Isin, had trained him carefully for his prospective kingship, and had shared responsibility with him. When the father died in 605, Nebuchadrezzar was with his army in Syria; he had just crushed the Egyptians in a cruel, bloody war and pursued them into the south. On receiving the news of his father's death, Nebuchadrezzar returned immediately to Babylon. In his numerous inscriptions he tells but rarely of his many wars. The Babylonian chronicle is extant only for the years 605–594, and not much is known from other sources about the later years of this famous king. He went very often to Syria and Palestine—at first to drive out the Egyptians. In 604 he took the Philistine city of Ashkelon. In 601 he tried to push forward into Egypt but was forced to pull back after a bloody, undecided battle and to regroup his army in Babylonia. After smaller incursions against the Arabs of Syria, he attacked Palestine at the end of 598. Joiakin of Judah had rebelled, counting on help from Egypt. According to the chronicle, Jerusalem was taken on March 16, 597, and King Joiakin, together with at least 3,000 Jews, were led into exile in Babylonia. They were treated well there, according to the documents. Zedekiah was appointed the new king. In 596, when danger threatened from the east, Nebuchadrezzar marched to the Tigris River and induced the enemy to withdraw. After a revolt in Babylonia had been crushed with much bloodshed, there were other campaigns in the west.

According to the Old Testament, Judah rebelled again in 589; Jerusalem was under siege. The city fell after 18 months of siege and was completely destroyed. Many thousands of Jews were forced into "Babylonian exile," and their country was reduced to a province of the Babylonian Empire. The revolt had been caused by an Egyptian invasion that pushed as far as Sidon. Nebuchadrezzar laid

siege to Tyre for 13 years without taking the city, because there was no fleet at his disposal. In 568/567 he attacked Egypt, again without much success, but from then on the Egyptians refrained from further attacks on Palestine. Nebuchadrezzar lived at peace with Media throughout his reign and acted as a mediator after the Median-Lydian war of 590–585.

The Babylonian Empire under Nebuchadrezzar extended to the Egyptian border. It had a well-functioning administrative system. Though he had to collect extremely high taxes and tributes in order to maintain his armies and carry out his building projects, Nebuchadrezzar made Babylonia one of the richest lands in western Asia—the more astonishing because it had been rather poor when it was ruled by the Assyrians. Babylon was the largest city of the "civilized world." The King maintained the existing canal systems and built many supplementary canals, making the land even more fertile. Trade and commerce flourished during his reign.

Nebuchadrezzar's building activities surpassed those of most of the Assyrian kings. He fortified the old double walls of Babylon, adding another triple wall outside the old wall. In addition, he erected another wall, the Median Wall, north of the city between the Euphrates and the Tigris rivers. According to Greek estimates, the Median Wall may have been about 30 metres (100 feet) high. He enlarged the old palace and added many wings, so that hundreds of rooms with large inner courts were now at the disposal of the central offices of the empire. Colourful glazed-tile bas-reliefs decorated the walls. Terrace gardens, called hanging gardens in later accounts, were added. Hundreds of thousands of workers must have been required for these projects. The temples were objects of special concern. He devoted himself first and foremost to the completion of the Tower of Babel. Its base measured 91 metres (about 300 feet) on each side and 91 metres in height. There were five terrace-like gradations surmounted by a temple, the whole tower being about twice the height of those of other temples. The wide street used for processions led along the eastern side by the inner city walls and crossed at the enormous Ishtar Gate with its world-renowned bas-relief tiles. The King also built many smaller temples throughout the country.

The Tower of Babel

The last kings of Babylonia. Awil-Marduk (in the Old Testament Evil-Merodach, 561–560), the son of Nebuchadrezzar, was unable to win the support of the priests of Marduk. His reign did not last long, and he was soon eliminated. His brother-in-law and successor, Nergal-sharusur (in classical sources, Neriglissar, 559–556), was a general who undertook a campaign in 557 into the "rough" Cilician land, which may have been under the control of the Medes. His land forces were assisted by a fleet. His still-minor son Labashi-Marduk was murdered not long after that, allegedly because he was not suitable for his job.

Nabonidus and Belshazzar

The next king was the Aramaean Nabonidus (555–539) from Haran, one of the most interesting and enigmatical figures of ancient times. His mother, Addagoppe, was a priestess of the god Sin in Haran, who came to Babylon and managed to get responsible offices for her son at court. The god of the moon rewarded her piety with a long life—she lived to be 102—and she was buried in Haran with all the honours of a queen in 547. It is not clear which powerful faction in Babylon supported the kingship of Nabonidus; it may have been one opposing the priests of Marduk, who had become extremely powerful. The King raided Cilicia in 555 and secured the surrender of Haran, which had been ruled by the Medes. He concluded a treaty of defense with Astyages of Media against the Persians, who had become a growing threat since 559 under their king Cyrus II. He also devoted himself to the renovation of many temples, taking an especially keen interest in old inscriptions. He gave preference to his god Sin and had powerful enemies in the priesthood of the Marduk temple. Modern excavators have found fragments of propaganda poems written against Nabonidus and also in support of him. Both traditions continued in Judaism.

Internal difficulties and the recognition that the narrow strip of land from the Persian Gulf to Syria could not be defended against a major attack from the east induced Nabonidus to leave Babylonia around 552 and to reside in Taima in northern Arabia. There he organized an Arabian province with the assistance of Jewish merecenaries. His viceroy in Babylonia was his son Belsharusur, the Belshazzar of the Book of Daniel in the Bible. Cyrus turned this to his own advantage by annexing Media in 550. Nabonidus, in turn, allied himself with Croesus of Lydia in order to fight Cyrus. But, when Cyrus attacked Lydia and annexed it in 546, Nabonidus was not able to help Croesus. Cyrus bode his time. In 542 Nabonidus returned to Babylonia, where his son had been able to maintain good order in external matters but had not overcome a growing internal opposition to his father. Consequently, Nabonidus' career after his return was of short duration, though he tried very hard to regain the support of the Babylonians. He appointed his daughter to be high priestess of the god Sin in Ur, thus returning to the Sumerian-Old Babylonian religious custom. The priests of Marduk looked to Cyrus, hoping to have better relations with him than with Nabonidus; they promised Cyrus the surrender of Babylon without a fight if he would grant them their privileges in return. In 539 Cyrus attacked northern Babylonia with a large army, defeating Nabonidus, and entered the city of Babylon without a battle. The other cities did not offer any resistance, either. Nabonidus surrendered, receiving a small territory in eastern Iran. Tradition has confused him with his great predecessor Nebuchadrezzar II. The Bible refers to him as Nebuchadrezzar in the Book of Daniel.

Babylonia's peaceful submission to Cyrus saved it from the fate of Assyria. It became a territory under the Persian crown but kept its cultural autonomy. Even the racially mixed western part of the Babylonian Empire submitted without resistance.

Culture of the Neo-Babylonian Empire

The Babylonians by 620 had grown tired of Assyrian rule. They were also weary of internal struggle. They were easily persuaded to submit to the order of the Chaldean kings. The result was a surprisingly quick social and economic consolidation, helped along by the fact that after the fall of Assyria no external enemy threatened Babylonia for more than 60 years. In the cities, the temples were an important element of the economy, having vast benefices at their disposal. The business class regained its strength, not only in the trades and commerce but also in the management of agriculture in the metropolitan areas. Livestock breeding—sheep, goats, beef cattle, and also horses—flourished, as did poultry farming. The cultivation of corn, dates, and vegetables grew in importance. Much was done to improve communications, both by water and by land, with the western provinces of the empire. The collapse of the Assyrian Empire had the consequence that many trade arteries were rerouted through Babylonia instead. Another consequence of the collapse was that the city of Babylon became a world centre.

The immense amount of documentary material and correspondence that has survived has not yet been fully analyzed. No new system of law or administration seems to have developed during that time. The Babylonian dialect gradually became Aramaicized, still written primarily on clay tablets that often bore added material in Aramaic lettering. Parchment and papyrus documents have not survived. In contrast to advances in other fields, there is no evidence of much artistic creativity. Aside from some of the inscriptions of the kings, especially Nabonidus, which were not comparable from a literary standpoint with those of the Assyrians, the main efforts were devoted to the rewriting of old texts. In the fine arts, only few monuments have any suggestion of new tendencies.

(W.T.v.S.)

MESOPOTAMIAN ART AND ARCHITECTURE

The name Mesopotamia has been used with varying connotations by ancient writers. If, for convenience, it is to be considered synonymous with the modern state of Iraq, it can be seen in terms of two fairly well defined provinces: a flat alluvial plain in the south and, in the north, the uplands through which the country's twin rivers flow in their middle courses. This geographical division of the

area is reflected in the history of its cultural development from the earliest times.

The first traces of settled communities are found in the northern region and date from the mid-6th millennium BC, a period that archaeologists associate with the transition from a Neolithic to a Chalcolithic age. It is of some importance that this period also corresponds to the earliest use of painted ornament on pottery vessels, since the designs used for this purpose are the most reliable criteria by which ethnological groupings and migratory movements can be distinguished. Archaeologically, such groupings are, for the most part, arbitrarily named after the site at which traces of them were first found, and the same names are sometimes attributed to the prehistoric periods during which they were predominant. Hence, Hassuna, Hassuna-Samarra, and Halaf in northern Iraq are names given to the first three periods during which known early settlements were successively occupied by peoples whose relations were apparently with Syria and Anatolia. The designs on their pottery, sometimes in more than one colour, usually consist of zones filled with "geometric" ornament in patterns reminiscent of woven fabrics, and they are often adapted to the shape of the vessels with creditable artifice. Only in Hassuna-Samarran pottery do devices occasionally appear that consist of animal, bird, or even human figures, ingeniously stylized and aesthetically attractive; but such motifs appear to be adopted from contemporary Iranian ceramics. The only other notable art form popular at this time is that of hominoid figurines of stone or clay, associated with primitive religious cults; however, their formal idiosyncracies vary greatly from group to group, and the meaning of their symbolism is unknown. Nor can they—or the pottery designs—be considered as ancestral to Mesopotamian art of historical times, the antecedents of which must be sought in southern Iraq.

Here, in the delta, the earliest phase of prehistory is associated with the name Ubaid 1; and, since this phase has a parallel in Susiana, north of the Iranian frontier, the first settlers in both areas may have a common origin. Among these settlers, according to some scholars, was the germ of Sumerian genius, but this is not indisputably authenticated until the end of the 4th millennium. By 3100 BC, however, the presence of the Sumerians is finally proved by the invention of writing as a vehicle for their own language; from then onward, successive phases in the evolution of Sumerian art can satisfactorily be studied.

Three factors influencing Mesopotamian art

Three factors may be recognized as contributing to the character of Mesopotamian art and architecture. One is the sociopolitical organization of the Sumerian city-states and of the kingdoms and empires that succeeded them. From the earliest times, cities were fortified by and adorned with public buildings; irrigation systems were organized and jealously protected; armies were efficiently equipped and troops trained in concerted action; victories were celebrated and treaties ratified. Because interestate warfare or foreign conquests were primary preoccupations of Mesopotamian rulers, it is understandable that in most periods a certain class of artworks was dedicated simply to the glorification of their military prowess. But a second and even more important factor is the major role played by organized religion in Mesopotamian affairs of state. For, particularly in Sumerian times, the municipal and economic organization of a city was the responsibility of the temple, with its hierarchical priesthood in which was vested an authority almost equal to that of the ruler and his advisory council of elders. Accordingly, in the early days of Sumer and Babylonia, architectural attention was paid primarily to religious buildings, and all sculpture served religious purposes. The elaboration and adornment of palaces was an innovation of Assyrian times (see below *Assyrian period*).

The third factor that contributed to the character of Mesopotamian art is the influence of the natural environment. The practical limitations imposed upon both artist and architect by the geology and climate of southern Iraq are immediately apparent. Since no stone or wood was available in the alluvial plain, sculptors were dependent on scarce imported material or compelled to use such substitutes as terra-cotta (baked clay). Architecture also was profoundly affected, first, by the restriction of building material to brickwork and, second, by problems of roof construction, only partially solved by the contrivance of brick vaulting, in the 2nd millennium BC. For the Assyrians, in the north, good quality stone was plentiful, but the cost of quarrying and transport, combined with an obstinate conservatism, caused it to be regarded as a luxury material and its use to be confined to sculptured ornament and conspicuous architectural features.

An equally apparent, if more abstract, association between Mesopotamian art and environment can be detected when the intellectual climate engendered by the latter is understood. In a country where summer and winter temperatures reach thermometrical extremes, where agriculture depends exclusively on the artificial distribution of river water and contends precariously with the timing of seasonal floods, where the herdsman is afflicted by the depredations of wild beasts and the cultivator by the menace of poisonous insects—in such a country, the inhabitants must have felt themselves in perpetual conflict with hostile and potentially destructive elements in nature. All of this confrontation and frustration is reflected in the melancholy undertones of their religious beliefs, particularly those of the Sumerians, for whom success and prosperity came to be identified with the principle of fertility and thus could only be attained by the appeasement of capricious deities. Such convictions are inherent in the fabric of their complicated mythology, which lends itself easily to expression in pictorial form and provides the predominant subject of almost all Sumerian art. Furthermore, since their mythical traditions and religious beliefs persisted for many centuries after the demise of the Sumerians themselves, they provided the basic imagery of almost all Mesopotamian art.

Sumerian period. *Architecture.* The beginnings of monumental architecture in Mesopotamia are usually considered to have been contemporary with the foundation of the Sumerian cities and the invention of writing, in about 3100 BC. Conscious attempts at architectural design during this so-called Protoliterate period (*c.* 3400–*c.* 2900 BC) are recognizable in the construction of religious buildings. There is, however, one temple, at Abū Shahrayn (ancient Eridu), that is no more than a final rebuilding of a shrine the original foundation of which dates back to the beginning of the 4th millennium; the continuity of design has been thought by some to confirm the presence of the Sumerians throughout the temple's history. Already, in the Ubaid period (*c.* 5200–*c.* 3500 BC), this temple anticipated most of the architectural characteristics of the typical Protoliterate Sumerian platform temple. It is built of mud brick on a raised plinth (platform base) of the same material, and its walls are ornamented on their outside surfaces with alternating buttresses (supports) and recesses. Tripartite in form, its long central sanctuary is flanked on two sides by subsidiary chambers, provided with an altar at one end and a freestanding offering table at the other. Typical temples of the Protoliterate period—both the platform type and the type built at ground level—are, however, much more elaborate both in planning and ornament. Interior wall ornament often consists of a patterned mosaic of terra-cotta cones sunk into the wall, their exposed ends dipped in bright colours or sheathed in bronze. An open hall at the Sumerian city of Uruk (biblical, Erech; Arabic, Tall al-Warkāʾ) contains freestanding and attached brick columns that have been brilliantly decorated in this way. Alternatively, the internal-wall faces of a platform temple could be ornamented with mural paintings depicting mythical scenes, such as at ʿUqair.

Sumerian platform temple

The two forms of temple—the platform variety and that built at ground level—persisted throughout the early dynasties of Sumerian history (*c.* 2900–*c.* 2400 BC). It is known that two of the platform temples originally stood within walled enclosures, oval in shape and containing, in addition to the temple, accommodation for priests. But the raised shrines themselves are lost, and their appearance can only be judged from facade ornaments discov-



Figure 1: Representation of lion-headed bird Imdugud and two stags, copper relief from al-'Ubaid that combines bas-relief with sculpture in the round, Early Dynastic period (*c.* 2900–*c.* 2400 BC). In the British Museum. Height 1.07 m.

By courtesy of the trustees of the British Museum

ered at al-'Ubaid (Figure 1). These devices, which were intended to relieve the monotony of sun-dried brick or mud plaster, include a huge copper-sheathed lintel, with animal figures modelled partly in the round; wooden columns sheathed in a patterned mosaic of coloured stone or shell; and bands of copper-sheathed bulls and lions, modelled in relief but with projecting heads. The planning of ground-level temples continued to elaborate on a single theme: a rectangular sanctuary, entered on the cross axis, with altar, offering table, and pedestals for votive statuary (statues used for vicarious worship or intercession). Considerably less is known about palaces or other secular buildings at this time. Circular brick columns and austerely simplified facades have been found at Kish. Flat roofs, supported on palm trunks, must be assumed, although some knowledge of corbelled vaulting (a technique of spanning an opening like an arch by having successive cones of masonry project farther inward as they rise on each side off the gap)—and even of dome construction—is suggested by tombs at Ur, where a little stone was available.

Sculpture. Practically all Sumerian sculpture served as adornment or ritual equipment for the temples. No clearly identifiable cult statues of gods or goddesses have yet been found. Many of the extant figures in stone are votive statues, as indicated by the phrases used in the inscriptions that they often bear: "It offers prayers" or "Statue, say to my king (god). . . ." Male statues stand or sit with hands clasped in an attitude of prayer. They are often naked above the waist and wear a woollen skirt curiously woven in a pattern that suggests overlapping petals (commonly described by the Greek word, *kaunakes,* meaning "thick cloak"). A toga-like garment sometimes covers one shoulder. Men generally wear long hair and a heavy beard, both often trimmed in corrugations and painted black. The eyes and eyebrows are emphasized with coloured inlay. The female coiffure varies considerably but predominantly consists of a heavy coil arranged vertically from ear to ear and a chignon behind. The hair is sometimes concealed by a headdress of folded linen. Ritual nakedness is confined to priests.

Sumerian stone sculpture

It has been thought that the rarity of stone in Mesopotamia contributed to the primary stylistic distinction between Sumerian and Egyptian sculpture. The Egyptians quarried their own stone in prismatic blocks, and one can see that, even in their freestanding statues, strength of design is attained by the retention of geometrical unity. By contrast, in Sumer, stone must have been imported from remote sources, often in the form of miscellaneous boulders, the amorphous character of which seems to be retained by the statues into which they were transformed.

Beyond this general characteristic of Sumerian sculpture, two successive styles have been distinguished in the middle and late subdivisions of the Early Dynastic period. One very notable group of figures, from Tall al-Asmar, dating from the first of these phases, shows a geometrical simplification of forms that, to modern taste, is ingenious and aesthetically acceptable (Figure 2). Statues characteristic of the second phase, on the other hand, though technically more competently carved, show aspirations to naturalism that are sometimes overly ambitious. In this second style, some scholars see evidence of occasional attempts at portraiture. Yet, in spite of minor variations, all these figures adhere to the single formula of presenting the conventional characteristics of Sumerian physiognomy. Their provenance is not confined to the Sumerian cities in the south. An important group is derived from the ancient capital of Mari, on the middle Euphrates, where the population is known to have been racially different from the Sumerians; yet, here also there appears to have been no deviation from the sculptural formula, and the Mari statues are distinguished only by technical peculiarities in the carving.

Metal sculpture

Deprived of stone, Sumerian sculptors exploited alternative materials. Fine examples of metal casting have been found, some of them suggesting knowledge of the cire perdue (lost-wax) process, and copper statues more than half life-size are known to have existed. In metalwork, however, the ingenuity of Sumerian artists is perhaps best judged from their contrivance of composite figures. The earliest and one of the finest examples of such figures—and of Sumerian sculpture as a whole—comes from a Protoliterate level of excavation at Tall al-Warkā'. It is the limestone face of a life-size statue (Iraq Museum, Baghdad), the remainder of which must have been composed of other materials; the method of attachment is visible on the surviving face. Devices of this sort were brought to perfection by craftsmen of the Early Dynastic period, the finest examples of whose work are to be seen among the treasures from the royal tombs at Ur (Figure 3): a bull's head decorating a harp, composed of wood or bitumen covered with gold and wearing a lapis lazuli beard (British Museum); a rampant he-goat in gold and lapis, supported by a golden tree (University Museum, Philadelphia); the composite headdresses of court ladies (British Museum; Iraq Museum, Baghdad; and University Museum, Philadelphia); or, more simply, the miniature figure of a wild ass, cast in electrum (a natural yellow alloy of gold and silver) and mounted on a bronze rein ring (British Museum). The inlay and enrichment of wooden objects reaches its peak in this period, as may be seen in the so-called standard or double-sided panel from Ur (British Museum), on which elaborate scenes of peace and war are depicted in a delicate inlay of shell and semiprecious stones (Figure 4). The refinement of craftsmanship in metal is also apparent in the famous wig-helmet of gold (Iraq Museum), belonging to a Sumerian prince, and in weapons, implements, and utensils.

Relief carving in stone was a medium of expression popular with the Sumerians and first appears in a rather crude form in Protoliterate times. In the final phase

By courtesy of the Oriental Institute, the University of Chicago

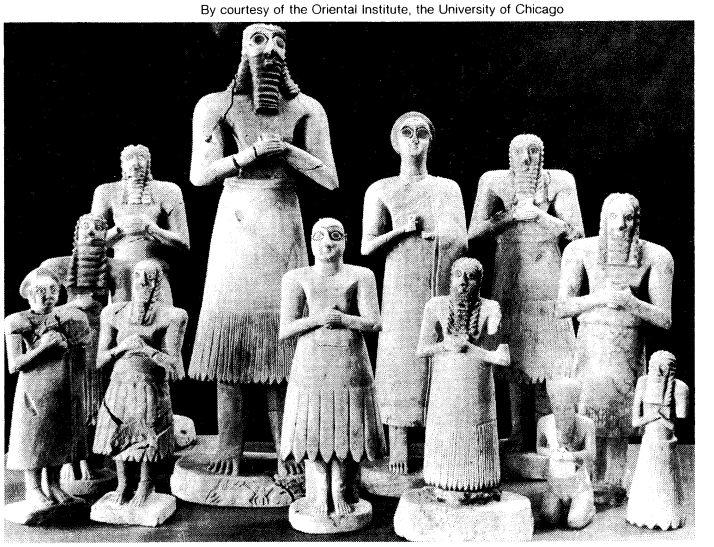


Figure 2: Stone statuettes of worshippers and perhaps deities, from the Temple of Abu, Tall al-Asmar, Iraq, Early Dynastic period (*c.* 2900–*c.* 2400 BC). In the Iraq Museum, Baghdad. Height of tallest statuette 72.1 cm.

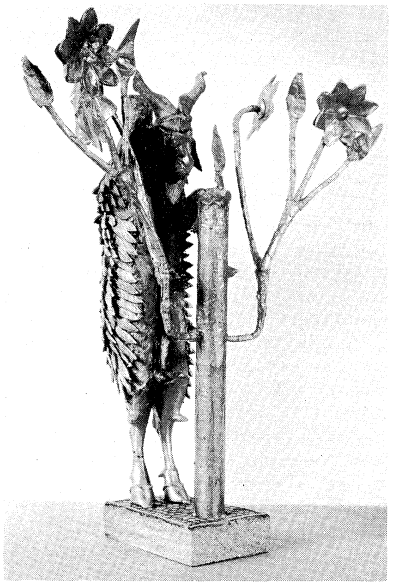


Figure 3: "Goat and Tree," offering stand from Ur (now Tall al-Muqayyar, Iraq) made of wood, gold, and lapis lazuli, *c.* 2500 BC. In the University Museum of the University of Philadelphia, Pennsylvania. Height 50.8 cm.
By courtesy of the University Museum, Philadelphia

of the Early Dynastic period, its style became conventional. The most common form of relief sculpture was that of stone plaques, a foot or more square, pierced in the centre for attachment to the walls of a temple, with scenes depicted in several registers (horizontal rows). The subjects usually seem to be commemorative of specific events, such as feasts or building activities, but representation is highly standardized, so that almost identical plaques have been found at sites as much as 500 miles apart. Fragments of more ambitious commemorative stelae have also been recovered; the Stele of Vultures (Louvre) from Tello, for example. Although it commemorates a military victory, it has a religious content. Emphasized by its size, there is a figure of a patron deity rather than that of the king. The formal massing of figures suggests the beginnings of mastery in design, and a formula has been devised for mutiplying identical figures, such as chariot horses.

Cylinder seals

In a somewhat different category are the cylinder seals so widely utilized at this time. Used like the more familiar stamp seal and, likewise, engraved in negative (intaglio), the cylinder-shaped seal was rolled over wet clay on which it left an impression in relief. Delicately carved with miniature designs on a variety of stones or shell, they rank as one of the higher forms of Sumerian art.

Prominent among their subjects is the complicated imagery of Sumerian mythology and religious ritual. Still only partially understood, their skillful adaptation to linear designs can at least be easily appreciated. Some of the finest date from the Protoliterate period (Figure 5, top). After a slight deterioration in the first phase of the Early Dynasties, when brocade patterns or files of running animals were preferred (Figure 5, bottom), mythical scenes returned. Conflicts are depicted between wild beasts and protecting demigods or hybrid figures associated by some with the Sumerian *Epic of Gilgamesh*. The monotony of animated motifs is occasionally relieved by the introduction of an inscription.

Akkadian period. Sargon of Akkad's (*c.* 2334–*c.* 2279 BC) unification of the Sumerian city-states and creation of a first Mesopotamian empire profoundly affected the art of his people, as well as their language and political thought. The increasingly large proportion of Semitic elements in the population were in the ascendancy, and their personal loyalty to Sargon and his successors replaced the regional patriotism of the old cities. The new conception of kingship thus engendered is reflected in artworks of secular grandeur, unprecedented in the godfearing world of the Sumerians.

Architecture. One would indeed expect a similar change to be apparent in the character of contemporary architecture, and the fact that this is not so may be due to the paucity of excavated examples. It is known that the Sargonid dynasty had a hand in the reconstruction and extension of many Sumerian temples (*e.g.*, Nippur) and that they built palaces with practical amenities (Tall al-Asmar) and powerful fortresses on their lines of imperial

By courtesy of (top) the Staatliche Museen, Berlin, (bottom) the Oriental Institute, the University of Chicago

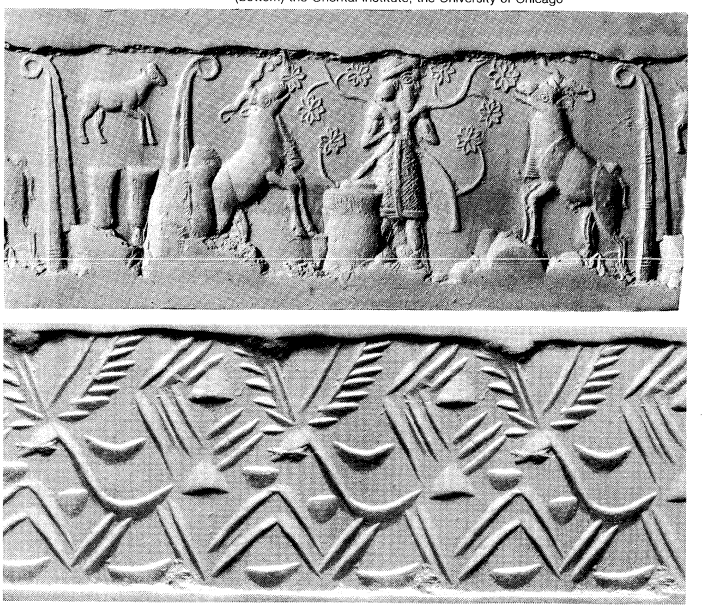


Figure 5: *Sumerian cylinder seal impressions.*
(Top) Feeding the sacred herd, seal impression from Tall al-Warkāʾ, Iraq, Protoliterate period (before *c.* 2900 BC).
(Bottom) Horned animals engraved in the brocade style, seal impression from Tall al-Asmar, Iraq, first Early Dynastic period (*c.* 2900–*c.* 2750 BC).

By courtesy of the trustees of the British Museum

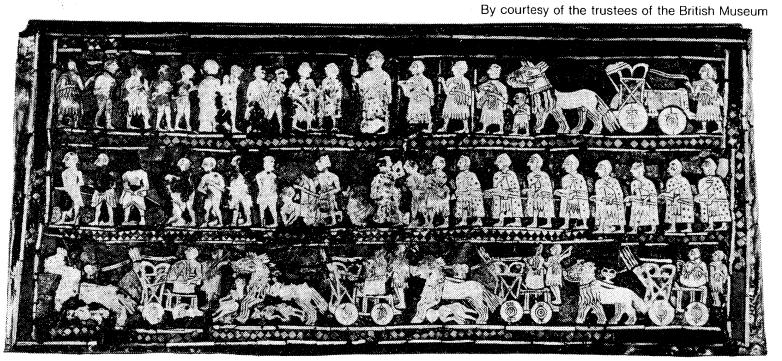


Figure 4: War side of the "Standard of Ur," mosaic of lapis lazuli, shell, coloured stone, and mother-of-pearl, from the Royal Cemetery, Ur, *c.* 2500 BC (Early Dynastic).



Figure 6: Bronze head of a king, perhaps Sargon of Akkad, from Nineveh (now in Iraq), Akkadian period, *c.* 2300 BC. In the Iraq Museum, Baghdad. Height 30.5 cm.
By courtesy of the Directorate General of Antiquities, Baghdad, Iraq

communication (Tell Brak, or Tall Birāk at-Tahtānī, Syria); their ruins, however, are insufficient to suggest either changes in architectural style or structural innovations.

Sculpture. Two notable heads of Akkadian statues have survived: one in bronze and the other of stone. The bronze head of a king (Figure 6), wearing the wig-helmet of the old Sumerian rulers, is probably Sargon himself (Iraq Museum). Though lacking its inlaid eyes and slightly damaged elsewhere, this head is rightly considered as one of the great masterpieces of ancient art. The Akkadian head (Iraq Museum) in stone, from Bīsmāyah, suggests that portraiture in materials other than bronze had also progressed. But, where relief sculpture is concerned, an even greater accomplishment is evident in the famous Naram-Sin (Sargon's son) stele (Louvre), on which a pattern of figures is ingeniously designed to express the abstract idea of conquest (Figure 7).

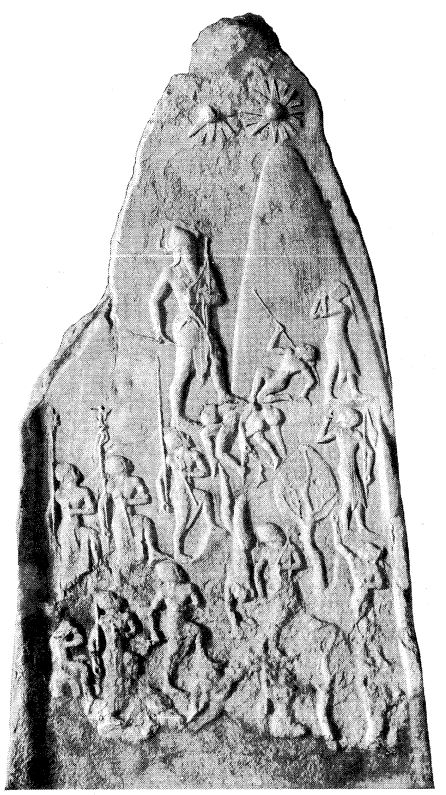


Figure 7: "Victory Stele of Naram-Sin," stone, Akkadian period, *c.* 2371–*c.* 2255 BC. In the Louvre, Paris. Height 1.98 m.
By courtesy of the Louvre, Paris

Other stelae and the rock reliefs (which by their geographical situation bear witness to the extent of Akkadian conquest) show the carving of the period to be in the hands of less competent artists. But two striking fragments in the Iraq Museum, which were found in the region of Nāṣirīyah, once more provide evidence of the improvement in design and craftsmanship that had taken place since the days of the Sumerian dynasties. One of the fragments shows a procession of naked war prisoners, in which the anatomical details are well observed but skillfully subordinated to the rhythmical pattern required by the subject.

Some compensation for the paucity of surviving Akkadian sculptures is to be found in the varied and plentiful repertory of contemporary cylinder seals. The Akkadian seal cutter's craft reached a standard of perfection virtually unrivalled in later times. Where the aim of his Sumerian predecessor had been to produce an uninterrupted, closely woven design, the Akkadian seal cutter's own preference was for clarity in the arrangement of a number of carefully spaced figures.

Posthumous flowering of Sumerian genius

The Akkadian dynasty ended in disaster when the river valley was overrun by the barbaric mountain tribes of northern Iran. Of all the Mesopotamian cities, only Lagash appears somehow to have remained aloof from the conflict and, under its famous governor, Gudea, to have successfully maintained the continuity of the Mesopotamian cultural tradition (Figure 8, left). In particular, the sculpture dating from this short interregnum (*c.* 2200 BC) seems to represent some sort of posthumous flowering of Sumerian genius. The well-known group of statues of the governor and other notables, discovered at the end of the 19th century, long remained the only criterion by which Sumerian art could be judged, and examples in the Louvre and British Museum are still greatly admired. The hard stone, usually diorite, is carved with obvious mastery and brought to a fine finish. Details are cleverly stylized; but the musculature is carefully studied, and the high quality of the carving makes the use of inlay unnecessary. The powerful impression of serene authority that these statues convey justifies their inclusion among the finest products of ancient Near Eastern art.

Sumerian revival. The short historical interlude represented by the Gudea sculptures was followed by a fullscale Sumerian revival, one which lasted for four centuries and culminated in the unification of the whole country under the rule of Hammurabi in the early 18th century BC. Dominated first by the powerful Third dynasty of Ur and later by the rival states of Isin and Larsa, the peoples of ancient Sumer reverted to their pre-Akkadian cultural traditions. On their northern frontiers the Sumerian culture was extended to increasingly prosperous younger city-states, such as Mari, Ashur, and Eshnunna (Tall al-Asmar), located on the middle courses of the Tigris and Euphrates rivers.

In all these cities, both old and new, the period is notable for the advances made in architectural planning and the large-scale reconstruction of ancient buildings. In the south, the early promise of Sumerian architecture had reached fulfillment, first in the great ziggurats, or stepped towers, rising above their walled temple enclosures at cities such as Ur, Eridu, Kish, Uruk, and Nippur. These huge structures, with their summit sanctuaries the appearance of which can only be guessed at, were faced with kiln-baked brick, panelled and recessed to break the monotony of their colossal facades, and were strengthened with bitumen and reinforced with twisted reeds. Tradition associates the ziggurat at Borsippa (Birs Nimrūd), near Babylon, with the biblical Tower of Babel. Surrounding temples at ground level were also much elaborated. The basic plan consisted of a tower-flanked entry, central court, antecella (or inner vestibule), and sanctuary, all arranged on a single axis; but in the larger examples this plan could be expanded by means of communicating courtyards. Facades were often elaborately decorated with panels of pilasters (recessed columns) or engaged half columns, skillfully modelled in mud-brick. At Ur, kiln-baked brick was again used to construct corbelled vaulting over huge subterranean tomb chambers, entered



Figure 8: *Ancient Mesopotamian official sculpture.* (Left) Diorite statue of Gudea, ruler of Lagash (now in Iraq), Post-Akkadian period, *c.* 2144–2124 BC. In the Louvre, Paris. Height 1.05 m. (Right) Inscribed diorite statue of Puzur-Ishtar, governor of Mari, Syria, Ur III period, *c.* 2000 BC. Reconstructed head (original in Ehemals Staatliches Museum, Berlin) with original body. In the Archaeological Museums of Istanbul. Height 1.73 m.

By courtesy of (right) the Archaeological Museum, Istanbul; photograph, (left) Cliches Musees Nationaux

through funerary chapels at ground level. Here, too, there are temple-palaces, the complicated planning of which is seldom self-explanatory.

Zimrilim palace

The better examples of residential palaces are found in the newer cities of the north, especially Mari, where a vast building with more than 200 rooms was constructed by a ruler called Zimrilim (*c.* 1779–*c.* 1761 BC). In this palace is found the standard reception unit common to all Babylonian palaces: a rectangular throne room that is entered by a central doorway from a square court of honour; and behind it a great hall, in this case apparently serving some religious purpose. There is also an immense outer courtyard, overlooked by a raised audience chamber, and, in the remotest corner of the building, a heavily protected residential suite. In some of the main chambers, mural paintings depicting ritual scenes and processions have been preserved.

The sculpture of this period is perhaps best represented by some well-preserved statues, also from the Mari palace (Figure 8, right). Their style owes much to the preceding Gudea period in the south, but they lack the authentic stamp of Sumerian design and workmanship. The same may be said of the few surviving pieces from the time of Hammurabi (early 18th century BC), whose conquests ended the epoch; *e.g.,* the relief at the head of the stela in the Louvre on which his law code is inscribed.

Assyrian period. *Architecture.* Ashur, a small Sumerian city-state on the middle Euphrates, began to gain political prominence during the pre-Hammurabi period just discussed. During the latter half of the 2nd millennium BC, the frontiers of Assyria were extended to include the greater part of northern Mesopotamia, and, in the city of Ashur itself, excavations have revealed the fortifications and public buildings constructed or rebuilt by a long line of Assyrian kings. The character of these buildings suggests a logical development of Old Babylonian architecture. There are certain innovations, such as the incorporation of small twin ziggurats in the design of a single temple, while in the temples themselves the sanctuary was lengthened on its main axis, and the altar itself was withdrawn into a deep recess. For the rest, the absence of ornament and the multiplication of buttressed facades with crenellated battlements tend to monotony.

Other forms of art are inconspicuous, except, perhaps, the contemporary cylinder seals, which show an interest in animal forms anticipating the relief carving of a later phase of Mesopotamian civilization. Sometimes known as Middle Assyrian, this later period corresponds to the occupation of southern Mesopotamia by the Kassites and to the Mitannian kingdom on the north Syrian frontier, neither of which have contributed greatly to the total development of ancient Near Eastern art.

Assyrian palaces

The fuller manifestation of Assyrian art and architecture is not seen until the 9th century BC, when Ashurnasirpal II transferred his capital from Ashur to Nimrūd (Calah). The rise of Assyria to imperial power during this and the centuries that followed gave increased vitality to

By courtesy of the Oriental Institute, the University of Chicago

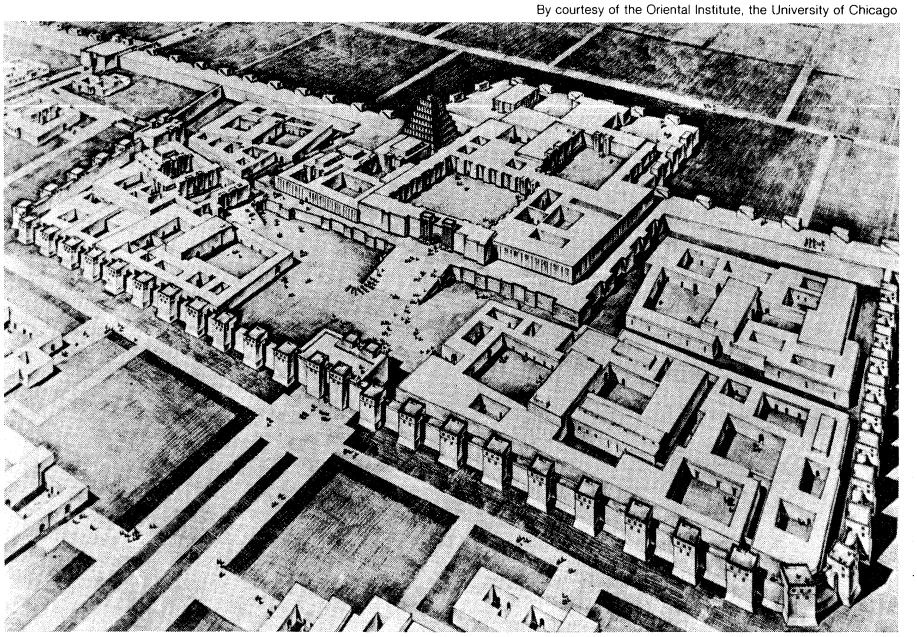


Figure 9: A reconstruction drawing of the citadel of Khorsabad, Iraq, as it may have appeared in the time of Sargon II (721–705 BC). Drawing by Charles Altman.

Mesopotamian architecture. The vast palaces brought to light in the 19th century emphasize the new interest in secular building and reflect the ostentatious grandeur of the Assyrian kings. Like the temples of earlier days, they are usually artificially raised up on a platform level with the tops of the city walls, astride which they often stand. Their gates are flanked by colossal portal sculptures in stone, and their internal chambers are decorated with pictorial reliefs carved on upright stone slabs, or orthostats. In addition to the 9th-century structure at Nimrūd, palace platforms have been exposed at Khorsabad (Dur Sharrukin), where Sargon II established a short-lived capital of his own in the late 8th century BC, and at Nineveh, which, after Sargon's rule, was rebuilt in the 7th century, first by his son Sennacherib and then by his grandson. On the platforms both at Nineveh and Nimrūd, palaces and temples were multiplied by successive kings.

The platform at Khorsabad is occupied by a single royal residence associated with a group of three modest temples and a small ziggurat. Similar buildings occupy a walled citadel at the foot of the platform, thus completing a complex that has been thoroughly excavated and provides the most informative example of typical contemporary architecture (Figure 9). Sargon's palace itself, like that of Zimrilim 1,000 years earlier (see above *Sumerian revival*), is planned, first, around a gigantic open courtyard accessible to the public and, second, around an inner court of honour. From the latter the great throne room is entered through triple doorways, around which, in common with the main outer entrance to the palace, are concentrated a fine array of portal sculptures. The throne room has an adjoining stairway leading to a flat roof and a suite of living apartments behind. Other staterooms, conventionally planned, open onto an open terrace facing the mountains beyond. All of the principal internal chambers are decorated with reliefs, except for the throne room itself, where mural painting seems to have been preferred. The individual purpose and function of the innumerable administrative and domestic offices must remain largely conjectural.

Sculpture. Any history of late Assyrian art must be concerned primarily with relief carving. Some statues in the round have been found, but the comparative ineptitude of the majority of them suggests that this form of expression did not come naturally to Assyrian sculptors. Portal sculptures, which many would consider the most characteristic Assyrian art form, are not statues in the round but "double-aspect" reliefs (*i.e.*, meant to be seen either from the front or the side), apparently derived from a Hittite invention of the 14th century BC. These impressive guardian figures—usually human-headed bulls or lions—decorate the arched gateways and are sometimes supplemented by others set at right angles on the adjoining facades, their heads facing sideways. Each is composed from a single block of stone weighing up to 30 tons, roughly shaped in the quarry and then carved *in situ*.

Less spectacular orthostat reliefs form a continuous frieze of ornament around the bases of interior wall-faces. There is evidence that they were placed in position before the walls that they decorate had been completed. Their carving *in situ* could thus be executed in full daylight. This form of architectural ornament dates from the first quarter of the 9th century BC and seems to have been a genuine Assyrian innovation. The earliest slabs, from the 9th-century palaces of Ashurnasirpal II and his son Shalmaneser III at Nimrūd, are about seven feet high, with the design arranged in two superimposed registers separated by a band of cuneiform inscription. In those from later buildings, such as Sargon II's palace at Khorsabad, the individual sculptured figures reach a height of nine feet.

Relief subjects and themes

The subjects of the designs on these reliefs are very rarely related in any way to religion. Superstitious symbols do occasionally appear in the form of benevolent winged beings, or genii, but the primary purpose of the picture is the glorification of the king himself, either by scenes of ceremonial homage or by extended pictorial narratives of his achievements. The most popular theme, giving rise to numerous variations, involves detailed scenes of military conquest and the ruthless suppression of revolt. These are often arranged episodically to represent successive events in the progress of a single campaign: the Assyrian army prepares for war; led by the king, it crosses difficult country on the way to attack a walled city; the city is taken, burnt, and demolished; the enemy leaders are punished with conspicuous brutality; and, finally, the victory is celebrated. Scenes such as these are distinguished above all by their stylistic vitality and fanciful detail. Animals as well as men are carefully observed and beautifully drawn. The principles of perspective, as later defined by the Greeks, are unknown, but attention is given to the relationship of figures in space and to devices for suggesting comparative distance. At Khorsabad, late in the 8th century BC; some notable stylistic changes are perceptible. The lively carving of narrative and historical subjects has been replaced by more tedious symbols of pomp and ceremony. In keeping with the winged bulls and genii of the portal sculptures, stiffly arranged files of courtiers, officials, and servants stand immobilized in the routine of ceremonial homage. The monotony of the figures is occasionally relieved by the sparing use of coloured pigment on the stone. In the 7th-century Ninevite palaces of Sennacherib and Ashurbanipal, the reliefs suggest a reaction in favour of narrative and violent activity. The slabs are covered to their full height by complicated battle scenes in which the progress of the fighting is suggested by episodic repetition. Types of landscape are depicted schematically, and significant episodes or individuals are identified by a short inscription, without impairing the overall rhythm of the design. In the intervals between their military campaigns, Assyrian kings appear to have been much preoccupied with hunting, and scenes from the chase provided an alternative subject for the reliefs. Lions hunted with spears from a light chariot and herds of wild asses (onager) or gazelles are subjects that stimulated the imagination and sensibility of the Assyrian artist (Figure 10).

Hirmer Fotoarchiv, Munchen

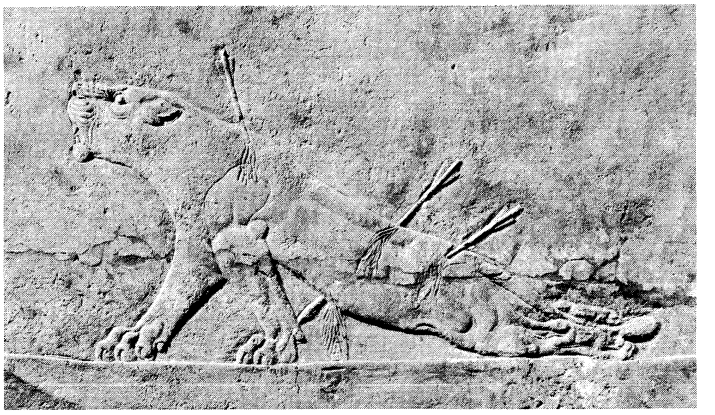


Figure 10: "Dying Lioness," detail of an alabaster mural relief from the North Palace of Ashurbanipal, Nineveh (now in Iraq), Assyrian period, *c.* 650 BC. In the British Museum.

A contrast to these descriptive carvings is provided by the formal monumentality of the Assyrian rock reliefs, secular or religious devices carved on vertical rock faces in localities such as Bavian and Maltai to commemorate historical events that took place there.

The Assyrian talent for relief ornament was not confined to sculpture in stone. First seen during the reign of Shalmaneser III (858–824 BC) are striking examples of relief modelling in bronze. The huge wooden gates of a minor palace at Imgur-Enlil (Balawat), near Nimrūd, were decorated with horizontal bands of metal, 11 inches high, each modelled by a repoussé process (relief hammered out from behind), with a double register of narrative scenes. Their subjects are much the same as the stone reliefs, but even greater ingenuity has been used in adapting the designs to so confined a space.

Relief modelling in bronze

Painting and decorative arts. When greater economy of labour and material was necessary, mural paintings were substituted for slab reliefs. At the time of Tiglath-pileser III (744–727 BC), a country palace at Til Barsip (modern

Tall al-Ahmar) was decorated in this way, with the conventional motifs of relief designs rather clumsily adapted to this very different medium. A few years later, such painting was extensively used to decorate both wall faces and ceilings in Sargon II's palace buildings at Khorsabad. One magnificent panel of formalized ornament has been reconstructed. It is painted in primary colours on a white ground.

There is evidence that the Assyrian palaces were well equipped with furniture. The wooden components have perished, but the ivory ornament with which it was enriched has survived in great quantities. Of these "Assyrian ivories"—relief panels, inlays, and other forms of ornament—only a small proportion can be attributed to indigenous workmanship. The remainder represent either loot from the cities of Syria and Phoenicia or the work of craftsmen imported from those regions. The carving is often technically superb, and the enrichment of the ivory with gold, semiprecious stones, or coloured paste by cloisonné or champlevé processes (whereby the applied decoration is outlined by raised metal strips or fills depressed areas of the surface that have been cut out to receive it) gives increased elegance. But the designs are for the most part a pastiche of misunderstood Egyptian symbolism and often less attractive than the purely Assyrian devices.

Neo-Babylonian period. During the half century following the fall of Nineveh, in 612 BC, there was a final flowering of Mesopotamian culture in southern Iraq, under the last dynasty of Babylonian kings. During the reigns of Nabopolassar (625–605 BC) and his son Nebuchadrezzar II (604–562 BC), there was widespread building activity. Temples and ziggurats were repaired or rebuilt in almost all the old dynastic cities, while Babylon itself was enormously enlarged and surrounded by a double enceinte, or line of fortification, consisting of towered and moated fortress walls. Inside the city the most grandiose effect was obtained by the disposal of public buildings along a wide Processional Way, leading through the centre of the town to the temple and ziggurat of its patron god, Marduk. Where the street passed through the innercity wall, the facades of the famous Ishtar Gate (Vorderasiatiches Museum, Berlin) and those facing the adjoining street were ornamented in brightly glazed brickwork, with huge figures of bulls, lions, and dragons modelled in relief. This form of decoration—a costly process, since each of the bricks composing the figures had to be separately cast—provided a solution for the problem of embellishing mud-brick facades. It appears again in the court of honour of Nebuchadrezzar's palace, using a more sophisticated design that suggests familiarity with Greek ornament. For the rest, there are few innovations in the planning of either palaces or temples during the Neo-Babylonian period. Also (strangely enough, in view of the prolonged excavations that took place at this site), examples of contemporary art are limited almost exclusively to cylinder seals and terracotta figurines of unpretentious design.

Glazed brickwork

(S.H.F.L.)

ASSYRO-BABYLONIAN LITERATURE

Akkadian, the language of ancient Babylonia and Assyria, is the earliest recorded Semitic language and hence of considerable importance for comparative linguistics. Its writing system was borrowed from Sumerian, the world's oldest written language. In the Assyro-Babylonian culture that flourished on the banks of the Tigris and Euphrates rivers, Akkadian replaced Sumerian as a spoken language in the 3rd millennium BC, though Sumerian continued in written usage almost to the end of the active use of Akkadian in the 1st century AD.

Many of the records of this civilization are of linguistic and historical rather than literary interest. There are numerous word lists of names of gods and of common words, the purpose of which was encyclopaedic: the systematic collection of all possible words and concepts. These lists may also have served pupils as spelling guides. Lists giving a word sign or signs with the Sumerian pronunciation and the Akkadian equivalent arranged in columns served as vocabularies, various types of which remained in use until the end of Babylonian scholarly activity. Of historical as well as linguistic interest are the earliest royal inscriptions, simple dedications of objects or buildings of the type "To god X has A, son of B, king of the city Y, given this vase" or "built this temple." The Assyrians further developed this pattern into what are known as "annals," detailed accounts of a king's military exploits for each year. Although popular and sometimes even mythological stories appear in some of these entries, their importance is chiefly historical.

Assyro-Babylonian literary texts have been found on tablets that are dated as early as *c.* 2500 BC and were about myths, epics about the deeds of early kings, hymns, and wisdom literature (proverbs). Of the greatest literary significance are two epics. *Enuma elish* (the *Epic of Creation*) told how Marduk, the city god of Babylon, slew the monster Tiamat and created the world out of her body, a deed for which the other gods rewarded him by granting him universal rule. The *Epic of Gilgamesh,* probably the most important work in Akkadian, related the story of the Sumerian hero Gilgamesh and his companion Enkidu.

The *Epic of Creation* and the *Epic of Gilgamesh*

The fullest text extant is that on tablets found at Nineveh in the library of Ashurbanipal, king of Assyria (668–627 BC). The epic was a creation of the Babylonians. It incorporated Sumerian stories about Gilgamesh and his friend and introduced a new motif, Gilgamesh's quest for immortality after the death of Enkidu. Of particular interest in the epic is the Babylonian version of the Flood, analogous to Noah's Flood in the Bible, which is recounted to Gilgamesh by a survivor, Utnapishtim. The Gilgamesh story, which was current in Asia Minor in languages other than Akkadian, may possibly have influenced the Greek *Odyssey;* and themes from it may occur also in the folklore of the Pacific.

(Ed.)

MESOPOTAMIA UNDER THE PERSIANS

Cyrus united Babylonia with his country in a personal union, assuming the title of "King of Babylonia, King of the Lands." His son Cambyses was appointed vice king and resided in Sippar. The Persians relied on the support of the priests and the business class in the cities. In a Babylonian inscription, Cyrus relates with pride his peaceful, bloodless conquest of the city of Babylon. At the same time, he speaks of Marduk as the king of gods. His moderation and restraint were rewarded: Babylonia became the richest province of his empire. There is no indication of any national rebellion in Babylonia under Cyrus and Cambyses (529–522). That there must have been an accumulation of discontent became clear at the ascension to the throne of Darius I (522–486), when a usurper seized the throne of Babylonia under the name of Nebuchadrezzar (III), only to lose both the throne and his life after ten weeks. Darius waived any punitory action. He had to take more drastic measures in 521, when a new Nebuchadrezzar incited another rebellion against him. This usurper's reign lasted for two months. Executions and plundering followed; Darius ordered the inner walls of Babylon to be demolished and reformed the organization of the state. But Babylon remained the capital of the new satrapy and also became the administrative headquarters for the satrapies of Assyria and Syria. One result was that the palace had to be enlarged.

Cyrus as king of Babylonia

Babylonia remained a wealthy and prosperous land, in contrast to Assyria, which was still a poor country. At the same time, the administration of the kingdom was more and more in the hands of the Persians, and the tax burdens grew heavier. This produced discontent, centring especially on the large temples in Babylon. Xerxes (486–465) had his residence in Babylon while he was crown prince and knew the country very well. When he assumed his kingship, he immediately curtailed the autonomy of the satrapies. This, in turn, gave rise to many rebellions. In Babylonia there were two short interim governments of Babylonian pretenders during 484–482. Xerxes retaliated by desecrating and partially destroying the holy places of the god Marduk and the Tower of Babel in the city of Babylon. Priests were executed, and the statue of Marduk was melted down. The members of the royal family still resided in the palaces of the city of Babylon, but Aramaic

became more and more the language of the official administration. One source of information for this period are the clay-tablet archives of the commercial house of Murashu and Sons of Nippur for the years of 455–403, which tell much about the important role the Iranians played in the country. The state domains were largely in their hands. They controlled many minor feudal tenants, grouped into social classes according to ancestry and occupation. The business people were predominantly Babylonians and Aramaeans, but there were also Jews. The documents become increasingly sparse after 400. The cultural life of Babylon became concentrated in a few central cities, particularly Babylon and Uruk; Ur and Nippur were also important centres. The work of astronomers continued, as evidenced in records of observations. Nabu-rimanni, living and working around 500, and Kidinnu, 5th or 4th century BC, were also known to the Greeks; both astronomers are famous for their methods of calculating the courses of the Moon and the planets. In the field of literature, religious poetic works as well as texts of omens and Sumero-Akkadian word lists were constantly copied, often with commentaries.

(W.T.v.S.)

BABYLONIA UNDER ALEXANDER, THE SELEUCIDS, AND THE PARTHIANS

The end of Babylonian culture

Babylonia was once more to become the centre of the civilized world. Alexander the Great decided to make Babylon one of the capitals of his empire after he conquered the Persians. He began to build a large harbour. In 331 he ordered the restoration of the Marduk temples, among them also the high temple of Etemenanki; Marduk was henceforth recognized as the equal of the Greek god Zeus. After Alexander's death, Babylon lost its position of importance. In 312 Seleucus I Nicator entered Babylon. But soon he founded a new capital, Seleucia, on the Tigris River. Babylon's temples and scholarly institutions continued to receive the care and consideration of the kings of the Seleucid dynasty as well as of some of the later Parthian kings. Antiochus I Soter (281–261) even ordered a Babylonian inscription to be included in the walls of the Temple of Nabu when he undertook the rebuilding of the city of Borsippa. Babylonian business documents from the time of the Seleucids, written on clay tablets, are still extant. Greek and Aramaean documents did not survive. The city of Uruk was second in importance to Babylon as a cultural centre at this time. Uruk's god Anu—the god of the heavens—was recognized along with Marduk as the equal of the Greek Zeus. Two large temples were built for Anu and Ishtar. Written rituals indicate that the cult had remained true to the old traditions. The citizens of Uruk participated in these rituals in various functions, lucrative as well as honourable and respected. Some documents indicate that religious benefices were handled as stock and often sold in shares. Other documents deal with the selling of promises. The exact measurements and markings of the boundary lines are stated on these. While the names of the parties are mostly Babylonian, it is not unusual to find a second, Greek name added to the original Babylonian one.

The Babylonian astronomers

An important result of the mixture of Babylonian and Greek culture in Babylon and Uruk was the further development of the science of astronomy. The Babylonians contributed their methods of mathematical calculation, as well as an accumulation of observations covering several centuries. The Greeks contributed their training in logical thinking. Many documents of the Hellenistic princes who ruled in Asia Minor and along the Aegean Sea during this time show the cooperative efforts of the Greek and Babylonian astronomers. They laid the basis upon which astronomers of later times were able to build. One of them deserves particular mention: the Marduk priest Berosus, who founded a school for astrology and astronomy on the island of Kos around the year 280. He was also the author of a book about Babylonia, its history, and its religion (*Babylōniaka*). The book was written in Greek, and fragments of it still exist.

The decline of the empire of the Seleucids after 150 led to the incorporation of Babylonia into the empire of the Parthians in eastern Iran between 141 and 129. The result was by no means a great political catastrophe. Nevertheless, it hastened the decline of the late Babylonian city culture. There had been no new creative or cultural ideas for a long time, and the old schools for scribes slowly became extinct. Few literary texts and documents survive from the time between 140 BC and 80 BC. The knowledge of the cuneiform script was lost under the Greeks, who were hardly interested in foreign languages. Not until the 19th and 20th centuries did the discovery of tens of thousands of cuneiform texts lead to the decoding of this entirely forgotten writing. This permitted scholars to repossess a large part of the lost literature of Mesopotamia.

Parthian and Sāsānian periods

Mesopotamia (especially the district of Nisibis) was attached to the growing dominions of Armenia temporarily in the 1st century BC but was recovered for Parthia by Phraates III (reigned 70–58/57 BC). With the rise of Roman power in the Near East, it assumed the character of a border territory between the Roman and Parthian spheres of influence, though it usually remained under Parthian control. The Roman leader Crassus led an army into Mesopotamia in 54 BC but was defeated by the Parthians at Carrhae in 53. In AD 114–117 the emperor Trajan conquered the area and made it a Roman province, but it was abandoned once again to Parthia by Hadrian after Trajan's death. Beginning in 165, the Romans once more occupied the northwest portion of Mesopotamia.

During most of the period from the early 3rd century until the rise of Islām in the early 7th century, Mesopotamia remained a battleground between the Romans (in the 6th and 7th centuries, the Byzantines) and the Sāsānian successors of the Parthians. As during the Parthian period, Roman influence was for the most part confined to the extreme northwestern part of the country—the area around Carrhae, Edessa, and Nisibis, near the Syrian and Anatolian borders. The Sāsānian kings made their capital at Ctesiphon, near the midpoint of the Tigris, and their influence predominated in the rest of Mesopotamia. During this period, Mesopotamia became one of the chief centres of Nestorian Christianity and had a sizable Jewish community. (For the history of the Parthian and Sāsānian empires, see IRAN: *History*.)

(Ed.)

IRAQ FROM THE ARAB CONQUEST TO 1918

From the time of the Arab conquest in the 7th century to the second decade of the 20th century, Iraq existed not as an independent state but as a geographical expression for the flatlands between Baghdad and the Persian Gulf—an area of unavoidable passage between Central Asia and the Mediterranean. The history of Iraq in this period is dominated by the rivalry of the three strongest ethnic elements of the Muslim world: the Arabs, the Iranians, and the Turks. During this period Iraq was also a great centre of Islāmic culture as well as the place of origin of several religious sects that divided the community of Islām. The political history of the area may be summarized as follows: after being the metropolitan province of the Arabo-Islāmic Empire under the caliphs' more or less effective authority, Iraq fell successively under the control of the Būyid dynasty of Iran and the Seljuq Turks; and then, after the Mongol invasion, which hastened its decline, it became a dominion of Mongol and later Turkmen and Persian dynasties established in Iran and remained for almost four centuries under Ottoman rule, to emerge again as an independent state as a result of World War I.

The Arab conquest. From the 3rd century AD, the Lakhmids, a small dynasty of South Arabian origin, ruled al-Ḥīrah, on the Euphrates, as vassals of the Sāsānid rulers of Iran, who had entrusted them with keeping the Bedouins in check and holding the borders of Iran against the Byzantines; the Lakhmid capital was a flourishing centre of trade between Persia and Arabia as well as a focus of Arabic culture. In addition, some elements of such important Arab tribes as the Taghlib, the Bakr, and the Tamīm had taken the steppes of Lower Mesopotamia for pastureland, and relations had been established between

the Prophet and those Bedouins, perhaps with a view to subsequent operations in the area.

The rich area stretching south of al-Ḥīrah was protected by strongholds that were easily reduced on the occasion of mere reconnaissances made by Arab parties under al-Muthannā ibn Ḥārithah, a leader of the Bakr, who is said to have called in the famous general Khālid ibn al-Walīd to help him. Khālid seized Al-Ḥīrah in 633–634, but he was sent to Syria. Al-Muthannā was left in charge of the Arab forces in Lower Iraq. For three years, the progress of the conquest proceeded slowly, until 637, when Sa'd ibn Abī Waqqāṣ won a decisive victory over the Sāsānids at al-Qādisīyah, 20 miles (32 kilometres) southeast of al-Ḥīrah. The Arabs sacked Ctesiphon, the Persian capital, and seem to have been welcomed by the native population of the region. Those successes marked the beginning of the collapse of the Sāsānid Empire and made it possible for the Arabs to spread eastward. In 641 both Nineveh and Mosul fell to the Arabs. Sāsānid power was eventually destroyed in 642 at the battle of Nehāvand. This victory enabled the Arabs to control western and central Persia in addition to Mesopotamia.

Establishment of Basra and Kūfah

At that time two military camps, Basra and Kūfah, had been established, the former on the right bank of the Shatt al-Arab in 638 and the latter on the right bank of the Euphrates, near al-Ḥīrah, in 638–639. Both camps were to develop into large cities and play an outstanding part in the history of Iraq until the foundation of Baghdad in 762. Both towns received new inhabitants either from Arabia or the East as a result of conquests that brought them captives and booty as well as revenue from two Persian provinces (*māh al-Baṣrah* and *māh al-Kūfah*) within their scope. The area to the east of Basra and Kūfah was called as-Sawād (the Black Region), in reference to its fertility. The mixing of various ethnic elements promoted economical and intellectual development in the rival cities, while political events gave rise to the great politico-religious parties that were to divide Islām into discordant groups. Moreover, it was in Basra and Kūfah that all literary, philological, and historical materials that were to form the basis of Arabo-Islāmic culture were collected from Bedouins: it was there that Islāmic sciences such as grammar and lexicography came into existence. Qur'ānic exegesis and science of traditions developed, while several schools of thought, particularly Mu'tazilism (see below), took shape. It is true that Kūfah did not outlive for long the founding of Baghdad, but Basra was far enough from the capital to retain its identity. The whole history of Islām has been to some extent conditioned by events that occurred at Basra and Kūfah or were in connection with either city. 'Uthmān ibn 'Affān, the third orthodox caliph, was murdered in 656. 'Ā'ishah, a Prophet's widow, set out for Basra with her followers with the object of avenging the caliph and fighting 'Alī ibn Abī Tālib, the latter's successor. With the help of troops from Kūfah, 'Alī defeated the rebels in the so-called Battle of the Camel (for 'Ā'ishah was riding a camel) fought in Basra in December of 656. Here two parties of Muslims struggled against each other for the first time. In the following year 'Alī tried to subdue Mu'āwiyah, the powerful governor of Syria who had refused to obey an order of dismissal. Leaving Kūfah, where he had established the seat of his government, he met his rival at Ṣiffīn, on the right bank of the Euphrates. When 'Alī was about to win the day, Mū'āwiyah's troops raised leaves of the Qur'ān on their spears, meaning that they wished to leave it to God's decision. 'Alī agreed to stop fighting and both parties agreed to arbitration. On his return to Kūfah, 'Alī had to suppress a rebellion instigated by some soldiers who contested the very principle of arbitration, and he crushed these Khārijites ("Seceders") on July 17, 658, in an-Nahrawān. The survivors, however, continued the movement, scattering in various areas and recruiting new followers; thus, it was in Iraq that Khārijism, the earliest sect to rise within Islām, came into existence.

The Umayyads. After being deposed in January 659 by the arbitrators met in Adhruḥ, 'Alī tried to preserve the caliphate, but he was murdered by a Khārijite on January 24, 661, and Mu'āwiyah, who had asserted his dominion over Syria and Egypt, was proclaimed in the Arabian province of Hejaz. Mu'āwiyah moved the seat of the Umayyad caliphate to Damascus in Syria. Iraq became a single province but a privileged one, as it developed into the military base for the Muslim conquest of Eastern countries. Iraq continued to play a major political role in the empire of the caliphate.

Movement of the caliphate to Damascus

In Umayyad times and afterward, the limits of Iraq were roughly marked out by a line running some distance to the west of the Shatt al-Arab and the Euphrates up to al-Anbār (sometimes up to al-Ḥādīthah), thence to Takrīt on the Tigris, then on through Ḥulwān to the Persian Gulf east of 'Abbādān (now Abadan, in Iran); al-Jazīrah and Mosul were not included in Iraqi provinces except during certain periods.

Under the Umayyads, Kūfah inclined toward 'Alī and his party; it was there that revolts raised by his descendants or their followers, the Shī'ah, broke out. Basra seemed to be in favour of a more democratic regime, but it was anxious to preserve independence and scarcely accepted governors sent from Damascus. It was at Basra that the Khārijites, the most vehement opponents to the established order, caused disturbances. After the murder of 'Alī and the failure of a plot woven simultaneously by them against Mu'āwiyah and the Governor of Egypt, they constantly threatened Iraq and especially Basra during the following decades. While Iraqi troops were launched to the conquest of Oriental areas as far away as India, Basra had to defend itself against the rebels. Mu'āwiyah publicly recognized Ziyād (an alleged illegitimate son of his father) as his brother and appointed him governor of Basra in 665 and then of Basra and Kūfah from 670 to 675. Ziyād fought both Khārijites and Shī'īs with efficiency and, by promoting important drainage projects, did much to restore the economic life of Lower Iraq. His son, 'Ubayd Allāh, continued to rule firmly in Iraq, but, when Mu'āwiyah died in 680, the succession of his son engendered further revolts. Al-Ḥusayn, 'Alī's second son, was invited by his followers to Kūfah, where a revolt had been initiated in his support; on October 10, 680, al-Ḥusayn was intercepted at Karbalā', near Kūfah, and slaughtered, together with his small escort. This event had an enormous impact on the Muslim world: it inspired unspeakable abhorrence among the Shī'ah and other orthodox Muslims toward the Umayyads. A few years after Karbalā', in January 685, the so-called Penitents (Tawwābūn) tried, with little success, to foster disturbances in Iraq with a view to avenging the murder of al-Ḥusayn; next, a rebellion broke out under al-Mukhtar, who backed the cause of Muḥammad, another son of 'Alī. Al-Mukhtār succeeded in winning the support of some 'Alids (supporters of 'Alī) and *mawālī* (non-Arabs who had accepted Islām) and gained control over Kūfah for a few months. At that time, the anti-caliph 'Abd Allāh ibn az-Zubayr, who had revolted in Medina and established himself in Mecca, had appointed Muṣ'ab suppressed the uprising, and al-Mukhtār was killed in 687. Four years later, the Umayyads recovered their control over Iraq.

Despite this internecine warfare, Iraq prospered. The province profited by remote conquest, carried out by its armies, and witnessed an important development. In 694, the caliph 'Abd al-Malik ibn Marwān appointed an outstanding statesman, al-Ḥajjāj, as governor of Kūfah. With the support of Syrian forces, he established respect for the caliphal authority and restored order throughout the country. Eager to achieve some sort of a balance between the two cities of Basra and Kūfah, he founded a new city, which he called Wāsiṭ (intermediary), and made it the seat of his governorship. At the same time, important public utility works were undertaken: canals were dug, swamps drained, and agriculture revitalized. Arabic replaced Persian as the official language of government administration. New coins with Arabic inscriptions were minted at Wāsiṭ to replace Sāsānid money.

The founding of Wāsiṭ

A serious problem arose concerning the tax system. According to a general rule, those lands that had been left to their non-Muslim owners were subjected to the *kharāj* (land tax), while Muslim landowners paid only the tithe; thus, a large number of peasants embraced Islām in order

to escape paying the *kharāj*. But some tax collectors continued to levy the land tax; this caused a mass movement of the peasantry into the towns. Al-Ḥajjāj forcibly resettled them on the land, and their embittered discontent erupted into several revolts against the Umayyads. Under the caliph Hishām ibn ʿAbd al-Malik (724–743) it was decided that land tax should be imposed independently of the religion of *kharāj* landowners.

After al-Ḥajjāj's death (714), Yazīd ibn al-Muhallab, one of his successors, who had been put in jail and escaped, fomented a rebellion that was crushed in 721; order was temporarily restored by Khālid ib ʿAbd Allāh al-Qasrī, but increasing disturbances stirred up by Khārijites and ʿAbbāsid followers as well brought the Umayyad rule to an end. In Khorāsān a campaign aiming at overthrowing the Umayyads in favour of the ʿAbbāsids, who inclined more toward the Persian element, was successful; the example of Khorāsān was followed in other areas; Kūfah was the first town to rise, and it was in its great mosque that as-Saffāḥ, the first ʿAbbāsid caliph, was proclaimed on October 30, 749. Thenceforth, the history of Iraq intermingles with the ʿAbbāsids'.

The ʿAbbasids. The reign of Abū alʿAbbās as-Saffāḥ (749–754) was taken up with fighting the last Umayyads and asserting ʿAbbāsid power. His successor, al-Manṣūr (754–775), ruthlessly harried the Shīʿīs, whose revolts he suppressed in 755 and 762–763; in 755 he put to death Abū Muslim, to whom the ʿAbbāsids were principally indebted for their accession to power. His fame rests on his founding of the new capital, Baghdad, completed in 766, and in laying the foundations on which his successors were to build an elaborate administration. Al-Mahdī (775–785) distinguished himself in fighting the so-called *zanādiqah—i.e.,* heterdox sects and individuals, especially Manichaeans. He was succeeded by al-Hādī (785–786) and Hārūn ar-Rashīd (786–809). The latter's image has been somewhat idealized by the *Thousand and One Nights.*

The founding of Baghdad

The Umayyads had been an Arab dynasty; the ʿAbbāsids, however, who often had Iranian blood on their mothers' side, naturally turned east and organized their administration and court on the same lines as the Sāsānids'. The first ʿAbbāsid period was the golden age of Persian officials. Although deeply Arabicized, they did not fail to remember the old glory of their fatherland and to assume hostile attitudes toward the Arabs, which produced a racial movement known as Shuʿūbīyah and various conflicts between Arabs and non-Arabs. On the other hand, it is obvious that converts to Islām were not always absolutely sincere, and many of them secretly retained their own faiths. That is the reason al-Mahdī hunted out the *zanā-diqah* and put a great many of them to death. Al-Mahdī acted in self-defense; it is not preposterous to discern a feeling of the same kind in ar-Rashīd's reaction against the Barmakids.

As-Saffāḥ had taken as his chief associate Khālid ibn Barmak, a noble Persian who remained in the service of the caliphs. Yaḥyā, Khālid's son, became the vizier of ar-Rashīd, while al-Faḍl and Jaʿfar, who were Yaḥyā's sons, were both entrusted with high government offices. Suddenly, ar-Rashīd decided to execute Jaʿfar (January 803). He imprisoned Jaʿfar's father and brothers and seized the property of the whole family. The reason for ar-Rashīd's actions is not quite made clear by the sources, which speak of an affair between Jaʿfar and ʿAbbāsah, ar-Rashīd's sister; whether this situation was so or not, such an act of cruelty was regarded as an unbearable slur cast on Persians, at a time when the Shīʿīs, after some unsuccessful attempts at seizing power, fostered deep resentment against the ʿAbbāsids.

On ar-Rashīd's death, his succession gave rise to bloody conflicts between his sons al-Amīn (809–813) and al-Maʾmūn (813–833); the latter defeated his brother but was compelled to besiege Baghdad to recover the throne. Aiming at reconciling Shīʿīs and Sunnīs, al-Maʾmūn contemplated the designation of ʿAlī ar-Riḍā, the eighth Shīʿī *imām,* or leader of the community of Islām, as heir apparent; he was forced, however, to go back on his decision in view of the strong Sunnī opposition in Baghdad, a fact that was again jeopardizing the balance between Arabs and Persians, the latter supporting the Shīʿīs; such policy was all the more a failure as Shīʿīs suspected the caliph of having poisoned ar-Riḍā (died 818). In 815, Abū as-Sarāyā, allegedly backing the ʿAlids Ibn Ṭabāṭabā and Muḥammad ibn Zayd, occupied Kūfah, where he minted money, seized Wāsiṭ, Basra, and other towns; and set out for Baghdad. He was eventually defeated and was beheaded on October 18, 815.

Later, however, the political power of the caliphate asserted itself. Economic prosperity grew, and the Arabo-Islāmic civilization centred in Baghdad radiated far beyond the limits of the empire. The seventh ʿAbbāsid caliph, al-Maʾmūn, founded, or at least developed, in Baghdad the Bayt al-Ḥikmah (The House of Wisdom), where Greek philosophical and scientific works were translated into Arabic. A rich library was attached to the House of Wisdom, as were several astronomical observatories. A harmonious blending of cultural factors, borrowed from Greece, Persia, and India, enabled Arabo-Islāmic culture to rise to a high degree of refinement in such fields as medicine, mathematics, philosophy, theology, literature, and poetry. Iraq came to regard itself as a kind of privileged heir to all former civilizations and felt bound to accept, sort out, and develop their legacies. Aristotle was considered the master by an intellectual elite fascinated by his method of logical reasoning. The influence of Aristotle was reflected in Muʿtazilism, a school of thought that flourished in Basra before reaching its peak in Baghdad. Without breaking with orthodoxy, the Muʿtazilīs aimed at harmonizing reason and faith; therefore, they assigned free will to man and regarded the Qurʾān as created by God, unlike strict orthodox Sunnīs, who hold it uncreated, coeternal with God. Anticipating that he could rely on the support of the intellectual elite against his opponents, al-Maʾmūn embraced the Muʿtazilī doctrine, but he came up against difficulties when he wanted to enforce the createdness of the Qurʾān as a formal tenet. Al-Maʾmūn's religious policy alienated both the Shīʿīs and the Sunnīs, and in order to maintain his precarious position he inaugurated a severe inquisition (*miḥnah*).

The House of Wisdom

In ordering the *miḥnah* al-Maʾmūn had proved that the ʿAbbāsid caliph could become an *imām,* a spiritual ruler, although without any right of legislation on purely religious matters such as worship and canon law; in addition, he was a monarch whose powers were fixed by the Sharīʿāh, the Islāmic law. Though these powers were held in the person of the caliph, he frequently delegated them to a vizier placed at the head of the hierarchical central administration, which included the army, the chancery, the ministry of finance, the postal system, and the intelligence service. A great *qāḍī,* with his seat at Baghdad, headed the administration of justice. Provincial towns often were governed by ʿAbbāsid princes, who were closely connected with the central government, the orders of which they carried out.

Powers of the caliph

From the economic point of view, the founding of Baghdad shifted international trade toward Iraq and developed Basra as an emporium and transit centre. The capital was the hub of Middle Eastern trade. From Baghdad, merchants sailed to India, Ceylon, and even China, while overland trade routes converged on Baghdad from Asia Minor, Syria, Egypt, Arabia, and Persia.

It was not long, however, before Iraq shared in the political and economic decline of the caliphate. Religious conflicts had to some extent weakened the authority of the caliph and divided the Muslim community. Yet a powerful administration for some time kept up a semblance of cohesion, in spite of centrifugal forces tending to detach several outlying provinces from Baghdad.

When al-Muʿtaṣim (833–842) succeeded to power, he enrolled a number of foreign mercenaries, mainly Turkish slaves, in his bodyguard. The officers of the new Turkish guard rapidly distinguished themselves in war and rose to influential positions. Since the excesses of these troops created a dangerous state of tension in Baghdad, al-Muʿtasim in 836 moved his residence to a new city, Sāmarrāʾ, about 60 miles (97 kilometres) north of Baghdad. Henceforth, Turkish mercenaries dominated the

empire and controlled the succession of caliphs. Under al-Wāthiq (842–847), the commander of the Turkish guard was even given the title of sultan. Al-Mutawakkil (847–861) succeeded to the throne with the help of two Turkish chiefs. In a desperate attempt to restore the caliph's authority, al-Mutawakkil endeavoured to conciliate the orthodox Muslims (Sunnīs). He persecuted not only the Muʿtazilites but also the Shīʿīs and imposed severe restrictions on Jews and Christians, compelling them to wear distinctive dress.

ʿAbbāsid decline

These measures did not prevent the decline of the dynasty. Between 861 and 870 four caliphs appeared in succession; they were to suppress a revolt of the Turks leading to the siege of Baghdad in 865–866 and, more especially, a serious uprising of Negro slaves (the Zanj) settled in the swamps of Lower Mesopotamia; the operations launched against them during 15 years (869–883) wore out the resources of the government. Basra was sacked (871), Lower Iraq occupied, Wāsiṭ taken (878), and the country devastated. This social revolt threatened the caliphate, which at the same time was faced with other rebellions in Khorāsān, Egypt, Ṭabaristān, and in other areas.

About 890 Lower Iraq was again shaken by revolt, this time by Ḥamdān Qarmat, whose followers won tremendous victories and gained control over parts of Iraq, Syria, and Palestine. Al-Muʿtamid (870–892) reestablished the seat of the caliphate in Baghdad (892) shortly before he died, and his successors did not return to Sāmarrāʾ. Al-Muktafī (902–908) stopped the Qarmaṭian movement temporarily, but the rebels resumed their activity and plundered Kūfah in 925.

In addition to such factors of internal decay, centrifugal forces compelled the ʿAbbāsids to grant independence to some provinces, even on the borders of Iraq. As early as 890, the Ḥamdānids, who were to come into prominence in Aleppo during the 10th century, created an independent amirate in Mosul, and the Caliph himself was sometimes induced to ask their help. Less than two centuries after the beginning of ʿAbbāsid rule, Iraq became practically the sole possession of the caliphate. Struggles among the Turkish commanders and between the Turks and the Daylamite and other elements of the army reduced even Baghdad and the neighbourhood to anarchy.

The Buyids. In 936 the caliph ar-Rāḍī (934–940) introduced the title of *amīr al-umarāʾ*, (amir of the amirs) and bestowed it on Muḥammad ibn Rāʾiq, governor of Basra, retaining for himself only the religious leadership of the Muslim community. The new title brought about serious rivalries between several pretenders to it, until the Būyids, a Shīʿī family from Daylam, a mountainous region in Iran, made themselves masters of Basra, Wāsiṭ, and Baghdad, ushering in a new stage and giving the caliphate, which had retained part of its previous prestige, a still unknown character. Although it reached its lowest ebb during the period of Būyid supremacy, the caliphate was not entirely eliminated by the Būyids, who probably hoped to use such glamour as the ʿAbbasids still possessed. Moreover, the survival of a sunnī caliphate left them with a free hand for their own political and religious aims. They were content with the rank of *amīr al-umarāʾ* and with honorary titles ending with ad-Dawlah ("of the dynasty," *e.g.*, "friend of the dynasty," or "support of the dynasty"), but it is obvious that they held all powers in their hands.

The reign of the first of them, Muʿizz ad-Dawlah (945–967), was largely occupied by internal rebellions and a series of expeditions against the Ḥamdānids; confiscations of the property of officials on death or dismissal were common, and offices were put up to the highest bidder. The Amīr al-Umarāʾ did nothing to improve conditions and had no care for the local population; he also made grants of land to his troops, the result of which was to bring agriculture into a hopeless state of disorganization. Under his successor ʿIzz ad-Dawlah (967–977), Turkish mercenaries revolted and seized power. In 977 ʿAḍud ad-Dawlah became the chief of the family; he ruled first in Fars, and then in Fars and Iraq (977–983). A capable ruler, ʿAḍud ad-Dawlah kept up some prosperity, improved roads, dug canals, reorganized the postal system, stamped out brigandage, and fostered trade. After his death the Būyid dynasty rapidly declined; the western provinces of their dominions were torn by internecine strife, the administrative system collapsed, and agriculture fell into decay.

Rise of ʿAḍud ad-Dawlah

In the field of finance, reforms initiated by the Būyids were connected with the pay of troops: the caliphs had tended to grant to the military an increasing amount of lands subject to the tithe. When the treasury proved unable to cover the cost of the army, the Būyids introduced the practice of granting *iqṭāʿ*s, or lands hitherto subject to land tax (*kharāj*). The beneficiaries of *iqṭāʿ*s levied the revenue of the *kharāj* themselves and paid nothing to the treasury. The debilitating consequences of such a system contributed to the political disintegration of the empire.

As for religion, the Būyids did not fail to spread their Shīʿī doctrines and practices, erecting in particular mausoleums on the tombs of the *imām*s. In spite of their Daylamite origin, the Būyids were eager supporters of Arabic culture, although they also favoured the development of new Iranian influences. It was at that time that rhymed prose gained ground in the Chancery and literature, the *maqāmah* appeared, and a new style was initiated in epistolography, yet most of that activity took place beyong Iraq. Baghdad, where Arabic culture had reached its height—with prose writers such as al-Jāḥiẓ Ibn Qutaybah, al-Masʿūdī, and many others; historians such as al-Yaʿqūbī, aṭ-Ṭabarī, or ad-Dīnawarī; poets such as Abū Nuwās, Abū Tammān, al-Buḥturī, and others; philosophers such as al-Fārābī; and scientists such as al-Khwārizmī—was not quite superseded, but it was challenged by other centres, such as Shīrāz and Isfahan, let alone Cairo, where the Fāṭimids established themselves in 969, attracting a large part of the flow of trade that previously had converged on Iraq.

Daylamite culture

In fact, a large district of Iraq escaped Būyid control; just as Mosul was in the hands of the ʿUqaylids, so, as early as 1012, the Mazyadids were recognized as *amīrs* over an area extending from Basra to Hīt; they threatened Baghdad itself, where the *ʿayyārūn* were terrorizing the inhabitants, being actually the masters of the capital. The activities of the *ʿayyārūn* lasted until 1055—when the Seljuq ruler, Toghrïl Beg, entered Baghdad—and perhaps a little longer.

The Seljuqs. A former high official of the Būyids, al-Basāsīrī, who had risen to power in Baghdad, fled on Toghrïl's coming, but it was not long before he marched on the city, occupied it on December 27, 1058, and proclaimed the Fāṭimid al-Mustanṣir caliph; the ʿAbbāsid al-Qāʾim took refuge with the ʿUqaylids, and al-Basāsīrī extended his control over Basra and Wāsiṭ, with Mazyadid support. In 1059 Toghrïl reconquered Baghdad, and the caliph returned.

While the power of the Būyids had been usurped, that of the Seljuq Turks was given a legal basis. Unlike the Būyids, the Seljuqs were Sunnī, hence the caliph recognized them as sultans and delegated all his powers to them, entrusting them with eradicating Shīʿīsm from Iraq and other parts of the empire. With their eagerness to restore and assert Sunnism, they must be credited with the founding of *madrasah*s—high colleges intended for training future officials in both religious and secular matters. The most famous among those *madrasah*s is the Niẓāmīyah, founded at Baghdad in 1067. Nevertheless, the Seljuqs did not take Baghdad as their capital, since they lived at Isfahan and later (from 1157 onward) at Hamadan. During the Seljuq period, the history of Iraq is that of a succession of Turkish sultans who regarded the ancient centre of the Islāmic Empire as a kind of protectorate under religious leadership of the caliph, who was not more than a local ruler, certainly less powerful, in the political and military field, than the Zangids in Mosul or the Mazyadids in Lower Iraq. The Arab dynasty of the Mazyadids came to prominence in al-Jāmiʿān (called al-Ḥillah from 1101 onward); they were Shīʿī, as were the majority of the Arab tribes in the region, and ready to support ʿAlid movements against the Seljuqs or play off some of them against the others. Profiting by the struggles between Berk-yaruq and his brother Muḥammad, Ṣadaqah ibn Manṣūr extended his power over almost all Iraq by the end of the 11th century; the

towns of Hīt, Wāṣit, Basra, and Takrīt fell into his hands, but he was killed in a battle against the Seljuqs in 1108, and the dynasty of the Mazyadids eventually came to an end in 1163.

In 1118 there came to power a branch of Seljuqs who established an autonomous principality in Iraq that lasted until 1194. During this period the caliphs made some attempts to dispute the supremacy of the sultans and shake off their tutelage, but these actions resulted in struggles, notably in the reign of Sultan Masʿūd (1134–52). Al-Mustarshid (1118–35) was the first caliph in Seljuq times to assemble an army and lead it in person. But al-Mustarshid was executed in 1135, as was his successor, ar-Rashīd, in 1136. Subsequently, on the death of Masʿūd, al-Muqtafī (1136–60) established himself as the legitimate sovereign in Iraq, exercising both temporal and religious power. Later, the caliph an-Nāṣir (1180–1225) tried to restore some importance to the caliphate by conciliating Sunnīs and Shīʿīs and using associations of various kinds (grouped under the general name of *futūwah*) as instruments of moral and social education, hoping to restore the moral and political unity of the ʿAbbāsid Empire. Furthermore, about 1192, he appealed to the *khwārezm-shāh* Tekesh for help against Toghrïl III, the last of the Seljuq sultans of Iraq, who was defeated in 1194 and his dynasty overthrown. An-Nāṣir contended with Tekesh for Seljuqs' spoils and succeeded in taking Khuzistan and other Persian provinces. But Tekesh proved a more dangerous rival than Toghrïl; shortly before his death (1200), he demanded that the *khuṭbah* (the Friday sermon in the great mosques) should be read in Baghdad in the name of the Khwārezm-Shāh. An-Nāṣir refused, and in 1196 fighting took place between the Khwārezm-Shāh's army and the Caliph's, to the disadvantage of the latter. Some years later, Muḥammad ibn Tekesh reiterated his father's demand and met with an uncompromising refusal. The conflict culminated in 1216, when the Caliph had a former official of Tekesh assassinated. In the following year Muḥammad ibn Tekesh set out for Baghdad but was unable to take the city, his army being annihilated by the Kurds. An-Nāṣir played off the Mongol leader Genghis Khan against the Khwārezm-Shāh, and the latter was defeated in 1219.

Reform organizations

The Mongols. Unfortunately, an-Nāṣir's successors, aẓ-Ẓāhir (1225–26), al-Mustanṣir (1226–42), and al-Mustaʿṣim (1242–58), were weak and unable to carry on his policy. In the meantime, Hülegü, a descendant of Genghis Khan, who established Mongol power in Iran, invited the caliph al-Mustaʿṣim to surrender and, on his refusal, marched on Baghdad, which he took on February 10, 1258, after a siege of some seven weeks. Al-Mustaʿṣim and several members of his family were put to death; the city was sacked and its inhabitants slaughtered. Iraq, the core of the ʿAbbāsid Empire, was devastated and depopulated. It did not recover for centuries.

Until 1339 the province of Iraq remained under the domination of the Īl-Khāns, a Mongol dynasty founded in Persia by Hülegü. Under the Īl-Khāns, the western frontier of Mesopotamia was the political boundary of Persia, Baghdad being the capital of Lower Mesopotamia and Mosul the seat of another government. In 1295, Baydu, a grandson of Hülegü and governor of Iraq, seized power, but he was overthrown in the same year by Ghāzān, governor of Khorāsān. When Abū Saʿīd, the 9th Īl-Khān, died in 1335, Ḥasan Buzurg (died 1356) founded in Baghdad the dynasty of the Jalāyirids, who achieved some success, chiefly in Fars and Upper Mesopotamia, but they were compelled to suppress a succession of revolts in Baghdad. Restless Bedouin tribes in the south and Turkmens and Kurds in the north complicated the state of affairs and hastened the decline of a moribund country. This small dynasty of Mongol origin ruled from 1339 to 1410 and incorporated Mosul into its dominions in 1364. On two occasions, however, in 1393 and particularly in 1401, Timur (Tamerlane) took and sacked Baghdad, slaughtered its inhabitants, and pulled down a great part of the town. Aḥmad the Jalāyirid (died 1410), driven out by Timur, recovered Baghdad in 1394 and again in 1405 and tried to make good the damages.

The Jalāyirids

Until 1432, Jalāyirid rulers retained some power in Lower Mesopotamia (al-Ḥillah, Basra, Khuzistan) but were unable to preserve Baghdad, which fell (1411) into the hands of the Kara Koyūnlu (the Black Sheep), a Turkmen dynasty that ruled Asia Minor from 1375 to 1468. This period is also marked by political fragmentation and rivalries between Baghdad and Mosul. The rule of the Turkmens was challenged by the politico-religious movement of the Mushaʿshaʿ, who came into prominence in Khuzistan, attempted to extend their control over Lower Iraq, and harried the Kara Koyunlu as far as the outskirts of Baghdad, which they were not able to take. In 1468, the Ag Koyunlu (the White Sheep) took over from the Kara Koyunlu, ruling from Tabriz; in 1503, the last member of the dynasty was compelled to seek refuge at Baghdad, where he ruled till 1508, when Iraq, including Mosul, fell under the sway of the Shīʿī Ṣafavids and remained in their possession until the Ottomans conquered it. In spite of the weakness of the local government, Iraq enjoyed a fairly prosperous economy during this short period.

The Ottomans. Although Sultan Selim I (1512–20) had crushed the Ṣafavid armies at Chāldirān in 1514 and taken Tabriz, Mosul, Diyarbakir, and several Kurdish areas east of the Tigris in 1515–17, Iraq proper remained under Persian control for some time. In Baghdad, the Kurdish leader Dhū al-Faqār raised a revolt in 1527 that induced Shāh Ṭahmāsp (1524–76) to assert Persian authority over central Iraq more firmly than ever, but hostilities started again between him and Süleyman (Sulayman the Magnificent, 1520–66); the latter entered Baghdad in 1534, and Iraq remained in the possession of the Ottomans until March 1917, though not without trouble: the peace was somewhat disturbed by intermittent struggles between Turks and Iranians, resulting in the occupation of Baghdad by the Ṣafavids from 1623 to 1638; the Turkish communications were at the mercy of turbulent tribesmen; and parts of the country were semi-independent at one time or another. The long period of Ottoman rule, however, witnessed some revival of Baghdad, improvement of living conditions, and administrative reorganization on Western lines as well as the beginning of European penetration.

Süleyman the Magnificent

The Ottomans had assumed the title of caliph, and the capture of the last ʿAbbāsid puppet in Cairo (1517) had emphasized their feeling that they were the successors of the great Baghdadian dynasty; therefore, the conquest of Baghdad increased Süleyman's prestige, but it had another consequence: on the one hand, it temporarily brought to an end the rivalries between Shīʿīs and Sunnīs and materialized the triumph of Sunnism in the country where such religious conflicts were frequent; on the other hand, the conquest of Egypt had made the Ottomans the masters of the Red Sea, which retained, in spite of the discovery of the Cape route around Africa at the end of the 15th century, much of its former importance in international trade. As masters of the Persian Gulf, the Ottomans could control all the traffic in the Middle East. Last but not least, Iraqi territories were a kind of buffer against not only the rulers of Iran but also the restless tribes of Kurds and Arabs.

Basra, always eager to preserve independence, played a certain part under the Ottoman rule. At first a vassal of the Sultan, it tried to shake his suzerainty, but two expeditions sent from Baghdad in 1546 resulted in its incorporation into the Ottoman possessions in Iraq, which was divided at the time into five *eyalets*: Baghdad, Basra, Mosul, Shahrazūr, and al-Ḥasā, on the western shore of the Persian Gulf.

For almost two centuries the Ottomans tried to exercise a sort of direct rule over Iraq. Süleyman organized local administration, made a survey of the lands, appointed governors, sent Janissaries, and displayed some building activity, mainly in Baghdad, which became a prosperous centre of trade much admired by European travellers. Early in the 17th century, however, a Janissary officer named Bakr Ṣū Bāshī revolted, held sway over Baghdad—which he terrorized—and entered upon negotiations with Shāh ʿAbbās I (1587–1629) in order to strengthen his position, but Iraq was taken by the Ṣafavids in 1623, the Ottomans being able to preserve only Mosul and Shahrazūr. The Persians destroyed many buildings, including

Sunnī mosques, and several thousands of inhabitants were killed or sold as slaves.

Just as in al-Ḥasā a family of governors ruled from the late 16th century until 1663, so in Basra Afrāsiyāb an army officer of unknown origin purchased the governorship from the local pasha in 1612 and founded a small dynasty of governors who ruled the town until 1668 and favoured European penetration by opening the doors to English, Portuguese, and Dutch merchants trading in the Persian Gulf. Afrāsiyāb offered strong resistance to the Ṣafavids but did not support Baghdad against them.

Beginning of European penetration

In the meantime, Baghdad had been retaken in 1638 by Sultan Murad IV (1612–40), and a great many Shīʿīs were massacred. A treaty concluded in 1639 at Qaṣr-e Shīrīn between the Ottomans and Ṣafavids fixed the boundaries of the two states.

This Ottoman victory did not bring stability to Iraq, and the rulers had to suppress many disturbances caused by the turbulent garrison troops and unruly Arab and Kurdish tribes. In Lower Iraq, after the Afrāsiyāb dynasty was overthrown, the Arab tribes of al-Muntafiq and Ḥuwayza threatened the Ottoman supremacy, while Iranians penetrated into this region. After the Treaty of Karlowitz (1699) with Austria, Russia, Poland, and Venice, the Ottomans were able to send some troops to Iraq and recover Basra. In spite of the 1639 settlement, Baghdad was still the stake of struggles between Turks and Persians because of its being the most important city in the east of the Ottoman possessions.

From 1704 a new government system was introduced in Iraq by Hasan Paşa (governor from 1704 to 1724), his son Ahmad Paşa (1724–1747), and their successors. Themselves of Georgian origin, they relied on young Georgian slaves, trained in a special school, to become the chiefs of the main administrative services; those *mamlūk*s (slaves), known in Turkish as *kölemen,* sometimes reached governorship in Baghdad and applied personal policies; they suppressed revolts, restored order, annexed Basra, and promoted some prosperity in the country. In 1733, however, Nāder Shāh (later ruler of Iran from 1736 to 1747) besieged Baghdad but was compelled to withdraw from the capital; in 1742, he also failed to seize Mosul, governed by a native pasha, ʿAbd al-Jalīl, whose family, the Jalīlīs, ruled there from 1726 to 1834, and an agreement in 1746 confirmed the terms of the Qaṣr-e Shīrīn settlement. In 1775, Basra, where an agency of the British East India Company had been established as early as 1763, was lost for five years to Ṣādiq Khān, a brother of Karīm Khān Zand, the ruler of Persia. Süleyman Paşa (later known as Büyük the Great), governor of Basra, was transported to Shīrāz and came back on Karīm's death, took possession of Baghdad, and was given the governorship of Baghdad, Basra, and Shahrazūr by the Sultan. Under Büyük (1780–1802), the *Kölemen* rule reached its zenith. He fostered agriculture and trade, built several monuments, and, in 1798, permitted a permanent British resident to be appointed in the capital. Shortly before he died, however, Iraq was attacked by the Wahhābīs of Arabia, who penetrated into al-Ḥasā and, with their iconoclastic zeal, raided and sacked in 1802 Karbalāʾ, the holy city of Shīʿīs, despite the support of the Muntafiq confederacy to Baghdad.

The *kölemen* regime lasted until 1831, when Sultan Mahmud (1730–54) turned out Dawud Paşa from Baghdad after famine and flood had devastated and depopulated the city. From 1831, the Sublime Porte had high hand over Iraq, the history of which is henceforth characterized by administrative, economic, and social reform, together with European penetration. English, French, and Italian agencies and consulates were established in Basra. A company for river traffic between Basra and Baghdad was created; a telegraphic line linked Istanbul with Baghdad; but the Ottoman authority over the tribesmen was still weak. A new stage came with the governorship of Midhat Paşa (1869–72), who enforced the Ottoman law of *vilayet*s, introduced a new administrative system, including an elected town council, created a press and the earliest Iraqi newspaper (*az-Zawrāʾ*, 1869), founded primary and secondary military schools, and inaugurated the Baghdad-Kāẓimayn tramway line. This was a meritorious attempt at modernizing the country, at least Baghdad. Progress continued until the early 20th century, when Arabic-speaking Iraqi leaders began to lose confidence in the Ottomans and to dream about independence within the framework of a Pan-Arab and anti-Ottoman movement.

On November 20, 1914, an Anglo-Indian expeditionary force occupied Basra; Gen. Charles Townshend seized Kūt al-ʿAmārah in 1915 but was compelled to surrender on April 29, 1916; finally, the British took possession of Baghdad in March 1917, Mosul remaining in the hands of the Ottomans until the armistice of Mudros (October 30, 1918).

(C.L.P.P.)

IRAQ SINCE 1918

British occupation and the mandatory regime. The merging of the three provinces of Mosul, Baghdad, and Basra into one political entity and the creation of a nation out of the diverse religious and ethnic elements inhabiting these lands was accomplished by the events and circumstances following World War I. Action undertaken by the British military authorities during the war and the upsurge of nationalism after the war helped determine the shape of the new Iraqi state and the course of events during the postwar years, until Iraq finally emerged as an independent political entity in 1932.

British control of Iraq, however, proved to be short-lived. After the war Britain debated both its general policy in Iraq and the specific type of administration to establish. Two schools of thought influenced policy makers in London. The first, advocated by the Colonial Office, stressed a policy of direct control to protect British interests in the Persian Gulf and India. Assessing British policy from India, this school may be called the Indian school of thought. The other school, hoping to reconcile Arab nationalism, advised indirect control. In Iraq itself, British authorities were divided on the issue. Some, under the influence of Sir Arnold Wilson, the acting civil commissioner, advocated direct control; others, alarmed by growing dissatisfaction with the British administration, advised indirect control and suggested the establishment of an indigenous regime under British supervision. Britain was still undecided on which policy it should follow in 1920 when external events in other Arab countries radically changed the internal conditions of Iraq.

Early in 1920 the *amīr* Fayṣal I, son of the *sharīf* Ḥusayn, who had led the Arab Revolt in 1916, established an Arab government in Damascus and was proclaimed king of Syria. Meanwhile, a group of Iraqi nationalists met in Damascus to proclaim the *amīr* ʿAbd Allāh, older brother of Fayṣal, as king of Iraq. Under the influence of nationalist activities in Syria, nationalist agitation followed first in northern Iraq and then in the tribal areas of the Middle Euphrates. By the summer of 1920 the revolt had spread to all parts of the country except the big cities of Mosul, Baghdad, and Basra, where British forces were stationed.

Amīr ʿAbd Allāh proclaimed king

In July 1920, Fayṣal came into conflict with the French authorities over control of Syria. France had been given the mandate over Syria and Lebanon in April and was determined to obtain Fayṣal's acceptance of the mandate. Nationalists urged Fayṣal to reject the French demands, and conflict ensued between him and the French, resulting in his expulsion from Syria. Fayṣal went to London to complain about the French action.

Although the revolt in Iraq was suppressed by force, it prompted both Iraq and Great Britain to reconcile their differences. In Britain a segment of public opinion wanted to "get out of Mesopotamia" and urged relief from further commitments. In Iraq, the nationalists were demanding independence. When Fayṣal arrived in London, the question of Iraq had been the subject of discussion between the Colonial and Foreign offices. In 1921 Britain offered the Iraqi throne to Fayṣal along with the establishment of an Arab government under British mandate. Fayṣal wanted the throne if it were offered to him by the Iraqi people. He also suggested the replacement of the mandate by a treaty of alliance. These proposals were accepted by the British government, and Winston Churchill, then colonial

secretary, promised to carry them out. He was advised by T.E. Lawrence, known for his sympathy for the Arabs.

In March 1921 a conference presided over by Churchill was held in Cairo to settle Middle Eastern affairs. Fayṣal was nominated to the Iraqi throne, provided that a plebiscite be held to confirm the nomination. Sir Percy Cox, recently appointed a high commissioner for Iraq, was entrusted with carrying out the plebiscite before Fayṣal's accession to the throne. A provisional government set up by Cox shortly before the Cairo Conference passed a resolution on July 11, 1921, declaring Fayṣal king of Iraq, provided that his "Government shall be constitutional, representative and democratic." The plebiscite confirmed this proclamation, and Fayṣal was formally crowned king on August 23, 1921.

The establishment of the monarchy was the first step in the establishment of a national regime. Two other steps followed immediately: the signing of a treaty of alliance with Great Britain and the drafting of a constitution. It was deemed necessary that a treaty should precede the constitution and define relations between Iraq and Britain. The treaty was signed on October 10, 1922. Without direct reference, it reproduced most of the provisions of the mandate. Iraq undertook to respect religious freedom and missionary enterprises and the rights of foreigners, to treat all states equally, and to cooperate with the League of Nations. Britain was obligated to offer advice on foreign and domestic affairs, such as military, judicial, and financial matters (defined in separate and subsidiary agreements). Although the terms of the treaty were open to periodic revision, they were to last 20 years. In the meantime, Britain agreed to prepare Iraq for membership in the League of Nations "as soon as possible."

Treaty of alliance with Britain

It soon became apparent that the substance, though not the form, of the mandate was still in existence and that complete independence had not been achieved. Strong opposition to the treaty in the press made it almost certain it would not be ratified by the Constituent Assembly. Nor was British public opinion satisfied with the commitments to Iraq. During the general elections of 1922 there was a newspaper campaign against British expenditures in Iraq. In deference to public opinion, both in Britain and Iraq, a protocol to the treaty was signed on April 30, 1923, reducing the period of the treaty from 20 to four years. Despite the shortening of British tutelage, the Constituent Assembly demanded complete independence when the treaty was put before it for approval. Ratification of the treaty was accomplished on June 11, 1924, after England's warning that nonapproval would lead to the referral of the matter to the League of Nations.

The Constituent Assembly then moved to consider a draft constitution drawn up by a constitutional committee. The committee tried to give extensive powers to the king. Discussion on the draft constitution by the Constituent Assembly lasted a month, and, after minor modifications, it was adopted in July 1924. The Organic Law, as the constitution was then called, went into effect right after it was signed by the King on March 21, 1925. It provided for a constitutional monarchy, a parliamentary government, and a bicameral legislature. The latter was composed of an elected House of Representatives and an appointed Senate. The lower house was to be elected every four years in a free manhood suffrage. The first Parliament met in 1925. Ten general elections were held before the downfall of the monarchy in 1958, and over 50 Cabinets formed during the same period reflected the instability of the system.

The Organic Law

From the establishment of a national government, there was keen interest in organizing political parties. Three parties, organized in 1921, one by the group in power and two by opposition parties, had similar social and economic views and essentially the same political objective: termination of the mandate and the winning of independence. They differed, however, in the means of achieving the objective. After the achievement of independence in 1932, these parties dissolved, because their *raison d'être* had disappeared. It was only when social issues were discussed that new political groupings, even if not formally organized as political parties, began to emerge. The power struggle among these groups became exceedingly intense after World War II.

The Iraqi nationalists, though appreciating the free expression of opinion permitted under a parliamentary system, were far from satisfied with the mandate. They demanded independence as a matter of right, as promised in war declarations and treaties, rather than as a matter of capacity for self-government as laid down in the mandate. Various attempts were made to redefine Anglo-Iraqi relations, as embodied in the 1926 and 1927 treaties, without fundamentally altering Britain's responsibility. The British treaties were deemed by the nationalists as not only impeding the realization of Iraq's nationalist aspirations, but also inimical to the economic development of the country. The nationalists viewed the situation as a "perplexing predicament" (*al-waḍ ʿash-shadh*) —a term that became popular in Parliament and the press. It referred to the impossibility of government by the dual authority of the mandate. The nationalists argued that there were two governments in Iraq, one foreign and the other national, and that such a regime was an abnormality that, though feasible in theory, was unworkable in practice.

The "perplexing predicament"

In 1929 Britain decided to end this stalemate and reconcile its interests with Iraq's national aspirations. It notified Iraq that the mandate would be terminated in 1932 and a new treaty of independence negotiated. A new government was formed, headed by Gen. Nūrī as-Saʿīd, who helped in achieving Iraq's independence.

The new treaty was signed on June 30, 1930. It provided for the establishment of a "close alliance" between Britain and Iraq, with "full and frank consultation between them in all matters of foreign policy which may affect their common interests." Iraq would maintain internal order and defend itself against foreign aggression, supported by Britain. Any dispute between Iraq and a third state, involving the risk of war, was to be discussed with Britain in the hope of a settlement in accordance with the Covenant of the League of Nations. In the event of an imminent menace of war, the two parties would take a common defense position. Iraq recognized that the maintenance and protection of essential British communications was in the common interest of both parties. Air-base sites for maintaining British troops were therefore granted and installed near Basra and west of the Euphrates, but these forces "shall not constitute in any manner an occupation, and will in no way prejudice the sovereign rights of Iraq." This treaty, valid for 25 years, was to come into effect after Iraq had become a member of the League of Nations. On October 3, 1932, Iraq was admitted to the League of Nations as an independent state.

Independence, 1932–39. Since conflict among Iraq's political leaders centred essentially on how to end the mandate rather than on the right of independence, King Fayṣal sought the cooperation of opposition leaders after independence. Shortly after Iraq's admission to the League, General Nūrī, who had been prime minister since 1930, resigned. After an interim administration, King Fayṣal invited Rashīd ʿĀlī al-Gaylānī, one of the opposition leaders, to form a new government. For a short while it seemed that all the country's leaders would close ranks and devote all their efforts to internal reforms.

But internal dissension soon developed. The first incident was the Assyrian uprising of 1933. The Assyrians, a small Christian community living in Mosul Province, were given assurances of security by both Britain and Iraq. When the mandate was ended, the Assyrians began to feel insecure and demanded new assurances. Matters came to a head in the summer of 1933 when King Fayṣal was in Europe. The opposition, now in power, wanted to impress the public through a high-handed policy toward a minority group. In clashes with the Iraqi troops several hundred Assyrians were brutally killed. The incident was brought to the attention of the League less than a year after Iraq had given assurances that it would protect minority rights. Had King Fayṣal been in the country, he would have counselled moderation. Upon his hasty return to Baghdad, he found deep-seated divisions and events beyond his control. Suffering from heart trouble, he returned to Switzerland, where he died in September 1933. The Assyrian incident brought about the fall of Rashīd ʿĀlī and his replacement by a moderate government.

The Assyrian uprising of 1933

Fayṣal was succeeded by his son, King Ghāzī (1912–39), who was still young and inexperienced—a situation that gave political leaders an opportunity to compete for power. Without political parties to channel their activities through constitutional processes, politicians resorted to extraconstitutional, or violent, methods. One extraconstitutional method involved clever political manoeuvring. Opposition leaders tried to embarrass those in power by press attacks, palace intrigues, or incidents that would cause Cabinet dissension and force the prime minister to resign. The first five governmental changes after independence, from 1932 to 1934, were produced by these methods.

The second device used was the incitement of tribal uprisings in areas where there were tribal chiefs unfriendly to the group in power. Tribes, though habitually opposed to authority, had been brought under control and remained relatively quiet after 1932. When opposition leaders began to incite them against the government in 1934, however, they rebelled and caused the fall of three governments from 1934 to 1935.

The third method was military intervention. The opposition tried to obtain the loyalty of army officers, plan a coup d'etat, and force those in power to resign. This method, often resorted to by the opposition, proved to be the most dangerous because, once the army intervened in politics, it became increasingly difficult to reestablish civilian rule. From 1936 until the army was defeated in a war with Britain, in 1941, the army dominated domestic politics. The army again intervened in 1958, and it continues to dominate the political scene.

The coup d'etat of 1936

Two different sets of opposition leaders produced the first military coup d'etat in 1936. The first group, led by Ḥikmat Sulaymān, was a faction of old politicians who sought power by violent methods. The other was the Ahalī group, composed mainly of young men who advocated Socialism and democracy and sought to carry out reform programs. It was Ḥikmat Sulaymān, however, who urged Gen. Bakr Ṣiqī, commander of an army division, to stage a surprise attack on Baghdad, in cooperation with another military commander, and forced the cabinet to resign. It seems that young King Ghāzī had also been disenchanted with the group in power and so allowed the government to resign. Ḥikmat Sulaymān became prime minister in October 1936, and Bakr Ṣidqī was appointed chief of the general staff. Neither the Ahalī group nor Ḥikmat Sulaymān could improve social conditions, however, because the army gradually dominated the political scene. Supported by opposition leaders, a dissident military faction assassinated Bakr Ṣidqī, but civilian rule was not reestablished. The significance of this first military coup lies in the fact that it introduced a new factor in politics. Lack of leadership after the assassination of Bakr Ṣidqī left the army divided, while jealousy among leading army officers induced each faction to support a different set of civilian leaders. The army continued to influence political decisions behind the scenes and became virtually the deciding factor in Cabinet changes from 1936 to 1941.

Despite political instability, material progress continued under King Ghāzī's short reign: the Kūt al-ʿAmarah irrigation project, begun in 1934, was completed; other projects, to be financed by oil royalties, were laid down; the pipelines from the Kirkūk oil fields to the Mediterranean were opened in 1935; the railroads, still under British control, were purchased in 1935, and the Baʿījī-Tal Küçük section, the only missing railway link between the Persian Gulf and Europe, was completed in 1938; and there was a noticeable increase in construction, foreign trade, and in educational facilities. Several outstanding disputes with Iraq's neighbours were settled, including the dispute with Iran over the Shatt al-Arab in 1936 and the boundary with Syria, which was settled in Iraq's favour to include within its boundary the Jabal Sinjār. A nonaggression pact, called the Saʿdābād Pact, between Turkey, Iran, Afghanistan, and Iraq was signed in 1937. Arab cooperation was stressed in a nonaggression agreement with Saudi Arabia and Yemen. In 1939, shortly before the outbreak of World War II, King Ghāzī was killed in a car accident and his son Fayṣal II ascended the throne. As Fayṣal was only four years old, Fayṣal's uncle, Amīr ʿAbd al-Ilāh, was appointed regent and served in this capacity for the next 14 years.

World War II and British intervention, 1939–45. General Nūrī, author of the 1930 treaty, was prime minister when World War II broke out. Believing that the Anglo-Iraqi alliance was the best guarantee for Iraqi security, he wanted to declare war on Germany, but his ministers counselled caution as British victory was then in doubt. General Nūrī accordingly declared Iraq nonbelligerent and severed diplomatic relations with Germany. When Italy entered the war, however, Nūrī, then Minister of Foreign Affairs in Rashīd ʿĀlī al-Gaylānī's Cabinet, was unable to persuade the Cabinet to break off diplomatic relations with Italy. Under the influence of Pan-Arab leaders, public opinion in Iraq radically changed after France's fall, becoming especially hostile to Britain because other Arab countries remained under foreign control. Pan-Arabs urged Iraqi leaders to free Syria and Palestine and achieve unity among Arab countries. Extremists advocated alliance with Germany as the country that would foster independence and unity among Arabs.

Hostility toward Britain

Rashīd ʿĀlī was at first unwilling to side with the extremists and gave lip service to the Anglo-Iraqi alliance. But dissension among the Iraqi leaders forced Rashīd ʿĀlī to side with the Pan-Arabs. Leading army officers also fell under Pan-Arab influences and encouraged Rashīd ʿĀlī to detach Iraq from the British alliance. During 1940 and 1941 Iraqi officers were unwilling to cooperate with Britain, and the Pan-Arab leaders began secret negotiations with the Axis Powers. Britain decided to send reinforcements to Iraq. Rashīd ʿĀlī, while allowing the landing of a small British force in 1940, was overthrown from power early in 1941, but he was reinstalled by the army in April and refused further British requests for reinforcements.

British contingents entered Iraq from the Persian Gulf as well as from the Ḥabbānīyah air base during April and May 1941; armed conflict with the Iraqi forces followed. The hostilities lasted only 30 days, during which period a few Iraqi leaders, including the regent and General Nūrī, fled the country. By the end of May the Iraqi Army capitulated, and Rashīd ʿĀlī and his Pan-Arab supporters left the country.

The return of the regent and moderate leaders through British intervention had far-reaching consequences. Britain was given what it demanded: the use of transportation and communication facilities and a declaration of war on the Axis Powers in January 1942. Rashīd ʿĀlī's supporters were dismissed from the service and some of them interned for the duration of the war. Four officers who were responsible for the war were hanged.

Postwar reconstruction and social upheavals, 1945–58. During World War II, liberal and moderate Iraqi elements began to play an active political role. The entry of the United States and the Soviet Union into the war on the side of Britain and their declarations in favour of democratic freedoms greatly enhanced the position of the Iraqi democratic elements. The people endured shortages and regulations restricting personal liberty and the freedom of the press, trusting that the end of the war would bring the promised better way of life. The government, however, paid no attention to the new spirit, and the wartime regulations and restrictions continued after the war. The regent called a meeting of the country's leaders in 1945 and made a speech in which he attributed public disaffection to the absence of a truly parliamentary system. He called for the formation of political parties and promised full freedom for their activities and the launching of social and economic reforms.

The immediate reactions to the regent's speech were favourable, but, when political parties were formed in 1946 and certain regulations abolished, the older politicians and vested interests resisted, and the new government formed in January 1946 to honour the regent's promises was overthrown within a few months of its inception. General Nūrī then became prime minister and tried to enlist the cooperation of political parties, but the general elections held under his government's supervision were no different from previously controlled elections. The parties boycotted

Ṣāliḥ Jabr's government

the elections. General Nūrī resigned in March 1947, and Ṣāliḥ Jabr formed a new government.

Jabr, the first Shī'ī politician to become a prime minister, included in his cabinet a number of young men, but he himself was unacceptable to some liberal and nationalist elements who had been roughly handled when he was wartime Minister of Interior. Jabr tried to help the Arabs in Palestine in order to improve his image in nationalist circles, but he mishandled opposition leaders. Most damaging was his attempt to replace the Anglo-Iraqi treaty of 1930 by another—without consultation with the country's leaders. When his attention was called to the need for consultation, he consulted only with older politicians and excluded younger leaders.

Jabr entered into negotiations with Britain with the intention of enhancing his own position. When he found that Britain wanted to retain control of its air bases in Iraq, he insisted that Britain accept the principle of Iraqi control of the bases and allow Iraq the use of the bases for defensive purposes. He threatened to resign if Britain refused his proposals. It was with this understanding that Jabr proceeded to London early in 1948 to negotiate a new treaty. Jabr and Ernest Bevin, the British foreign secretary, quickly came to an agreement and signed a new 20-year treaty at Portsmouth on January 15, 1948. It provided for a new alliance between Iraq and Britain on the basis of equality and complete independence and stated "each of the high contracting parties undertake not to adopt in foreign countries an attitude which is inconsistent with the alliance or which might create difficulties for the other party." The annex to the treaty stressed the importance of the air bases as "an essential element in the defense of Iraq," but Britain's use of the air bases in the event of war, or threat of war, would be dependent on Iraq's invitation. The treaty also provided for the establishment of "a joint defence board," equally represented by Britain and Iraq, for common defense and consultation. Both parties agreed to grant each other necessary facilities for defense purposes. An improvement of the 1930 treaty, this treaty sought an alliance on the basis of mutual interests. The two air bases, which were often the subject of criticism, were returned to Iraq. British forces were to be evacuated, and Iraq would be supplied with arms and military training.

Anti-British rioting

Despite these advances, the treaty was repudiated immediately in a popular uprising. Street demonstrations had been going on before the treaty was signed, in defense of Arab rights in Palestine, but, when the news of the signing of the new treaty was broadcast in London, rioting and demonstrations in Baghdad led to clashes between students and the police. The regent called a meeting at the royal household on January 21, which was attended by the country's leaders. After deliberations, it was decided to repudiate the treaty. Jabr returned to Baghdad to defend his position but to no avail. Rioting, demonstrations, and clashes with the police increased, and Jabr was forced to resign. Obviously the new treaty was not the root cause of the uprising. The uprising was the culmination of events in a struggle between the liberal and young leaders who wanted to participate in political activities and the older leaders who insisted on excluding the younger elements from public affairs. The conflict between the older and younger leaders continued after the treaty had been rejected. The older politicians returned to power under General Nūrī's leadership. By 1952 another popular uprising, stirred by opposition leaders and carried out by students and extremists, again manifested itself in rioting and street demonstrations. The police were unable to control the mob, and the regent called on the army to maintain public order. The chief of the general staff governed the country under martial law for more than two months. Civilian rule was restored at the beginning of 1953, but there was no sign that the country's older leaders were prepared to share authority with their opponents. Meanwhile, King Fayṣal II, who had come of age, began to exercise his formal powers, and the period of regency came to an end. It was hoped that the regent would withdraw from active politics and allow the political forces of the country to create a new order. The former regent, who became the crown prince, continued to control political events from behind the scene. The struggle for power among the leaders continued with increasing intensity until the downfall of the monarchy in 1958.

Development of Iraq oil in the 1950's

Despite political instability, Iraq achieved material progress during the 1950s, thanks to a new oil agreement that increased royalties and the establishment of the Development Board. The oil agreement between the Iraqi government and the Iraq Petroleum Company (IPC), originally signed in 1925, had yielded relatively modest royalties, owing to certain technical limitations, such as construction of pipelines, and to war conditions. It was not until 1952 that construction of pipelines to Banias was completed. There were certain points of dispute between the government and the IPC that were not entirely resolved. The nationalization of the oil industry in Iran and the announcement of the 1950 agreement between Saudi Arabia and Aramco (Arabian American Oil Company), on a half-and-half basis of payment, induced the Iraqi government and IPC to negotiate a new agreement on the new pattern of the division of profits. Some of the opposition leaders demanded the nationalization of the oil industry, but the Iraq government and IPC, forestalling any serious move for nationalization, agreed to negotiate on the basis of the fifty-fifty formula, to the mutual advantage of Iraq and the company. The new agreement was signed in 1952, allowing Iraq to take in kind as part of its share and to receive an increasing amount of royalties specifically agreed upon between the two parties. It was then estimated that Iraq would receive no less than £30 million in 1953 and 1954 and no less than £50 million in 1955 and subsequent years.

Creation of the Development Board

In 1950 the government created the independent Development Board, an agency immune from political pressures and responsible directly to the prime minister. This agency provided for a body of six executive members, three of whom were to be experts in some branch of the development program. The prime minister, as chairman, and the minister of finance were ex-officio members. An amendment to the law increased membership by two and provided for a minister of development responsible directly to the head of the Cabinet. These members were appointed by the Cabinet, given equal voting rights, and might not hold any other official position. Two foreign members held positions as experts, and the Iraqi members were selected on merit and past experience. The board was composed of a council and ministry. Its staff was divided into technical sections and the ministry into a number of departments. The technical sections were for irrigation, flood control, water storage, drainage, transportation, and industrial and agricultural development. The board was financed from 70 percent of oil royalties and from loans and revenues from the board's own projects. The estimated expenditure of the first six-year program (1951–56) was £155,000,000, and in 1950 the World Bank provided a loan of $12,800,000 for the Wādī ath-Tharthār flood-control project.

The Tharthār project, completed in 1956, connects the River Tigris at Samarrā' (some 60 miles north of Baghdad) with the Wādī ath-Tharthār Depression. A barrage was constructed on the Euphrates at Ramādī, making it possible to divert water into Lake Ḥabbānīyah. Other flood-control plans, such as that at Bakhma (completed after the revolution of 1958), were laid down. Extensive work on bridges and public buildings, including schools, hospitals, a new Parliament building, and a royal house, was started, and it was completed after 1958. This work, especially the work on dams and irrigation projects, was a long-term investment, and many short-term projects of more direct benefit to the population were neglected. Opposition leaders attacked the Development Board for the stress on long-term projects whose beneficiary, they held, were the vested interests—landowners and tribal chiefs. Despite criticism, the board maintained an independent status rarely enjoyed by any other government department. Nevertheless, the public remained unaware of the far-reaching effects of the projects undertaken, while the opposition attacked the board for squandering funds on contracts given to wealthy landlords and influential politicians.

The revolution of 1958. Despite material progress, the monarchy failed to win public support and, in particular,

the confidence of the younger generation. Before the revolution, Iraq lacked an enlightened leadership capable of achieving progress and inspiring public confidence. The new generation offered such leadership, but the older leaders resisted and embarked on an unpopular foreign policy, including an alliance with Britain through participation in the Baghdad Pact and opposition to the establishment of the United Arab Republic (U.A.R.).

Rise of the Free Officers

The failure of younger civilians to obtain power aroused the concern of some young military officers who, required by military discipline to take no part in politics, called themselves the Free Officers and began to organize in small groups and to lay down revolutionary plans. The number of Free Officers was relatively small, but there was a considerably larger number of sympathizers. The officers worked in cells, and the identity of the participants was kept secret. Only the Central Organization, which supplied leadership of the movement, was known to all the Free Officers. The Central Organization was composed of 14 officers, headed by Abdul Karim Kassem, who held the highest military rank.

Of the several plots proposed, that laid down by Kassem and his close collaborator ʿAbd as-Salām ʿĀrif proved the most appropriate. The general staff issued an order to one of the brigades, in which ʿĀrif served, to proceed to Jordan on July 14, 1958, to reinforce Jordanian forces against alleged threats by Israel. Brigadier Kassem, in command of another brigade, was to protect the brigades going to Jordan. Kassem and ʿĀrif agreed that as the brigade proceeding to Jordan passed through Baghdad it would capture the city.

On July 14 the revolutionary forces captured the capital, declared the downfall of the monarchy, and proclaimed a republic. The leading members of the royal house, including the King and Crown Prince, and General Nūrī were executed. Kassem, head of the revolutionary force, formed a Cabinet, over which he presided, and appointed himself commander of the National Forces. He also assumed the portfolio of defense and appointed ʿĀrif both minister of the interior and deputy commander of the National Forces. A Council of Sovereignty, composed of three persons, was to act as head of state.

A provisional constitution declared that Iraq formed an integral part "of the Arab nation," and that "Arabs and Kurds are considered partners in this homeland." Iraq was declared a republic and Islām the religion of the state; all executive and legislative powers were entrusted to the Sovereignty Council and the Cabinet. It soon became clear, however, that power rested in Kassem's hands, supported by the army.

Pan-Arabism versus Iraqi nationalism

Conflicts among the officers developed, first between Kassem and ʿĀrif, and then between Kassem and his supporters. ʿĀrif championed the Pan-Arab cause and advocated Iraq's union with the U.A.R. Kassem rallied the forces against Arab unity—Kurds, Communists, and others—and stressed Iraq's own identity and internal unity. ʿĀrif was dropped from power in October, but in the following year Kassem's power was threatened by other factions. Kassem tried to divert public attention to foreign affairs by putting forth Iraq's claim to Kuwait's sovereignty in June 1961. This claim brought him into conflict not only with Britain and Kuwait but also with the other Arab countries. He opened negotiations with the Iraq Petroleum Company to increase Iraq's royalties, but his extreme demands resulted in the breakdown of negotiations in 1961. Public Law 80 was enacted to prohibit the granting of concessions to any foreign company and to transfer control over all matters connected with oil to an Iraq National Oil Company (INOC).

By 1963 Kassem had become isolated internally as well as externally; the only great power with which he remained friendly was the Soviet Union. When one faction of the army, in cooperation with one Arab nationalist group—the Iraqi regional branch of the Arab Socialist Baʿth ("Renaissance") Party—started a rebellion in February 1963, the Kassem regime suddenly collapsed, and Kassem was executed.

Recurrence of military coups, 1963–68. The military faction that brought about the collapse of the Kassem regime preferred to remain behind the scenes rather than to assume direct responsibility. The Baʿth Party, a group of young activists who advocated Arab nationalism and Socialism, was entrusted with power. Baʿth leaders invited ʿAbd as-Salām ʿĀrif to assume the presidency. A National Council for Revolutionary Command (NCRC), composed of civilian and military leaders, was established to assume legislative and executive powers. The premiership was entrusted to Col. Aḥmad Ḥassan al-Bakr, a Baʿthist officer.

Some of the Baʿth leaders wanted to carry out Baʿth Socialist ideas; others advised more caution. A compromise was finally reached in which the party's goals—Arab unity, freedom, and Socialism—were reaffirmed in principle, but it was decided to adopt a transitional program. Industrialization and economic development were stressed, and the role of the middle class was recognized. The dissension among Baʿth leaders, however, soon led to the collapse of the regime. President ʿĀrif, whose powers initially had been restricted by the Baʿth leaders, rallied the military forces to his side. On November 18, 1963, he placed the leaders of the Baʿth Party under arrest and took over control, becoming in fact, as in name, the real ruler of the country. In May 1964, a new provisional constitution was promulgated in which the principles of Arab unity and Socialism were adopted, and, in July, the banks and a number of the country's industries were nationalized.

ʿĀrif's alliance with the military forces

The idea of Arab Socialism attracted only a small group in Iraq, and ʿĀrif began to discover its unfavourable effects on the country. ʿĀrif himself had never been a believer in Socialism, but he had adopted it under the influence of Egypt. The adverse influence of nationalization gave him an excuse to replace the ideological group that supported Socialism with others who would pay attention to the reality of Iraq's economic conditions. Nor had ʿĀrif been happy with the group of officers who had elevated him to power. He began to prepare the way to entrust power to civilian hands willing to be guided by him as chief executive.

In September 1965 ʿĀrif invited Abd ar-Rahman al-Bazzaz, a distinguished lawyer, diplomat, and writer on Arab nationalism, to form a new government. Bazzaz did not feel that he should abolish Arab Socialism, but he offered to increase production and create a balance between the public and private sectors.

The Arab Baʿth Socialist regime since 1968. After the Baʿth fell from power, an internal conflict between Marxists and nationalists virtually crippled the party. In 1964 Aḥmad Ḥassan al-Bakr, an army officer highly respected by the party, was elected leader and put an end to internal dissension.

Premiership of Bazzaz

ʿĀrif died suddenly in a helicopter accident in April 1966. Even before his death, Premier Bazzaz, known for his opposition to military interference in politics, had begun to talk about the need to hold elections for a representative assembly; military officers pressured the new president, ʿAbd ar-Raḥman ʿĀrif, elder brother of the late president, to remove Bazzaz, and the Cabinet resigned in August 1966. Power remained in military hands, but factionalism in the army was accentuated and leadership frequently changed hands. The Arab defeat in the Arab–Israeli war of 1967, in which Iraq took only a nominal role, led to intense unrest within the country and within the party. The Baʿth, joined by other opposition leaders, called for the formation of a coalition government and the holding of general elections for a National Assembly. President ʿĀrif paid no attention to their demands.

The revolution of 1968. By July 1968, for the fourth time since the overthrow of the monarchy in 1958, the army had become the arbiter between the group in power and the opposition. Col. ʿAbd ar-Raḥman ad-Daud, Col. Hammad Shihab, and Col. ʿAbd ar-Razzaq an-Nayīf, leaders of the military officers, agreed to cooperate with the Baʿth Party, which accepted their help as a means to achieve power but which intended to bridle them at the earliest possible moment. It was agreed that Bakr, the Baʿth leader, would be president of the republic, Nayīf prime minister, and the Cabinet a coalition of various shades of opinion. The uprising, led by Bakr, took place on July 17, 1968.

The Revolutionary Command Council

The first official step was the establishment of a Revolutionary Command Council (RCC), which assumed supreme authority. Within the Cabinet a struggle for power began immediately between the Ba'th and the Nayīf-Daud group, chiefly over Socialism and foreign policy. The Ba'th leaders decided to remove the Nayīf-Daud group. On July 30, 1968, a group of officers led by Saddam Hussein at-Takrītī arrested Nayīf, who agreed to resign. Daud, on a mission in Jordan, agreed to remain there.

Bakr assumed the post of premier in addition to the presidency and presided over the RCC and the Regional Command of the Ba'th Party. Most of the Cabinet posts were given to Ba'th leaders. A temporary constitution, defining the regime as democratic and Socialist, was issued on September 21. Until the calling of a National Assembly, all powers were entrusted to the RCC. The party itself, already highly organized, began to infiltrate the government and almost all national organizations.

National Front. An uprising led by the head of the security police in 1973 had far-reaching effects on the Ba'th regime, which purged some of the security police and laid down rules curtailing security activities. Specific steps were undertaken to strengthen the regime by the establishment of a National Front and a settlement of the conflict with the Kurds.

Before the front could be established, it had been necessary to eliminate misunderstandings between the Ba'th and Communist parties. A Charter for National Action was made public in 1971, and in March 1972 the Ba'th and Communist leaders agreed to draw up a statute that defined the front as a voluntary union among progressive parties and groups that accepted cooperation under the charter. The statute also provided for a 16-member central executive committee, called the High Committee of the Progressive National Front (PNF), that included the Ba'th (eight seats), the Iraqi Communist Party (ICP; three), the Kurdish Democratic Party (three), and the Progressive Nationalists and Independent Democrats (one each). The Progressive National Front was declared in 1973.

Though President Bakr continued to preside over the RCC and the Iraqi Regional Command, Saddam Hussein, leader of the coup of July 30, 1968, emerged as the unrivalled leader not only in Ba'th circles but also in the country as a whole. In July 1979 Bakr resigned and Hussein succeeded as president of the republic in accordance with the temporary constitution. He promptly suppressed the ICP and other opponents.

Economic development to 1980. Perhaps the greatest assets of the Ba'th regime were the ambitious plans for reconstruction and the development laid down by its leaders. The struggle for power during 1958–68 had left little or no time for constructive work, and the Ba'th Party not only sought to transform the economic system from free enterprise to collectivism but also to assert the country's economic independence. The immediate objectives were to increase production and to raise the standard of living, but the ultimate objective was to establish a Socialist society in which all citizens would enjoy the benefit of progress and prosperity.

Socialist goals

The Five-Year National Economic Plan of 1965–70, followed by the Five-Year Plan of 1971–75, concentrated on raising the level of production, both in agriculture and industry, and aimed at reducing dependence on oil revenues as the primary source for development. But agriculture lagged far behind target, and industrial development was slow. In the third Five-Year Plan (1976–80), greater emphasis on agricultural production was noticeable, and industrial production slowed down.

Nationalization of the oil industry

The nationalization of the oil industry was considered by the Ba'th leaders as perhaps their greatest achievement. Between 1969 and 1972 a number of agreements with foreign powers—the Soviet Union and others—were concluded to provide the Iraq National Oil Company (INOC) with capital and technical skill to exploit the oil fields. In 1972 the operation of the North Rumayla field, rich in oil, started, and an Iraqi Oil Tankers Company was established to deliver oil to several foreign countries. Also in 1972 the Iraq Petroleum Company (IPC) was nationalized (with compensation) and a national company, the Iraqi Company for Oil Operations, was established to operate the fields. In 1973, when the fourth Arab–Israeli war broke out, Iraq nationalized U.S. and Dutch companies, and in 1975 it nationalized the remaining foreign interests in the Basra Petroleum Company.

Agrarian reform

The initial step in agrarian reform had been taken with the Agrarian Reform Law of 1958, which provided for the distribution to peasants of lands in excess of a certain minimum ownership. A decade later, less than half of the land had been distributed. In 1969 a revised Agrarian Reform Law relieved the peasants from payments for their land by abolishing compensation to landowners, and a year later a new Agrarian Reform Law was designed to improve the conditions of the peasantry, increase agricultural production, and correlate development in rural and urban areas. The results were disappointing, however, because of the difficulty of persuading the peasants to stay on their farms and their inability to improve the quality of agricultural production.

The Ba'th regime also completed work on irrigation projects that had already been underway and began new projects in areas where water was likely to be scarce in the summer. In the Five-Year Plan of 1976–80, funds were allocated to complete dams on the Euphrates, Tigris, Diyālā, and Upper Zab rivers and the lake known as Baḥr ath-Tharthār (in northern Iraq). Recognizing that a rapid transformation to full Socialism was neither possible nor in the country's best interest, the Ba'th provided for a private, though relatively small, sector for private investors, and a third, mixed sector was created in which private and public enterprises could cooperate.

Foreign relations. In 1972 a treaty of alliance with the Soviet Union was signed by which the two countries agreed to cooperate in political, economic, and military affairs, including the supply to Iraq of Soviet weapons. Iraq supported aims of the Palestine Liberation Organization, and in 1973 Iraq joined Egypt and Syria in the war they launched against Israel. Iraq disagreed with the United States on the settlement of the Arab–Israeli conflict, but economic and cultural relations between the two nations improved.

The war with Iran. In 1975 Iraq and Iran reached an agreement on their frontier differences in the Shatt al-Arab, but when the Iranian monarchy was overthrown in 1979, Iraq, threatened by the spread of revolution from Iran, denounced the settlement and demanded restoration of the eastern bank of the Shatt al-Arab as the border. The dispute erupted into all-out war with an Iraqi invasion of Iran in September 1980. Although air attacks on oil facilities proved spectacular, casualties on both sides at first remained low, owing to inefficient organization and lack of technical skill. As the conflict continued, however, a seesaw pattern of resurgences led to mounting losses of both equipment and personnel.

Economic consequences

Economic development slowed considerably as military spending increased. The damage to oil facilities and transportation lines led to a sharp decline of oil exports and revenue, and the government was forced to borrow heavily and to adopt austerity measures at home.

For later developments in the political history of Iraq, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL. (M.Kh./Ed.)

BIBLIOGRAPHY

General works on Iraq: A good general summary in English is RICHARD F. NYROP (ed.), *Iraq: A Country Study* (1979). Also useful is *The Middle East and North Africa* Europa publication. A political history in the 20th century is GEORGE LENCZOWSKI, *The Middle East in World Affairs,* 3rd ed. (1962). Geography and related aspects are presented in G.B. CRESSEY, *Crossroads: Land and Life in Southwest Asia* (1960); *Asia's Lands and Peoples,* 3rd ed. (1963); and S.H. LONGRIGG, *The Middle East,* 2nd ed. (1970). Information on the people and the social system is found in DORIS G. ADAMS, *Iraq's People and Resources* (1958); FUAD BAALI, *Relation of the People to the Land in Southern Iraq* (1966); GABRIEL BAER, *Population and Society in the Arab East* (1964; originally published in Hebrew, 1960; repr., 1976); FREDRIK BARTH, *Principles of Social Organization in Southern Kurdistan* (1953; repr., 1977); E.S.S. DROWER, *The Mandeans of Iraq and Iran* (1962); ROBERT A. FERNEA, *Shaykh and Effendi: Changing Patterns of Authority*

Among the El Shabana of Southern Iraq (1970); and S.M. SALIM, *Marsh Dwellers of the Euphrates Delta* (1962). Information on the national economy and related questions is found in S.N. ALITOVSKY, *Agrarian Changes in the Republic of Iraq, 1958–1965* (1966, originally published in 1966); KATHLEEN M. LANGLEY, *The Industrialization of Iraq* (1961); FAHIM I. QUBAIN, *The Reconstruction of Iraq, 1950–1957* (1958); A.R. AL-RUBAAI, *The Monetary and Banking System in Iraq* (1962); TAHER HAMDI KANAAN, *Input-Output and Social Accounts of Iraq, 1960–1963* (1965); UNITED STATES DEPARTMENT OF AGRICULTURE, *The Agricultural Economy of Iraq* (1965); and DOREEN WARRINER, *Land Reform and Development in the Middle East: A Study of Egypt, Syria, and Iraq,* 2nd ed. (1962; repr., 1976). Political, social, and administrative questions are dealt with in KAMEL ABU JABER, *The Arab Baʿth Socialist Party: History, Ideology and Organization* (1966); ARAB INFORMATION CENTER, *Education in the Arab States* (1966); H. BATATU, *The Old Social Classes and the Revolutionary Movements of Iraq* (1978); E. BE'ERI, *Army Officers in Arab Politics and Society* (1970; originally published in Hebrew, 1966); LORD BIRDWOOD, *Nuri al-Said: A Study in Arab Leadership* (1959); C.J. EDMONDS, *Kurds, Turks and Arabs* (1957); S.N. FISHER (ed.), *The Military in the Middle East: Problems in Society and Government* (1963); W.J. GALLMAN, *Iraq Under General Nuri: My Recollections of Nuri al-Said, 1954–1958* (1964); G.M. HADDAD, *Revolutions and Military Rule in the Middle East* (1965); MAJID KHADDURI, *Independent Iraq, 1932–1958: A Study of Iraqi Politics,* 2nd ed. (1960), *Republican Iraq: A Study in Iraqi Politics Since the Revolution of 1958* (1969), and *Socialist Iraq: A Study in Iraqi Politics Since 1968* (1978); DERK KINNANE, *The Kurds and Kurdistan* (1964); A. KELIDAR, *The Integration of Modern Iraq* (1979); S.H. LONGRIGG, *Iraq, 1900–1950: A Political, Social and Economic History* (1956); and with FRANK STOAKES, *Iraq* (1958); ABID AL-MARAYATI, *A Diplomatic History of Modern Iraq* (1961); F.J. SNELL ("CARACTACUS"), *Revolution in Iraq* (1959); DESMOND STEWART and JOHN HAYLOCK, *New Babylon: A Portrait of Iraq* (1956); and the UNITED STATES DEPARTMENT OF LABOR, *Labor Law and Practice in Iraq* (1962).

Mesopotamia until 1600 BC: See the relevant chapters of *The Cambridge Ancient History,* rev. ed. (1962–68), each written by an individual author and richly documented; *Fischer Weltgeschichte, Die Altorientalischen Reiche,* 3 vol. (1965–67; Eng. trans., *The Near East: The Early Civilizations,* 1967), with contributions by JEAN BOTTERO (Akkad), D.O. EDZARD (Early history, Ur III, and Old Babylonian Period), and A. FALKENSTEIN (Prehistory and Protohistory), which gives a balanced picture of political, social, and economic history; J. MELLAART, *The Neolithic of the Ancient Near East* (1975); ROBERT W. EHRICH (ed.), *Chronologies in Old World Archaeology* (1965), with contributions on Mesopotamia by R.H. DYSON, JR., E. PORADA, and P.J. WATSON; J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament,* 3rd ed. with suppl. (1969), a collection of important literary sources in English translation with bibliography; A. FALKENSTEIN, "La Cité-temple sumérienne," in *Cahiers d'histoire mondiale,* 1:784–814 (1954), a description and discussion of Sumerian temple economy and its political implications; I.M. DIAKONOFF (ed.), *Ancient, Mesopotamia: Socio-Economic History* (1969), English translation of representative articles by the editor and others on Mesopotamian history, with emphasis on social and economic aspects; W.W. HALLO, "Gutium-lexicon" article (English) in *Reallexikon der Assyriologie,* vol. 3, pp. 708–720 (1957–71), the best résumé available of the state of knowledge on the Gutians; D.O. EDZARD, *Die zweite Zwischenzeit Babyloniens* (1957), detailed monograph (the title is misleading) on the history of the Old Babylonian Period from Ur III to the end of Hammurabi; PAUL GARELLI, *Les Assyriens en Cappadoce* (1963), a standard work on the Old Assyrian trade colonies in Asia Minor; I.J. GELB, *Hurrians and Subareans* (1944), still the best book on the third cultural element in ancient Mesopotamian history.

Mesopotamia from 1600 BC to AD 630: See the relevant chapters of *The Cambridge Ancient History;* J. OATES, *Babylon* (1979); which deals with history and civilization; *Fischer Weltgeschichte* (*op. cit.*), an up-to-date paperback account by different authors; A. LEO OPPENHEIM, *Ancient Mesopotamia: Portrait of a Dead Civilization* (1964), including some controversial views; WOLFRAM VON SODEN, *Herrscher im alten Orient,* ch. 6–14 (1954), treats the more important kings; HARTMUT SCHMOKEL (ed.), *Kulturgeschichte des alten Orient* (1961), a good general account; and FIORELLA IMPARATI, *I Hurriti* (1964), a short synopsis in Italian. Standard works, now partly out of date, include A.T. OLMSTEAD, *History of Assyria* (1923, reprinted 1960); and BRUNO MEISSNER, *Babylonien und Assyrien,* 2 vol. (1920–25). J.A. BRINKMAN, *A Political History of Post-Kassite Babylonia, 1158–722 B.C.* (1968), is an extensive special study, with complete documentation.

Iraq from the Arab Conquest to 1918: There is no comprehensive history of medieval Iraq in English; some periods or towns form the subject of special works: GUY LE STRANGE, *Baghdad During the Abbasid Caliphate* (1900, reprinted 1924); REUBEN LEVY, *A Baghdad Chronicle* (1929); STEPHEN H. LONGRIGG, *Four Centuries of Modern Iraq* (1925). Travellers' accounts are of some importance: J.-B. TAVERNIER, *Les Six Voyages,* 2 vol. (1675; Eng. trans., 3 vol., 1677–78); EDWARD IVES, *Journey from Persia to England by an Unusual Route* (1773); KER PORTER, *Travels in Syria, Persia, Armenia, Ancient Babylonia,* 2 vol. (1817–20); JAMES S. BUCKINGHAM, *Travels in Mesopotamia* (1827); J.R. WELLSTED, *Travels to the City of the Caliphs . . . ,* 2 vol. (1840); HENRY SOUTHGATE, *Narrative of a Tour Through Armenia, Kurdistan, Persia and Mesopotamia,* 2 vol. (1840). For Iraq during World War I, see the ADMIRALTY NAVAL STAFF, *A Handbook of Mesopotamia,* 4 vol. (1916–17); and also *The Encyclopaedia of Islam,* new ed. (1960–). JACOB LASSNER, *The Shaping of ʿAbbāsid Rule* (1980), concentrates on the founding of Baghdad.

Iraq since 1918: The most comprehensive history of Iraq is STEPHEN H. LONGRIGG, *Iraq, 1900–1950* (1953). The period of the Mandate is adequately treated in PHILIP W. IRELAND, *Iraq: A Study in Political Development* (1937). A study of the political development of Iraq after independence is covered in considerable detail in MAJID KHADDURI, *Independent Iraq, 1932–1958: A Study in Iraqi Politics,* 2nd ed. (1960), *Republican Iraq: A Study in Iraqi Politics Since the Revolution of 1958* (1969), and *Socialist Iraq: A Study in Iraqi Politics Since 1968* (1978). For the economic development of Iraq, see the INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *The Economic Development of Iraq* (1952); KATHLEEN M. LANGLEY, *The Industrialization of Iraq* (1961); FAHIM QUBAIN, *The Reconstruction of Iraq, 1950–1957* (1958); and RONY E. GABBAY, *Communism and Agrarian Reform in Iraq* (1978). For a detailed reference book, see HANNA BATATU, *The Old Social Classes and the Revolutionary Movements of Iraq: A Study of Iraq's Old Landed and Commercial Classes and of its Communists, Baʿthists, and Free Officers* (1978). EDMUND GHAREEB, *The Kurdish Question in Iraq* (1981), details the government's dealings with the Kurds since 1968.

Ireland

The republic of Ireland, also known by its Irish name of Éire, occupies the greater part of an island lying to the west of Great Britain, from which it is separated—at distances ranging from 11 to 120 miles (from 18 to 193 kilometres)—by the Irish Sea. Located in the temperate zone between latitude 51°30′ and 55°30′ N, and longitude 6°00′ and 10°30′ W—as far north as Labrador or British Columbia in Canada, and as far west as Portugal, or the West African state of Liberia—it constitutes the westernmost outpost of the Atlantic fringe of the Eurasian landmass. Ireland, which, like Britain, once formed part of this landmass, lies on the European continental shelf, surrounded by seas that are generally less than 650 feet (200 metres) in depth. The greatest distance from north to south in the island is 302 miles (486 kilometres), and from east to west it is 171 miles (275 kilometres). The area of the 26 counties of the Republic of Ireland is 27,137 square miles (70,285 square kilometres). The capital is Dublin.

The magnificent scenery of Ireland's Atlantic coastline faces a 2,000-mile-wide expanse of ocean, and its geographical isolation has helped it to develop a rich heritage of culture and tradition linked initially to a separate language. By the late 20th century the perennial concerns of the republic—emigration, cultural and political identity, relations with Northern Ireland (that portion of the island remaining within the United Kingdom)—had been augmented by economic and political problems linked both to a deterioration in the international competitive position of the nation and to its entry in 1973 into the European Economic Community (EEC), or Common Market. The discussions and political developments discernible within the nation as a result of these difficulties evoked a high level of public interest not unaccompanied by a measure of emotion and conflict.

Ireland was an integral part of the United Kingdom from 1800 to 1922 when, by virtue of the Anglo-Irish agreement of Dec. 6, 1921, the Irish Free State was established as an independent member of the British Commonwealth. A new constitution, adopted by the people in a 1937 plebiscite, declared Ireland to be "a sovereign, independent, democratic state," and the last remaining link with the Commonwealth was severed by the Republic of Ireland Act of 1948.

Of the four traditional provinces of Ireland, the republic comprises the whole of Leinster (counties Carlow, Dublin, Kildare, Kilkenny, Laois [Leix], Louth, Longford, Meath, Offaly, Westmeath, Wexford, and Wicklow), the whole of Munster (counties Clare, Cork, Kerry, Waterford, Limerick, and Tipperary), the whole of Connacht (counties Galway, Leitrim, Mayo, Roscommon, and Sligo) and three counties (Cavan, Donegal, and Monaghan) of the nine counties of Ulster. The remaining six counties of Ulster constitute Northern Ireland. For information on the geography and history of Northern Ireland, see UNITED KINGDOM: *Northern Ireland.*

This article is divided into the following sections:

Physical and human geography 951
- The land 951
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 955
- The economy 955
  - Resources
  - Agriculture
  - Industry
  - Trade unions
  - Finance
  - Transportation
- Administration and social conditions 956
  - Government and justice
  - Armed forces
  - Education
  - Health and welfare services
- Cultural life 958

History 958
- Early Ireland 958
  - Neolithic Age
  - Bronze Age
  - Iron Age
  - Early Celtic Ireland
  - Early Christianity
  - The Norse invasions and their aftermath
- First centuries of English rule (*c.* 1166–*c.* 1600) 961
  - The Anglo-Norman invasion
  - The 14th and 15th centuries
  - The Kildare ascendancy
  - The Reformation period
  - Ireland under Elizabeth I
- Modern Ireland under British rule 963
  - The 17th century
  - The 18th century
  - Social, economic, and cultural life in the 17th and 18th centuries
  - The 19th and early 20th centuries
- Independent Ireland 967
  - Establishment of the Irish Free State
  - The Cosgrave ministries
  - De Valera's governments (1932–48) and the establishment of Éire
  - The Republic of Ireland

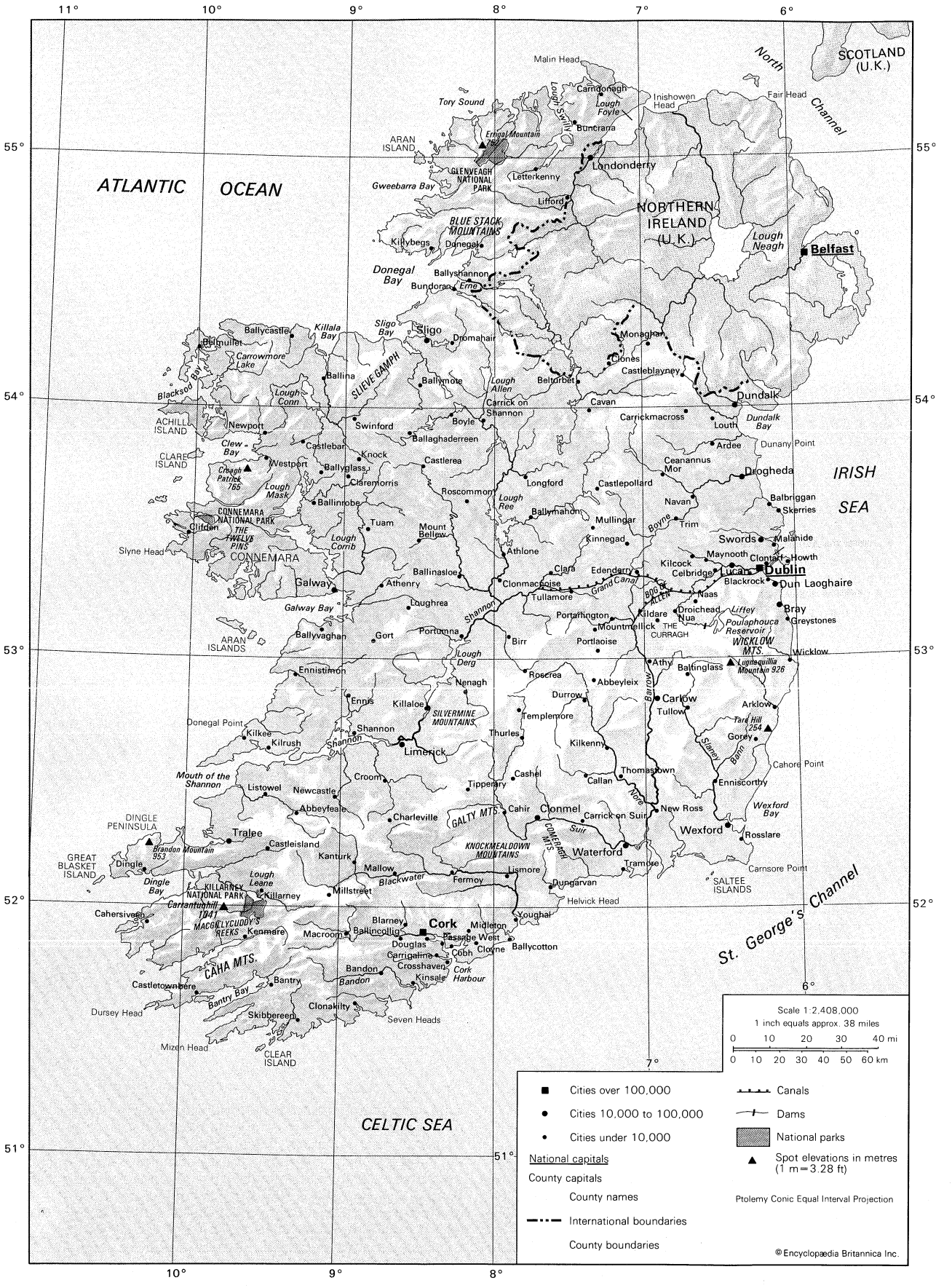
Bibliography 970

Physical and human geography

THE LAND

Relief. The territory of the republic consists of a broad and undulating central limestone plain, ringed almost completely by coastal highlands, which vary considerably in geological structure. The flatness of the central lowland—which lies for the most part between 200 and 400 feet above sea level—is relieved in many places by low hills between 600 and 1,000 feet high, while a multiplicity of lakes, large bog areas, and low ridges lend it considerable scenic attractions. The principal mountain ranges are the Wicklow Mountains (Lugnaquillia 3,039 feet [926 metres]), the Macgillycuddy's Reeks in Kerry (Carrantuohill 3,414 feet [1,041 metres]), the Knockmealdown and Comeragh mountains in Waterford, and The Twelve Pins in Connemara. In the west and southwest the wild and beautiful coast is heavily indented where the mountains of Donegal, Mayo, and Kerry thrust out into the Atlantic, separated by deep, wide-mouthed bays, some of which—Bantry Bay and Dingle Bay, for example—are, in fact, drowned river valleys. Bantry Bay, which is 20 miles long and between three and five miles wide, is one of the principal deep-sea anchorages in western Europe. The oil terminal at Whiddy Island in the bay, although capable of handling the largest tankers afloat, was closed following the explosion of a tanker there in 1979. The east coast, on the other hand, is little indented, but most of the country's trade passes through its ports because of their proximity to British and continental markets.

The western coastline


ATLANTIC OCEAN
CELTIC SEA
IRISH SEA
St. George's Channel
North Channel
SCOTLAND (U.K.)
NORTHERN IRELAND (U.K.)
Malin Head
Tory Sound
Carndonagh
Lough Foyle
Inishowen Head
Fair Head
Buncrana
Lough Swilly
Errigal Mountain 752
ARAN ISLAND
GLENVEAGH NATIONAL PARK
Gweebarra Bay
Letterkenny
Londonderry
Lifford
BLUE STACK MOUNTAINS
Killybegs
Donegal
Donegal Bay
Ballyshannon
Bundoran
Erne
Lough Neagh
Belfast
Ballycastle
Killala Bay
Sligo Bay
Sligo
Dromahair
Monaghan
Clones
Castleblayney
Belmullet
Carrowmore Lake
Blacksod Bay
Ballina
SLIEVE GAMPH
Ballymote
Lough Allen
Belturbet
Dundalk
Lough Conn
Carrick on Shannon
Cavan
Carrickmacross
Dundalk Bay
ACHILL ISLAND
Newport
Swinford
Boyle
Louth
Clew Bay
Castlebar
Ballaghaderreen
Ardee
Dunany Point
CLARE ISLAND
Knock
Castlerea
Ceanannus Mor
Croagh Patrick 765
Westport
Ballyglass
Claremorris
Longford
Castlepollard
Drogheda
Lough Mask
Roscommon
Lough Ree
Balbriggan
Skerries
Ballinrobe
Ballymahon
Navan
CONNEMARA NATIONAL PARK
THE TWELVE PINS
Clifden
Tuam
Mount Bellew
Mullingar
Boyne
Trim
Slyne Head
CONNEMARA
Lough Corrib
Athlone
Kinnegad
Swords
Malahide
Maynooth
Clontarf
Howth
Kilcock
Celbridge
Lucan
Dublin
Ballinasloe
Clara
Edenderry
Galway
Athenry
Clonmacnoise
Grand Canal
Blackrock
Dun Laoghaire
Naas
Tullamore
Galway Bay
Loughrea
Shannon
Portarlington
Kildare
Droichead Nua
Liffey
Bray
Mountmellick
THE CURRAGH
Poulaphouca Reservoir
Greystones
ARAN ISLANDS
Ballyvaghan
Gort
Portumna
Birr
Portlaoise
WICKLOW MTS.
Wicklow
Lough Derg
Athy
Baltinglass
Lugnaquillia Mountain 926
Ennistimon
Roscrea
Abbeyleix
Nenagh
Barrow
Durrow
Carlow
Ennis
Killaloe
SILVERMINE MOUNTAINS
Templemore
Tullow
Arklow
Donegal Point
Kilkee
Shannon
Thurles
Tara Hill 254
Kilrush
Limerick
Kilkenny
Gorey
Slaney
Bann
Cahore Point
Mouth of the Shannon
Croom
Cashel
Thomastown
Listowel
Newcastle
Tipperary
Callan
Enniscorthy
Abbeyfeale
Charleville
GALTY MTS.
Cahir
Clonmel
Carrick on Suir
New Ross
Wexford Bay
DINGLE PENINSULA
Tralee
Castleisland
Brandon Mountain 953
Kanturk
KNOCKMEALDOWN MOUNTAINS
COMEERAGH MTS.
Suir
Waterford
Wexford
Rosslare
GREAT BLASKET ISLAND
Dingle
Mallow
Blackwater
Fermoy
Lismore
Tramore
Carnsore Point
Dingle Bay
Lough Leane
KILLARNEY NATIONAL PARK
Killarney
Millstreet
Dungarvan
SALTEE ISLANDS
Carrantuohill 1041
Helvick Head
Cahersiveen
MACGILLYCUDDY'S REEKS
Blarney
Cork
Youghal
Kenmare
Macroom
Ballincollig
Midleton
Passage West
Douglas
Cobh
Cloyne
Ballycotton
Carrigaline
Crosshaven
Cork Harbour
CAHA MTS.
Bandon
Kinsale
Castletownbere
Bantry
Bantry Bay
Dursey Head
Clonakilty
Skibbereen
Seven Heads
Mizen Head
CLEAR ISLAND
Scale 1:2,408,000
1 inch equals approx. 38 miles
0 10 20 30 40 mi
0 10 20 30 40 50 60 km
Cities over 100,000
Cities 10,000 to 100,000
Cities under 10,000
National capitals
County capitals
County names
International boundaries
County boundaries
Canals
Dams
National parks
Spot elevations in metres (1 m=3.28 ft)
Ptolemy Conic Equal Interval Projection
©Encyclopædia Britannica Inc.

MAP INDEX

Political subdivisions

Cabhan, An, see Cavan
Carlow ..... 52 40 N 6 50 W
Cavan (Cabhan, An) ..... 53 55 N 7 15 W
Ciarrai, see Kerry
Cill Chainnigh, see Kilkenny
Cill Dara, see Kildare
Cill Mhantain, see Wicklow
Clare (Clar) ..... 52 50 N 9 00 W
Cork ..... 51 58 N 8 35 W
Donegal (Dun na nGall) ..... 54 55 N 8 00 W
Dublin-Belgrad .. 53 16 N 6 26 W
Dublin County Borough ..... 53 21 N 6 15 W
Dublin-Fingal .... 53 31 N 6 16 W
Dun Laoghaire-Rathdown ..... 53 15 N 6 13 W
Dun na nGall, see Donegal
Galway (Gaillimh) . 53 20 N 9 00 W
Iarmhi, An, see Westmeath
Kerry (Ciarrai) ... 52 10 N 9 45 W
Kildare (Cill Dara) . 53 10 N 6 45 W
Kilkenny (Cill Chainnigh) ..... 52 35 N 7 15 W
Laois (Laoighis, Leix) ..... 53 00 N 7 24 W
Leitrim (Liatroim) . 54 07 N 8 00 W
Leix, see Laois
Liatroim, see Leitrim
Limerick (Luimneach) ... 52 30 N 8 45 W
Loch Garman, see Wexford
Longford ..... 53 40 N 7 45 W
Louth (Lu) ..... 53 50 N 6 30 W
Luimneach, see Limerick
Mayo (Maigh Eo) . 53 54 N 9 15 W
Meath (An Mhi, Mide) ..... 53 40 N 6 40 W
Monaghan ..... 54 15 N 7 00 W
Offaly (Uibh Fhaili) ..... 53 15 N 7 30 W
Port Lairge, see Waterford
Roscommon .... 53 45 N 8 15 W
Sligo ..... 54 15 N 8 40 W
Tipperary North Riding ..... 52 45 N 8 00 W
Tipperary South Riding ..... 52 25 N 7 50 W
Uibh Fhaili, see Offaly
Waterford (Port Lairge) ..... 52 15 N 7 30 W
Westmeath (Iarmhi, An) .... 53 30 N 7 30 W
Wexford (Loch Garman) ..... 52 30 N 6 40 W
Wicklow (Cill Mhantain) ..... 53 00 N 6 25 W

Cities and towns

Abbeyfeale ..... 52 23 N 9 18 W
Abbeyleix ..... 52 54 N 7 21 W
Ardee ..... 53 51 N 6 32 W
Arklow (tInbhear Mor, An) ..... 52 48 N 6 09 W
Athenry ..... 53 18 N 8 45 W
Athlone ..... 53 26 N 7 57 W
Athy ..... 52 59 N 6 59 W
Balbriggan ..... 53 36 N 6 11 W
Ballaghaderreen . 53 54 N 8 35 W
Ballina ..... 54 07 N 9 10 W
Ballinasloe ..... 53 20 N 8 13 W
Ballincollig ..... 51 53 N 8 35 W
Ballinrobe ..... 53 38 N 9 14 W
Ballycastle ..... 54 17 N 9 22 W
Ballycotton ..... 51 50 N 8 01 W
Ballyglass ..... 53 44 N 9 10 W
Ballymahon ..... 53 34 N 7 46 W
Ballymote ..... 54 05 N 8 31 W
Ballyshannon .... 54 30 N 8 11 W
Ballyvaghan .... 53 07 N 9 09 W
Baltinglass ..... 52 56 N 6 42 W
Bandon (Droichead na Bandan) ... 51 45 N 8 44 W
Bantry ..... 51 41 N 9 27 W
Belmullet ..... 54 13 N 10 00 W
Belturbet ..... 54 06 N 7 27 W
Binn Eadair, see Howth
Birr (Biorra) ..... 53 06 N 7 54 W
Blackrock (Carraig Dubh) . 53 18 N 6 10 W
Blarney ..... 51 56 N 8 34 W
Boyle ..... 53 58 N 8 18 W
Bray (Bre) ..... 53 12 N 6 06 W
Buncrana ..... 55 08 N 7 27 W
Bundoran ..... 54 28 N 8 17 W
Cabhan, An, see Cavan
Cahersiveen .... 51 58 N 10 13 W
Cahir ..... 52 22 N 7 55 W
Callan ..... 52 32 N 7 23 W
Carlow (Ceatharlach) ... 52 50 N 6 56 W
Carndonagh .... 55 15 N 7 16 W
Carraig Dubh, see Blackrock
Carraig na Siuire, see Carrick on Suir
Carrick on Shannon ..... 53 57 N 8 05 W
Carrick on Suir (Carraig na Siuire) ..... 52 21 N 7 25 W
Carrickmacross .. 53 58 N 6 43 W
Carrigaline ..... 51 48 N 8 23 W
Cashel ..... 52 31 N 7 53 W
Castlebar ..... 53 51 N 9 18 W
Castleblayney ... 54 07 N 6 44 W
Castleisland .... 52 14 N 9 28 W
Castlepollard .... 53 40 N 7 18 W
Castlerea ..... 53 46 N 8 30 W
Castletownbere .. 51 39 N 9 55 W
Cavan (Cabhan, An) ..... 54 00 N 7 22 W
Ceanannus Mor (Kells) ..... 53 43 N 6 52 W
Ceatharlach, see Carlow
Celbridge ..... 53 20 N 6 33 W
Charleville (Rath Luirc) ..... 52 21 N 8 41 W
Cill Airne, see Killarney
Cill Chainnigh, see Kilkenny
Cill Dalua, see Killaloe
Cill Dara, see Kildare
Cill Mhantain, see Wicklow
Cionn tSaile, see Kinsale
Clara ..... 53 20 N 7 37 W
Claremorris ..... 53 43 N 9 00 W
Clifden ..... 53 29 N 10 01 W
Clonakilty ..... 51 37 N 8 53 W
Clones ..... 54 11 N 7 14 W
Clonmacnoise (Cluain Mhic Nois) ..... 53 20 N 7 59 W
Clonmel (Cluain Meala) ..... 52 21 N 7 42 W
Clontarf ..... 53 21 N 6 10 W
Cloyne ..... 51 51 N 8 07 W
Cluain Meala, see Clonmel
Cluain Mhic Nois, see Clonmacnoise
Cobh ..... 51 51 N 8 17 W
Cork (Corcaigh) .. 51 54 N 8 28 W
Croom ..... 52 31 N 8 43 W
Crosshaven ..... 51 48 N 8 18 W
Dingle ..... 52 08 N 10 15 W
Donegal ..... 54 39 N 8 07 W
Douglas ..... 51 52 N 8 26 W
Drogheda (Droichead Atha). 53 43 N 6 21 W
Droichead na Bandan, see Bandon
Droichead Nua .. 53 11 N 6 48 W
Dromahair ..... 54 14 N 8 18 W
Dublin ..... 53 20 N 6 15 W
Dun Laoghaire .. 53 17 N 6 07 W
Dundalk (Dun Dealgan) ..... 54 00 N 6 25 W
Dungarvan ..... 52 05 N 7 37 W
Durlas, see Thurles
Durrow ..... 52 50 N 7 24 W
Edenderry ..... 53 20 N 7 03 W
Ennis (Inis) ..... 52 51 N 8 59 W
Enniscorthy ..... 52 30 N 6 34 W
Ennistimon ..... 52 56 N 9 18 W
Fermoy ..... 52 08 N 8 17 W
Galway (Gaillimh) . 53 17 N 9 03 W
Gorey ..... 52 40 N 6 18 W
Gort ..... 53 04 N 8 49 W
Greystones ..... 53 09 N 6 04 W
Howth (Binn Eadair) ..... 53 23 N 6 04 W
Inis, see Ennis
Kanturk ..... 52 10 N 8 54 W
Kells, see Ceanannus Mor
Kenmare ..... 51 53 N 9 35 W
Kilcock ..... 53 24 N 6 40 W
Kildare (Cill Dara) . 53 10 N 6 55 W
Kilkee ..... 52 41 N 9 38 W
Kilkenny (Cill Chainnigh) ..... 52 39 N 7 15 W
Killaloe (Cill Dalua) ..... 52 48 N 8 27 W
Killarney (Cill Airne) ..... 52 03 N 9 31 W
Killybegs ..... 54 38 N 8 27 W
Kilrush ..... 52 38 N 9 29 W
Kinnegad ..... 53 27 N 7 06 W
Kinsale (Cionn tSaile) ..... 51 42 N 8 31 W
Knock ..... 53 47 N 8 55 W
Letterkenny ..... 54 57 N 7 44 W
Lifford ..... 54 50 N 7 29 W
Limerick (Luimneach) ... 52 40 N 8 37 W
Lismore (Lios Mor) ..... 52 08 N 7 55 W
Listowel ..... 52 27 N 9 29 W
Loch Garman, see Wexford
Longford ..... 53 44 N 7 48 W
Loughrea ..... 53 12 N 8 34 W
Louth (Lu) ..... 53 57 N 6 32 W
Lucan ..... 53 21 N 6 27 W
Luimneach, see Limerick
Macroom ..... 51 54 N 8 57 W
Malahide ..... 53 27 N 6 09 W
Mallow ..... 52 08 N 8 38 W
Maryborough, see Portlaoise
Maynooth ..... 53 23 N 6 36 W
Midleton ..... 51 55 N 8 10 W
Millstreet ..... 52 04 N 9 04 W
Monaghan ..... 54 15 N 6 58 W
Mount Bellew ... 53 28 N 8 30 W
Mountmellick .... 53 07 N 7 19 W
Mullingar ..... 53 32 N 7 21 W
Naas (Nas, An) .. 53 13 N 6 40 W
Navan (Uaimh, An) ..... 53 39 N 6 40 W
Nenagh ..... 52 52 N 8 12 W
New Ross (Ros Mhic Thriuin) ... 52 23 N 6 56 W
Newcastle ..... 52 27 N 9 03 W
Newport ..... 53 53 N 9 33 W
Passage West ... 51 52 N 8 20 W
Port Lairge, see Waterford
Portarlington .... 53 09 N 7 11 W
Portlaoise (Maryborough, Portlaoighise) .. 53 02 N 7 18 W
Portumna ..... 53 05 N 8 13 W
Rath Luirc, see Charleville
Ros Mhic Thriuin, see New Ross
Roscommon .... 53 38 N 8 11 W
Roscrea ..... 52 57 N 7 48 W
Rosslare ..... 52 17 N 6 23 W
Shannon ..... 52 42 N 8 52 W
Skerries ..... 53 35 N 6 07 W
Skibbereen ..... 51 33 N 9 16 W
Sligo ..... 54 16 N 8 29 W
Swinford ..... 53 57 N 8 57 W
Swords ..... 53 27 N 6 13 W
Templemore .... 52 47 N 7 50 W
Thomastown .... 52 31 N 7 08 W
Thurles (Durlas) .. 52 41 N 7 48 W
Tinbhear Mor, An, see Arklow
Tipperary (Tiobraid Arann) . 52 28 N 8 10 W
Tralee ..... 52 16 N 9 43 W
Tramore ..... 52 09 N 7 09 W
Trim ..... 53 33 N 6 48 W
Tuam ..... 53 31 N 8 51 W
Tullamore ..... 53 16 N 7 29 W
Tullow ..... 52 48 N 6 44 W
Uaimh, An, see Navan
Waterford (Port Lairge) ..... 52 15 N 7 06 W
Westport ..... 53 48 N 9 31 W
Wexford (Loch Garman) ..... 52 20 N 6 28 W
Wicklow (Cill Mhantain) ..... 52 59 N 6 03 W
Youghal ..... 51 57 N 7 51 W

Physical features and points of interest

Achill Island .... 53 57 N 10 00 W
Allen, Bog of (Moin Alieine) .. 53 18 N 6 55 W
Allen, Lough .... 54 07 N 8 03 W
Aran Island ..... 54 59 N 8 32 W
Aran Islands (Arainn) ..... 53 06 N 9 40 W
Atlantic Ocean .. 55 00 N 10 00 W
Bandon, *river* ... 51 43 N 8 37 W
Bann, *river* ..... 52 32 N 6 33 W
Bantry Bay ..... 51 38 N 9 42 W
Barrow, *river* .... 52 14 N 6 58 W
Blacksod Bay ... 54 05 N 10 01 W
Blackwater, *river* . 51 59 N 7 51 W
Blue Stack Mountains ..... 54 45 N 8 06 W
Boyne, *river* ..... 53 43 N 6 15 W
Brandon Mountain ..... 52 14 N 10 15 W
Caha Mountains . 51 45 N 9 43 W
Cahore Point .... 52 33 N 6 12 W
Carnsore Point .. 52 10 N 6 22 W
Carrantuohill (Carrauntoohill), *peak* ..... 52 00 N 9 45 W
Carrowmore, Lake ..... 54 12 N 9 48 W
Celtic Sea ..... 51 00 N 8 30 W
Clare Island ..... 53 48 N 10 00 W
Clear Island ..... 51 26 N 9 30 W
Clew Bay ..... 53 50 N 9 48 W
Comeragh Mountains ..... 52 17 N 7 34 W
Conn, Lough .... 54 00 N 9 15 W
Connemara, *region* ..... 53 22 N 9 40 W
Connemara National Park ... 53 32 N 9 55 W
Cork Harbour ... 51 51 N 8 14 W
Corrib, Lough ... 53 26 N 9 14 W
Croagh Patrick (Cruach Phadraig), *peak* . 53 45 N 9 39 W
Curragh, The, *plain* ..... 53 08 N 6 50 W
Derg, Lough (Deirgeirt, Loch) . 52 57 N 8 19 W
Dingle Bay ..... 52 05 N 10 14 W
Dingle Peninsula . 52 10 N 10 10 W
Donegal Bay .... 54 32 N 8 30 W
Donegal Point ... 52 43 N 9 38 W
Dunany Point ... 53 51 N 6 15 W
Dundalk Bay .... 53 57 N 6 15 W
Dursey Head, *cape* ..... 51 35 N 10 15 W
Erne, *river* ..... 54 30 N 8 16 W
Errigal Mountain . 55 02 N 8 07 W
Foyle, Lough (Feabhail, Loch), *bay* ..... 55 06 N 7 08 W
Galty Mountains . 52 23 N 8 07 W
Galway Bay .... 53 10 N 9 10 W
Gamph, Slieve, *mountains* ..... 54 07 N 8 54 W

Glenveagh National Park ....55 00 N 8 00 W
Grand Canal .....53 20 N 8 13 W
Great Blasket Island .....52 06 N 10 32 W
Gweebarra Bay .....54 52 N 8 27 W
Helvick Head, *cape* .....52 03 N 7 32 W
Inishowen Head, *cape* .....55 14 N 6 55 W
Irish Sea .....53 40 N 5 30 W
Killala Bay .....54 15 N 9 08 W
Killarney National Park .....52 00 N 9 35 W
Knockmealdown Mountains .....52 14 N 7 55 W
Leane, Lough ....52 02 N 9 33 W
Liffey, *river* .....53 21 N 6 14 W
Lugnaquillia Mountain .....52 58 N 6 27 W
Macgillycuddy's Reeks, *mountains* .....52 00 N 9 42 W
Malin Head, *cape* .55 22 N 7 24 W
Mask, Lough ....53 37 N 9 21 W
Mizen Head, *cape* .....51 27 N 9 49 W
Moin Alieine, see Allen, Bog of
Nore, *river* .....52 25 N 6 57 W
Poulaphouca Reservoir .....53 08 N 6 30 W
Ree, Lough .....53 32 N 7 58 W
St. George's Channel .....52 00 N 6 00 W
Saltee Islands ....52 07 N 6 36 W
Seven Heads, *cape* .....51 35 N 8 43 W
Shannon, *river* ....52 35 N 9 41 W
Shannon, Mouth of the, *bay* ....52 30 N 9 50 W
Silvermine Mountains (Silvermines) ....52 46 N 8 15 W
Slaney, *river* .....52 20 N 6 27 W
Sligo Bay .....54 18 N 8 42 W
Slyne Head, *cape* .....53 24 N 10 12 W
Suir, *river* .....52 17 N 7 00 W
Swilly, Lough, *bay* .....55 10 N 7 32 W
Tara Hill .....52 41 N 6 13 W
Tory Sound .....55 13 N 8 12 W
Twelve Pins, The, *mountains* ..53 30 N 9 50 W
Wexford Bay .....52 17 N 6 22 W
Wicklow Mountains .....53 00 N 6 27 W

The coastal mountain fringe illustrates the country's complex geological history. In the west and northwest and in the east, the mountains are mainly composed of granite. Old Red Sandstone predominates in the south, where the mountain ridges are parallel east–west foldings separated by limestone river valleys. Ireland experienced two general glaciations, one covering most of the country and the other extending as far south as a line linking Limerick, Cashel, and Dublin. The characteristic diversity of Irish scenery owes much to the glacial influence, and the large areas of peat bog to be found throughout the country are a notable feature of the landscape. In the past they were the rural population's principal—if not its only—source of domestic fuel. By the 1980s, virtually all rural households were connected to the national electricity network, and peat, or turf, production had become mechanized and industrialized.

Drainage. The rivers that rise on the seaward side of the coastal mountain fringe are naturally short and rapid. The inland streams, however, flow slowly, often through marshes and lakes, and enter the sea—usually by way of waterfalls and rapids—long distances from their sources. The famed Shannon, for example, rises in the plateau country near Sligo Bay and flows sluggishly south-southwestward for 161 miles, reaching tidewater level at Limerick and draining a wide area of the central lowland on its way. Other major inland streams—some of them famous for their salmon fisheries—are the Slaney, Liffey, and Boyne in the east; the Nore, Barrow, and Suir in the southeast; the Blackwater, Lee, and Bandon in the south; and the Clare and the Moy in the west. Because of the porosity of the underlying Carboniferous limestones, an underground drainage system has developed, feeding the interlacing surface network of rivers and lakes. The state has implemented major arterial drainage projects, preventing flooding—and making more land available for cultivation—by improving the flow of water in the rivers and thereby lowering the levels of lakes. There are also state-aided farm drainage schemes designed to bring wasteland and marginal land into production.

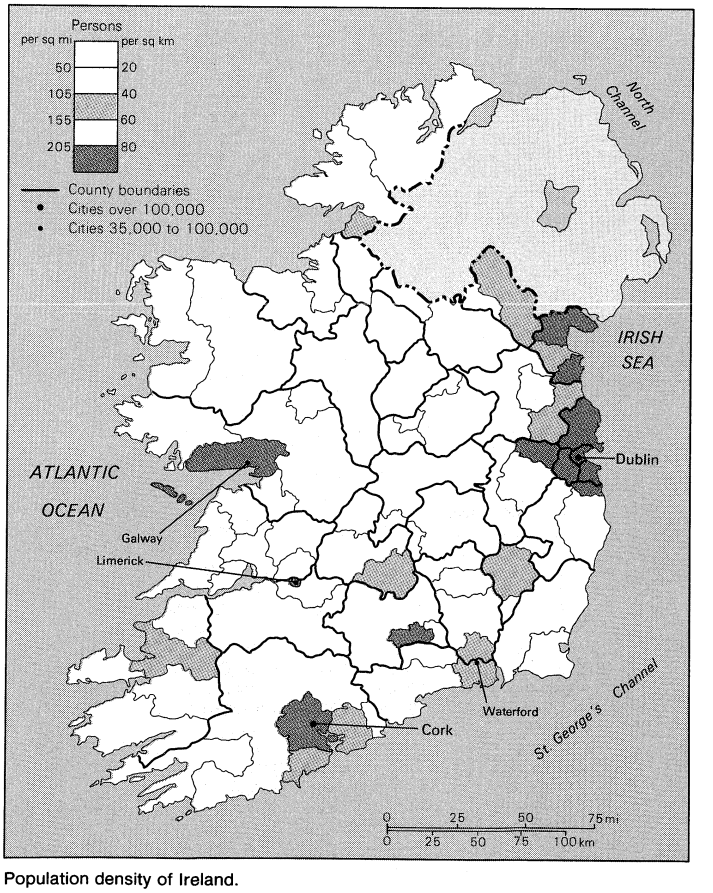
Drainage improvement

Soils. Most Irish soils originate from drift, the ice-scoured waste formerly frozen to the base of the advancing glaciers. Some of the older rocks in the country's geological formation—quartzites, certain granites, and shales—weather into infertile and unproductive soils. In many places, however, these have been overlaid by patches of the ice-borne drift, mostly limestone-bearing, which are farmed with considerable success. The bare limestone regions remaining in western areas are a reminder of how much glacial drift cover has meant to the Irish agricultural economy.

Climate. Ireland's climate is of the type known to geographers as western maritime. The predominant influence is the Atlantic Ocean, which is no more than 70 miles from any inland location. The mild southwesterly winds and warm Gulf Stream waters contribute to the moderate quality of the climate. Temperature is almost uniform over the entire island. Average air temperatures lie mainly between limits of 39° and 45° F (4° and 7° C) in January and February, the coldest months of the year. In July and August, the warmest months, temperatures usually range between 57° and 61° F (14° and 16° C), although occasionally considerably higher readings are recorded. The sunniest months are May and June, when there is sunshine for an average duration of 5.5 and 6.5 hours a day, respectively, over most of the country, and the ancient patchwork of fields and settlements making up the landscape glows under a clear, vital light. Average annual rainfall varies from about 30 inches (760 millimetres) in the east, to more than 100 inches in the western areas exposed to the darkening clouds that often come scudding in from the Atlantic. The rainfall, combined with the equable climate, is particularly beneficial to the grasslands, which are the mainstay of the country's large livestock population. Snow is infrequent except in the mountains, and prolonged or severe snowstorms are rare.

Plant and animal life. Ireland was almost completely covered by the ice cap during one of the two glaciations it experienced, and its plant and animal life are consequently mainly—but not quite entirely—the result of the migration of species from other areas. As long as there was a land connection between Ireland and what was to become the rest of the British Isles, most species arrived overland from northern Europe. Irish plant and animal life nevertheless possesses certain unique features due partly to climatic conditions and partly to the fact that Ireland became separated from Britain by the Irish Sea some time before Britain itself became separated from the European continent. Common English animals such as the weasel and the mole do not exist in Ireland, which also has no snakes; tradition ascribes their absence to banishment at the hands of St. Patrick, the island's patron saint. In addition, there are only two kinds of mice—as against four in Britain—while the only reptile found in Ireland is a

Migration of species



Population density of Ireland.

species of lizard. Apart from flora imported from northern Europe, several plants common in Ireland are believed to have reached the country from the Mediterranean, along a subsequently drowned coastal route, and others appear to have migrated from North America, probably by way of Greenland and Iceland.

Settlement patterns. The smallness of the country makes for homogeneity and helps explain the nation's distinctive character; it also works against the development of significant regional or local divisions. One important regional distinction is, however, that between the part of the country east of the Shannon—with its expanding industrial employment, fertile farmlands, economic growth, and rising standard of living—and the poorer areas of the west—particularly west Donegal, Leitrim, west Mayo, west Galway, Clare, west Cork, and south Kerry—where incomes are among the lowest in the country, and the fertility of the land is in many cases insufficient to provide an acceptable standard of living for the people. These western areas include the districts, known collectively as the Gaeltacht, in which the Irish language and the traditional national culture are best preserved. Heavy emigration, however, constitutes a threat to the survival of this cultural heritage.

THE PEOPLE

Although Ireland has been invaded and colonized within historic times by Celts, Norsemen, Normans, English, and Scots, no racial or ethnic distinctions exist in the republic today. A large majority of the republic's population is Roman Catholic, with other religious groups (including Church of Ireland Episcopalians, Presbyterians, Methodists, and Jews) very much in a minority. There is no established church in Ireland; freedom of conscience and the free practice of religion are guaranteed by the constitution.

The Irish language

The constitution provides that Irish, as the national language, is the first official language; it recognizes English as the second official language. All official documents are published in both Irish and English. The Irish language, which is very similar to Scots Gaelic, was widely spoken up to the time of the Great Famine of the 1840s and the subsequent emigrations. From then on, its use declined until 1922, when the teaching of Irish was introduced into all schools. Although its use as a vernacular has decreased and is restricted to small Gaeltacht (*i.e.,* Irish-speaking) areas, it is certainly more widely read, spoken, and understood today than at any previous time in the 20th century. English is also taught in all the schools and is universally spoken.

The republic's marriage, birth, and death rates are broadly comparable with those of nearby countries, but the rate of emigration historically has been greatly in excess of the next highest rate in Europe. As a result of emigration, the number of Irish-born people living outside their native land is reckoned to total about 50 percent of the population of the republic.

Emigration

THE ECONOMY

The republic has a mixed economy. The constitution provides that the state shall favour private initiative in industry and commerce, but when the necessary private initiative is not forthcoming, the state itself undertakes essential services and promotes development projects. Thus, state corporations operate the country's air services, its rail and road transport, its television and radio stations, its electricity generation and distribution system, its sugar industry, its peat industry, its petroleum refinery, and other major national undertakings. State companies are also active in the fields of agricultural and industrial credit, food processing, fertilizers, hotels, and life insurance. In some cases, these companies are in direct competition with private enterprise.

State and private enterprise

Economic relations between the republic and the United Kingdom continue to be close. About half of the republic's imports come from Great Britain and Northern Ireland, and tourism and business relations are facilitated by the absence of a passport barrier between the two countries. Since the republic joined the EEC in 1973, the proportion of Irish exports destined for EEC countries other than the United Kingdom has increased, though Britain remains a principal export market. The Irish pound, or *punt,* which had been linked to the British pound sterling, was replaced by a separate currency in 1979 when the republic joined the European Monetary System. Ireland is also a member of the European Investment Bank.

Resources. The country is not rich in mineral resources. Discoveries of silver, lead, zinc, gypsum, and barite have been successfully developed, but the country's dependence on imports for its energy requirements is high. In the early 1980s offshore oil and natural gas wells began production in the Celtic Sea south of County Cork.

Agriculture. Agriculture, once the mainstay of the national economy, continues to be important. Most of Ireland's agricultural land is used as pasture or for growing hay. The country's climate produces abundant vegetable and other plant growth and is particularly beneficial to the rich grasslands that enable grazing stock to be kept on pasture almost the year round, only a limited amount of animal housing being required in winter. Most farms are

Porterfield-Chickering—Photo Researchers, Inc.



Blackfaced sheep on a small farm on the Atlantic coast of County Kerry, Ireland.

The family farm

family farms; only a small percentage of those employed in agriculture work as hired labour. Mixed farming is the general pattern, with the production of beef cattle tending to predominate in the midlands and dairy farming in the south. Cereal growing is an important activity in the east and southeast. Sheep raising is widespread on the rugged hills and mountain slopes throughout the country and particularly in the western area.

Most of the gross agricultural output consists of livestock and livestock products, cattle being the biggest single item, followed by milk and pigs. Other important products are cereals (particularly barley and wheat), poultry and eggs, sheep and wool, and root crops, including potatoes. Enough beets are grown to meet the country's sugar requirements. The bloodstock (Thoroughbred) industry is a thriving sector of the agricultural economy and has won worldwide fame for the Irish horse. Adverse conditions in export markets following World War II handicapped the expansion of Irish agriculture, and the subsequent growth of agricultural output was notably slower than that in the industrial and service sectors. This situation was ameliorated with the republic's entrance into the EEC in 1973. In the late 1970s, however, farm incomes began to decline again.

When the independent nation was founded in 1922, woodland represented less than 1 percent of the total land area, but state replanting since World War II has increased almost fivefold the acreage under forests and woodland. The sea-fishing industry, on the other hand, remains relatively underdeveloped. Judged by western European standards, home consumption of fish, particularly shellfish, is remarkably low.

Industry. Until World War II and for some years after it, official policy on manufacturing was nationalistic and protectionist. High tariffs and quotas protected young industries, which provided badly needed employment and helped to supply the home market, but which had little or no export potential. From the middle 1950s onward, the protectionist policy was progressively reversed. The principal basis of the government's Programmes for Economic Expansion was an industrial development policy designed, by means of tax concessions, financial grants, and other incentives, first, to encourage existing industries to increase their competitive strength and seek markets abroad and, second, to attract new manufacturing enterprises, whether foreign or Irish-owned, to the republic.

Protectionist and expansionist policies

The policy achieved a large measure of success. By the late 20th century a larger proportion of the labour force was employed in manufacturing than in agriculture, and the industrial sector accounted for most of Ireland's total export earnings. A competitive economy has become all the more desirable in view of the governmental obligation to demobilize protective tariffs in accordance with the Anglo-Irish Free Trade Agreement, to say nothing of the effect of Ireland's entry into the EEC. A wide range of manufactured products is exported, including electrical goods and apparatus, processed foods, beer and spirits, clothing, footwear, machinery, metal ores and concentrates, glassware, pharmaceuticals, and floor coverings. Machinery and transport equipment, chemical products, petroleum and petroleum products, food products, and textiles are among the major imports.

Tourist facilities

Tourism is the largest item in the country's net invisible earnings. Its value has increased considerably in recent years, due mainly to the promotional efforts of the Irish Tourist Board (Bord Fáilte Éireann) and to that body's work in encouraging new hotel construction, developing resort areas, extending sporting facilities, and increasing tourist amenities.

Trade unions. Almost all Irish trade unions are affiliated to the Irish Congress of Trade Unions (ICTU). There are also several employers' unions, some organized on a craft, and others on a regional, basis. The employers' central negotiating organization is the Federated Union of Employers. Wages and conditions of employment are normally matters for free collective bargaining, although industrial disputes may be referred to the Labour Court, an official body set up in 1946 with the main function of helping disputants to reach agreements.

Finance. The Central Bank of Ireland, established in 1942, is the national monetary authority and has responsibilities in regard to safeguarding the integrity of the currency. It does not transact business with the public, but it exerts a considerable influence on the volume of bank credit through the "advice" it gives to the commercial (or, to use the Irish term, the associated) banks. These are the Bank of Ireland, Allied Irish Banks, the Ulster Bank, and the Northern Bank. All four institutions have branches throughout the republic and in Northern Ireland as well. The Bank of Ireland and Allied Irish Banks are Irish-owned, while the other two banks are subsidiaries of London clearing banks. There is a trading floor of the London Stock Exchange in Dublin.

Transportation. *Roads and railways.* As a result of its scattered rural population, Ireland has a large road system. Most local roads are well-surfaced, and continuous progress has been made toward bringing the arterial roads up to the best modern standards.

The public transport system in the republic, including rail, road, and inland waterway services, is operated by a state company, the Irish Transport System (Córas Iompair Éireann; CIE). There are rail services between the principal cities and towns, but, as in other countries, many branch lines have become uneconomical and have been replaced by road services for passengers and goods. The city bus services in Dublin and the larger towns as well as a countrywide network of bus services are operated by the CIE. In addition, under contracts with the appropriate government authorities, it provides free transport for schoolchildren and free travel for old-age pensioners.

Waterborne traffic. There is no longer any commercial traffic on Irish canals. Of the two major canals in the country, the Royal, which joins the River Shannon with the Irish Sea via Mullingar and Dublin, is closed to navigation, but the Grand Canal, which also runs from the Shannon to the Irish Sea with a branch to the River Barrow, is maintained for use by pleasure craft.

Some of the smaller ports are of considerable importance to the local business communities concerned, but most of the country's seaborne trade tends to be conducted through the principal east and south coast ports, particularly Dublin, the country's major port, Waterford, and Cork. Dun Laoghaire, Dublin, Rosslare, and Cork are served by cross-channel passenger services to Britain. Overseas traffic in owner-driven automobiles has greatly increased since the introduction of ferry services to Britain at Dublin and Dun Laoghaire, and to Britain and the European continent at Rosslare.

The state-owned British and Irish (B & I) Steam Packet Company carries passengers, merchandise, and road vehicles between Irish and British ports. The trend toward larger vessels and the shipment of goods in containers has adversely affected the smaller Irish ports, as well as the smaller privately owned shipping companies. About half of the trade between the republic and continental ports is still carried in Irish ships, but the percentage of the country's total foreign trade carried in Irish vessels is considerably less.

Air facilities. There are international airports at Dublin, Shannon, Cork, and Knock. Shannon was the world's first duty-free airport; a state-sponsored company offers substantial tax and other advantages to manufacturing and warehousing concerns proposing to establish plants within the duty-free industrial estate at the airport. Aer Lingus-Irish International Airlines is a state-owned company.

ADMINISTRATION AND SOCIAL CONDITIONS

Government and justice. The Irish republic is a parliamentary democracy, although, unlike the United Kingdom, it has a written constitution, promulgated in 1937. Amendment to the constitution is by referendum.

The president of Ireland, head of state and first citizen, is elected by direct vote of the people for a term of seven years and is eligible for reelection for a second term. He normally acts on the advice of the government but acts in consultation with an advisory Council of State in the exercise of certain functions. The president signs and promulgates bills passed by the Oireachtas (Parliament) and,

when so advised by the *taoiseach* (prime minister), summons and dissolves Parliament. He may, however, refuse to dissolve Parliament on the advice of a *taoiseach* who has ceased to command a majority in the Dáil (House of Representatives). The president is the guardian of the constitution and may, in certain circumstances, submit a bill passed by Parliament to the people in a referendum, or refer it to the Supreme Court to decide on its constitutionality.

The president's role

There are two houses of the Oireachtas—the Dáil and the Seanad (Senate). Members of the Dáil are elected by adult suffrage in a secret ballot held at least every five years. There are 166 members of the Dáil and 60 members of the Seanad—11 appointed by the *taoiseach,* six elected by the Irish universities, and 43 elected to represent various economic, vocational, and cultural interests. The Seanad may delay, for a period of 90 days, bills passed by the Dáil, or it may suggest changes in them, but it cannot indefinitely block their passage into law.

The government, which consists of not fewer than seven or more than 15 members, is the executive power of the state and is headed by the *taoiseach,* who presides over its meetings. The *taoiseach,* the *tánaiste* (deputy prime minister), and the minister for finance must be members of the Dáil. The other ministers must be members of either house, but not more than two of them may be senators.

The major political parties are Fianna Fáil, Fine Gael, and the Labour Party. Fianna Fáil is a republican party founded by Eamon de Valera, who opposed the Anglo-Irish Treaty of 1921. The party boycotted the Dáil until 1927 but won the general election of 1932, when de Valera became prime minister, a position he held, with two intermissions, until 1959, when he became president. Fine Gael is the party of the Irish nationalists Arthur Griffith, Michael Collins, and W.T. Cosgrave, who supported the Treaty of 1921 and founded the Irish Free State. Fine Gael held office from 1922 to 1932 and, following World War II, participated in two coalition governments opposed to Fianna Fáil, as did the Labour Party.

Coalition government

The local government system comprises four county borough corporations, seven borough corporations in the major cities, and 27 county councils (one for each county except Tipperary, which for administrative purposes is divided into a North Riding and a South Riding, each having a county council). Each of these has a council elected, at regular intervals, by adult suffrage. County councils are responsible for physical planning, roads, health and public assistance services, sewerage and water supplies, housing, public libraries, fire services, and courthouses. County borough corporations have the same functions except that, in their case, health and public assistance are the responsibility of ad hoc authorities. Local government authorities in the republic have no functions in relation to police or education.

Important policy decisions—as on local taxes, borrowing, the making of bylaws—are made by the elected councils. Administration, on the other hand, is the responsibility of the county (or city) manager, who usually consults with members of the council before discharging important executive functions. There is a city manager for each county borough council and, for each county council, there is a county manager, who also acts as manager for the lesser local authorities within the country. Noncounty boroughs, urban districts, and towns have more limited duties, and, in regard to functions outside their scope, they form part of the administrative counties in which they are situated. The local government system is supervised by the national Department of Local Government.

Irish law is based partly on common law and partly on statute law. All judges are appointed by the president and normally serve until death or retirement. They may be removed from office only in the case of incapacity or misdemeanour and then only by resolution of both houses of the Oireachtas.

The judicial system

There are district courts and circuit courts, as well as a high court and a supreme court that acts as the court of final appeal. The supreme court consists of the chief justice, five other judges, and the president of the high court. The circuit court has jurisdiction to try all serious offenses except murder, treason, and piracy. Such trials take place before a jury. The central criminal court, which has unlimited criminal jurisdiction, consists of a high court judge sitting with a jury. Appeal lies from its decisions to the court of criminal appeal, which consists of a supreme court judge and two high court judges chosen by the chief justice. A special criminal court was established in 1972 with jurisdiction over cases of terrorism; appeal is made to the court of criminal appeal. All members of the judiciary, including district justices, have prescribed legal qualifications: there is no equivalent in the republic of the lay magistracy in England.

There are no local police forces in the republic. The Gárda Síochána (Civic Guard), established in 1922, is a nationwide force headed by a commissioner who is responsible to the minister for justice. A few hundred members of the force are assigned to detective duties; they are usually plainclothes officers and, when necessary, are armed. The rest of the force is uniformed and unarmed. Policewomen are employed on special duties.

Armed forces. The defense forces, including the air corps and the naval service, have played an active part in United Nations peacekeeping operations. Irish officers or forces served as UN observers in such places as Lebanon and other areas of the Middle East, Kashmir, the Congo (now Zaire), and Cyprus. There is no conscription in the republic; enlistment in the defense forces and the reserve force, as well as participation by enlisted men in overseas operations, is voluntary.

Education. Primary education is free, compulsory, and almost entirely denominational. There are five state-aided teachers' training colleges. The secondary school system is almost entirely denominational and privately owned. State grants are available to most secondary schools to enable them to waive payment of fees. Parallel to the secondary schools are the vocational schools, which are designed to provide primary schools' leavers with apprentice training and courses leading to qualifications in architecture, accountancy, engineering, and similar professions. The vocational education system includes schools of art, music, domestic science, and hotel training. Regional technical colleges provide advanced vocational courses. University education is given at the University of Dublin (Trinity College), founded in 1591, and at the National University of Ireland, founded in 1908. The latter has constituent colleges at Dublin, Cork, and Galway and several recognized colleges. National institutes for higher education in Limerick and Dublin emphasize applied studies in varied, flexible course structures. A higher education authority has been established to deal with the financial and organizational problems of higher education.

Health and welfare services. Health services are administered by eight regional health boards under the general supervision of the minister for health. School health examinations, child welfare clinics, and the treatment of tuberculosis and infectious diseases are available to all without charge. Otherwise, the cost of public health services to the patient depends on his means. Persons who cannot afford to pay are entitled to a comprehensive health service free of charge. A middle income group—insured workers, smaller farmers, and others of restricted means—is entitled to a free maternity and child welfare service and to free hospital and specialist services. People who are more comfortably off normally arrange and pay for their own medical advice and hospital services; but a voluntary health insurance program was established by law in 1957. Owing in large measure to the world-famous Irish Hospitals' Sweepstake, the republic developed an excellent system of hospitals.

The Irish Sweepstake

Pay-Related Social Insurance (PRSI) is paid by most employees aged 16 and over. Benefits include widows' and orphans' pensions, unemployment and disability benefits, deserted wives' allowances, and old-age pensions. Persons of inadequate means receive certain benefits on a noncontributory basis. These include widows' and orphans' pensions, old-age pensions, home assistance, unemployment assistance, and pensions for those disabled or blind. Children's allowances are paid to all households for each child under 16, irrespective of means.

CULTURAL LIFE

Although a few highly successful novels and plays have been written in Irish, the language has yet to produce a modern literature of significant proportions. The language steadily lost ground during the 18th and 19th centuries, but it never ceased to exert a strong, if imperceptible, influence on the English-speaking culture that was progressively taking its place. Many of the traditional modes of thought and expression characteristic of the Irish language were gradually absorbed into the English spoken in Ireland, and they served to enrich the creative imaginations and the powers of expression of those who wrote and spoke it. The remarkable contributions that Anglo-Irish literature and drama have made to the Western world may be ascribed, at least in part, to this cross-fertilization of two cultures. Be this as it may, it is noteworthy that so small a country should have nurtured so much creative literary genius—men like the great satirist Jonathan Swift, the orator and political theorist Edmund Burke, the novelist George Augustus Moore, the poet William Butler Yeats, and the modern prose masters James Joyce, Liam O'Flaherty, Frank O'Connor, Sean O'Faolain, Flann O'Brien, and Samuel Beckett. In the theatre, too, Irish talent has won world acclaim in the persons of William Congreve, Oliver Goldsmith, Richard Brinsley Sheridan, Oscar Wilde, George Bernard Shaw, and Sean O'Casey.

Literary heritage

The theatre continues to enjoy strong popular support. Dublin's famous Abbey Theatre, now rebuilt, stages Abbey classics as well as the work of new dramatists, in Irish as well as English. The Peacock Theatre, incorporated in the new Abbey building, concentrates on experimental theatre and theatre in Irish. There are a half dozen other small experimental theatres in Dublin alone, and Galway's Druid Theatre, which performs plays in Irish, is the most active provincial company. A vigorous amateur dramatic movement has spread throughout the country. The Wexford Opera Festival, held each year in the autumn, has an international reputation.

Many cultural institutions are specifically concerned with the popularization and preservation of various aspects of the traditional national culture—the Gaelic Athletic Association with the encouragement of the vigorous national games of hurling and football; the Gaelic League and other organizations with the spread of the Irish language; other bodies with the organization of folk-song and folk-music festivals (*feiseanna*) at which there are competitions in traditional storytelling and dancing as well as in music and singing. Some of these activities command large followings, particularly among the young.

Folk festivals

Other institutions are concerned with more general cultural interests—the Royal Irish Academy (1785) with science and humanities; the Royal Hibernian Academy (1823) with the fine arts; the Royal Dublin Society (1731) with the advancement of agriculture and the promotion of art and science; and the Royal Irish Academy of Music (1848). The Arts Council receives an annual grant from the state and in its turn assists the arts and artists by way of grants, scholarships, and awards. Individual writers, artists, and composers are also aided by tax concessions and by recognition and financial support from the Aosdána organization.

There is a healthy activity in the fine arts, and many younger artists and designers have gained international acclaim. The work of Irish painters, sculptors, and stained-glass artists is shown each year at three major exhibitions—the Exhibition of the Royal Hibernian Academy, the Oireachtas Exhibition, and the Exhibition of Living Art. The work of the Georgian Society and An Taisce (the National Trust of Ireland) has not been without impact in countering the effects of property developers on the architectural heritage of Dublin—one of the best-preserved 18th-century cities in Europe.

All radio and television stations in the republic are operated by Radio Telefís Éireann, an autonomous statutory body financed, as to capital, by repayable state grants, and as to current expenditure, by revenue from licenses and the sale of advertising time. Much of the country also can receive the British television networks.

For statistical data on the land and people of Ireland, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (F.H.B./J.O'B.R.)

History

Ireland, lying to the west of Britain, has always been to some extent cut off by it from direct contact with other European countries, especially the area between Sweden and the Rhine. Approach from France, Spain, and Portugal and even from Norway and Iceland has been more possible. Internally, the four ecclesiastical provinces into which Ireland was divided in the 12th century realistically denoted the main natural divisions of the country. Of these, the north had in the earliest times been culturally connected with Scotland, the east with Roman Britain and Wales, the south with Wales and France, and the southwest and west with France and Spain. In later times, despite political changes, these associations continued in greater or lesser degree.

The position of Ireland, geographically peripheral to western Europe, became central and thus potentially more important once Europe's horizons expanded, in the 15th and 16th centuries, to include the New World. But paradoxically it was in the earlier period that Ireland won especial fame as a notable and respected centre of Christianity, scholarship, and the arts. After the Middle Ages, subjugation to Britain stultified, or the struggle for freedom absorbed, much of Ireland's native energy. But its influence was always exercised as much through its emigrants as in its achievements as a nation. During the centuries of British occupation the successors of the great missionaries and scholars who, from the 7th to the 9th century, fostered Christianity and learning among the Germanic peoples of the Continent were those who formed a considerable element in the armies and clergy of Roman Catholic countries and had an incalculable influence on the later development of the United States. In British history innumerable great men of Anglo-Irish origin or nurture have, as statesmen or soldiers, played vital parts; the influence of Ireland itself on Britain has been constant and profound.

Ireland's influence

EARLY IRELAND

The human occupation of Ireland did not begin until a late stage in the prehistory of Europe. It has generally been held that the first arrivals were Mesolithic hunter-fisher people, represented largely by flintwork found mainly in ancient beaches in Antrim, Down, Louth, and Dublin. These artifacts have been named Larnian for the type site at Larne, County Antrim; dates from 6000 BC onward have been assigned to them. More recent work, however, casts considerable doubt on the antiquity and affinities of the people who were responsible for the Larnian industry; association with Neolithic remains suggests that they should be considered not as a Mesolithic people but rather as groups contemporary with the Neolithic farmers. The Larnian could then be interpreted as a specialized aspect of contemporary Neolithic culture. Lake and riverside finds, especially along the River Bann, show a comparable tradition. A single radioactive carbon date of 5725 ± 110 BC from Toome Bay, north of Lough Neagh, for woodworking and flint has been cited in support of a Mesolithic phase in Ireland, but such a single date cannot be relied on.

Neolithic Age. The general pattern of radioactive-carbon determinations suggests that the Neolithic period in Ireland began about 3000 BC. As in Britain, the most widespread evidence of early farming communities is long-barrow burial. The main Irish long-barrow series consists of megalithic tombs called court tombs because an oval or semicircular open space, or court, inset into the end of the long barrow, precedes the burial chamber. These court tombs number more than 300. They occur in the northern half of Ireland, and the distribution is bounded on the south by the lowlands of the central plain. Timber-built rectangular houses belonging to the court-tomb builders have been discovered at Ballynagilly, County Tyrone, and at Ballyglass, County Mayo. The court tombs are intimately related to the British long-barrow series of the

Megalithic tombs

Severn-Cotswold and chalk regions and probably derive from more or less common prototypes in northwestern France.

In Ireland a second type of megalithic long barrow—the so-called portal tomb—developed from the court tomb. There are more than 150 examples. They spread across the court tomb area in the northern half of Ireland and extend into Leinster and Waterford and also to western Wales and Cornwall.

A notable feature of the Irish Neolithic is the passage tomb. This megalithic tomb, unlike the long-barrow types, is set in a round mound, sited usually on hilltops and grouped in cemeteries. The rich grave goods of these tombs include beads, pendants, and bone pins. Many of the stones of the tombs are elaborately decorated with engraved designs. The main axis of the distribution lies along a series of great cemeteries from the River Boyne to Sligo (Boyne and Loughcrew in County Meath, Carrowkeel and Carrowmore in County Sligo). Smaller groups and single tombs occur largely in the northern half of the country and in Leinster. A specialized group of later, indeed advanced, Bronze Age date, near Tramore, County Waterford, is quite closely akin to a large group on the Scilly Islands and Cornwall. The great Irish passage tombs include some of the most magnificent megalithic tombs in all Europe—for example, Newgrange and Knowth in County Meath. While the passage tombs represent the arrival of the megalithic tradition in its fullest and most sophisticated form, the exact relation between the builders of these tombs and the more or less contemporary long-barrow builders is not clear. The passage tombs suggest rather more clearly integrated communities than do the long barrows.

To the final stage of the Neolithic probably belong the rich house sites of both rectangular and circular form at Lough Gur, County Limerick. The pottery shows a strong connection with the tradition of the long barrow (court tomb and portal tomb).

Bronze Age. Two great incursions establish the early Bronze Age in Ireland. One, represented by approximately 400 megalithic tombs of the wedge tomb variety, is associated with Beaker pottery. This group is dominant in the western half of the country. Similar tombs also associated with Beaker finds are common in Brittany, and the origin of the Irish series is clearly from this region. In Ireland the distribution indicates that these tomb builders sought well-drained grazing land, such as the Burren limestones in Clare, and also copper deposits, such as those on the Cork-Kerry coast and around the Silvermines area of Tipperary.

In contrast, in the eastern half of the country a people in the single-burial tradition dominate. Their burial modes and distinctive pottery, known as food vessels, have strong roots in the Beaker tradition that dominates in many areas of western Europe. They may have reached Ireland via Britain from the lowland areas around the Rhine or farther north.

Throughout the early Bronze Age Ireland had a flourishing metal industry, and bronze, copper, and gold objects were exported widely to Britain and the Continent. In the middle Bronze Age (about 1500 BC) new influences brought urn burial into eastern Ireland. From about 1200 BC elements of a late Bronze Age appear, and by about 800 BC a great late Bronze Age industry was established. A considerable wealth of bronze and gold is present, an example of which is the great Clare gold hoard. Nordic connections have been noted in much of this metalwork.

Iron Age. The period of the transition from the Bronze Age to the Iron Age in Ireland is fraught with uncertainties. There is little sign of a continental-like Hallstatt Iron Age in Ireland. La Tène culture, which may date in Ireland from 300 BC or earlier, is represented in metalwork and some stone sculpture, mainly in the northern half of the country. Connections with northern England are apparent. Hill-fort building seems also characteristic of the Iron Age. The problem of identifying archaeological remains with language grouping is notoriously difficult, but it seems on the whole likely that the principal Celtic arrivals occurred in the Iron Age. Irish sagas, which probably reflect the pagan Irish Iron Age, reflect conditions in many respects similar to the descriptions of the ancient classical authors, such as Poseidonius and Julius Caesar.

Early Celtic Ireland. *Political and social organization.* Politically, Ireland was organized into a number of petty kingdoms, or *tuatha,* each of which was quite independent under its elected king. Groups of *tuatha* tended to combine, but the king who claimed overlordship in each group had a primacy of honour rather than of jurisdiction. Not until the 10th century AD was there a king of all Ireland (*árd rí Éireann*). A division of the country into five groups of *tuatha,* known as the Five Fifths (Cuíg Cuígí), occurred about the beginning of the Christian Era. These were Ulster (Ulaid), Meath (Midhe), Leinster (Laigin), Munster (Muma), and Connaught (Connacht). The Five Fifths

Surrounding a king was an aristocracy (*airi aicme,* the upper class), whose land and property rights were clearly defined by law and whose main wealth was in cattle. Greater landowners were supported by *céilí,* or clients. These and other grades of society, minutely classified and described by legal writers, tilled the soil and tended the cattle. Individual families were the real units of society and exercised collectively powers of ownership over their farms and territory. At law the family (*fine*) did not merely act corporatively but was, by one of the oldest customs, held responsible for the observance of the law by its kindred, serfs, and slaves.

Rural economy and living conditions. There were no urban centres, and the economic basis of society was cattle rearing and agriculture. The principal crops were wheat, barley, oats, flax, and hay. The land was tilled with plows drawn by oxen. Sheep appear to have been bred principally for their wool, and the only animal reared specifically for slaughter was the pig. Fishing, hunting, fowling, and trapping provided additional food. Transport of goods over land was by packhorse, for wheeled vehicles appear to have been few. Sea transport was by currach, a wicker-framed boat covered with hides; the normal freshwater craft was the dugout.

The dwellings of the time were built on the post-and-wattle technique, and some, at least, were situated within the protected sites called by archaeologists ring forts. Excavations have shown that some of these may have existed even in the Bronze Age and that they remained a normal place of habitation down to medieval times. Advantage was also taken of the relative security of islands in rivers or lakes as dwelling places; and artificial islands, called crannogs, were also extensively made.

The Irish laws point to a large development of rural industry in the period in which they were first written down, shortly before the Norse invasions beginning at the end of the 8th century. They deal minutely not only with the management of land and animal rearing but also with innumerable further details of husbandry, including milling, dyeing, dairying, malting, meat curing, and spinning and weaving. Wool was spun with a wooden spindle weighted with a whorl of bone or stone, and it was woven on a loom. The outer garment worn by both men and women was a large woolen cloak (*brat*), fastened on shoulder or breast with a pin or brooch. The inner garment was a long linen tunic (*léine*), girded at the waist with a belt. Shoes of rawhide or tanned leather were worn, at least by the upper classes and the higher professional ranks. A large amount of metalwork reveals the adaptation by Irish craftsmen of many techniques originating in Britain or on the Continent. An instinct for design, added to the skillful use of these techniques, enabled them to produce many superb objects, of which the Tara brooch, dating from about the mid-8th century, is an outstanding example. The chief musical instrument of the period was the harp.

Early political history. The documentary history of Ireland begins only in the 7th century, which saw the production both in Latin and Irish of sufficiently rich and numerous records of all sorts. For events before that time historians must rely on literary sources such as the sagas, many of whose characters may represent only poetic imagination and in which the social or political circumstances portrayed reflect the fantasies of their authors rather than historical reality. Nevertheless, the traditions seem to indicate, during the early centuries of the Christian Era, a

process of political cohesion in Ireland through which the *tuatha* ultimately became grouped into the Five Fifths. Among these, Ulster seems at first to have been dominant; but by the time of Niall of the Nine Hostages (died early 5th century), hegemony had passed to his midland kingdom of Meath, which was then temporarily associated with Connaught. In the 6th century, descendants of Niall, ruling at Tara in northern Leinster, were claiming to be overkings of three provinces, Ulster, Connaught, and Meath. Later, they claimed to be kings of all Ireland, although their power rarely extended over Munster or the greater part of Leinster. Two branches of Niall's descendants, the Cenél nEogain, of the northern Uí Néill, and the Clan Cholmáin, of the southern Uí Néill, alternated as kings of Ireland from 734 to 1002, a fact that suggests a formal arrangement between the two septs. Inevitably, claims to a high kingship came to be contested by the rulers of Munster, who, from their capital at Cashel, had gradually increased their strength, depriving Connaught of the region that later became County Clare. But not until the reign of Brian Boru in the 11th century was Munster sufficiently strong to secure a real high kingship over all Ireland.

The Irish kingship

Irish raids and migrations. From about the middle of the 3rd century Latin writings make frequent reference to raiding expeditions carried out by the Irish, who were now given the new name, Scoti, rather than the older one, Hiberni. Native Irish traditions also suggest that such attacks took place. In the second half of the 4th century, when Roman power in Britain was beginning to crumble seriously, the raids became incessant, and settlements were made along the west coast of Britain and extensively in Wales and Scotland. From the early 5th century the rulers of Dalriada in northern Antrim extended their power over the Irish already settled in Argyll and the neighbouring islands. Ultimately the Scottish kingdom of Dalriada became separated from the Irish; and when, in the 9th century, it overcame the Picts, it gave its name, Scotland, to the whole area.

Early Christianity. *Conversion.* Little is known of the first impact of Christianity upon Ireland. Traditions in the south and southeast refer to early saints who allegedly preceded St. Patrick, and their missions may well have come through trading relations with the Roman Empire. The earliest firm date is AD 431, when St. Germanus, bishop of Auxerre in Gaul, proposed, with the approval of Pope Celestine I, to send a certain Palladius to "the Scots believing in Christ." Subsequent missionary history in Ireland is dominated by the figure of St. Patrick, whose 7th-century biographers, Tirechán and Muirchú, credited him with converting all the Irish to Christianity and won for him the status of national apostle.

St. Patrick

A 9th-century record, the *Book of Armagh,* includes a work by Patrick himself, the *Confessio* ("Confession," a reply to charges made by British ecclesiastics), in which he describes his life at a Roman villa in Britain, his capture by Irish raiders, and his seven years of slavery in Ireland. Recovering his freedom, he was educated and ordained to the priesthood and eventually managed to be sent as a missionary to Ireland. He concentrated on the north and west of the country, achieving remarkable success; he did not himself claim to have converted all Ireland. Confusion exists regarding the chronology of Patrick's life, and it is seriously contended that tradition came to merge the experience of two men, the continental Palladius and the Patrick of the *Confessio.* No sufficient evidence supports the traditional date (AD 432) for the beginning of Patrick's mission, while of the rival dates (AD 461/462 and AD 492/493) given for his death in annals and biographies, the latter is now preferred.

Irish monasticism. Though monks and monasteries were to be found in Ireland at the time of Patrick, their place was then altogether secondary. But in the course of the 6th and 7th centuries a comprehensive monastic system developed in Ireland, partly through the influence of Celtic monasteries in Britain, such as Candida Casa at Whithorn in Galloway and Llangarvan in Wales. Early attempts to organize the Irish Church on the usual Roman system—by which each bishop and his clergy exercised exclusive jurisdiction within a diocese—seem to have given way to one in which groups of Christian settlements were loosely linked together, usually under the auspices of some one or other of the great saints. Careful study of the lives of the early saints reveals the manner in which their reputations developed in proportion to the power of the political dynasties that became connected with them.

By the end of the 6th century, enthusiasm for Christianity was leading Irishmen to devote themselves to a most austere existence as monks, as hermits, and as missionaries to pagan tribes of Scotland, in the north of England, and in a great area of west central Europe, particularly between the Rhine, Loire, and Rhône rivers. St. Columba's foundation (*c.* 563) of the monastery of Iona off the northwest Scottish coast provided the best-known base for the Celtic Christianization of Scotland; and its offshoot, Lindisfarne, lying off the coast of the Anglo-Saxon kingdom of Northumbria, was responsible for the conversion of that area. Of the continental missionaries, the best known is St. Columban (*c.* 543–615), whose monastic foundations at Luxeuil near Annegray in the Vosges and at Bobbio in northern Italy became important centres of learning. Columban, however, by his individualism and austere puritanism, came into conflict not only with the Merovingian rulers of Gaul but also with the local ecclesiastical administration; his limitations exemplify those of the Irish monastic system as a whole and explain why, in the end, it was supplanted by the ordinary administrative system of the church.

Learning and art. Both at home and abroad the saints were succeeded by scholars, whose work in sacred and classical studies and particularly in elaborating an Irish Christian mythology and literature was to have profound effects upon the Irish language and was to be a major factor in its survival. The Irish monasteries became notable centres of learning; in Ireland itself those of Clonmacnoise (in County Offaly) and Clonard (in County Meath) were among the most famous. Christianity had brought to Ireland the Latin tongue and command of the Latin authors; not only Church Fathers but also classical writers were read and studied. Irish scribes produced manuscripts written in the clear hand known as Insular; this usage spread from Ireland to Anglo-Saxon England and to Irish monasteries on the Continent. Initial letters in the manuscripts were illuminated, usually with very intricate ribbon and zoomorphic designs. The most famous of the Irish manuscripts is the *Book of Kells,* a copy of the Four Gospels dating from about the 8th or 9th century. The earliest surviving illuminated manuscript, the *Book of Durrow,* was probably made about a century earlier.

Illuminated manuscripts

The adoption of Christianity made it necessary to relate the chronology of Irish tradition, history, and genealogies to the events recorded in the Bible. The *Book of Invasions* (*Leabhar Gabhála*), in which Irish history was linked with events in the Old Testament, was a notable example of this process. In this way Latin civilization in Ireland became linked to the Gaelic; and the association became closer under the impact of the Viking wars. Gradually the Latin products of the Christian schools became replaced by Irish works; Latin lives of the saints, for example, are almost always earlier in date than those written in Irish. Recurring bouts of puritanism and reforming movements in the church tended to remove secular literature from monastic control; ultimately, there developed a class of professional families who were its custodians from the 12th to the 17th century. The medieval secular writers, employing a degenerate form of Old Irish usually known as Middle Irish, were responsible for a large proportion of Irish literary achievement; their historical works, the annals, and the great genealogies, supplemented by the law collections, have enabled historians to reconstruct early Irish social history.

The Norse invasions and their aftermath. The first appearance of the Norsemen on the Irish coast is recorded in 795. Thereafter the Norsemen made frequent plundering raids, sometimes far inland. They seized and fortified two ports, Annagassan and Dublin, in 838; and the 840s saw a series of large-scale invasions in the north of the country. These invaders were driven out by Aed Finnliath, high

king from 862 to 879, but meanwhile the Norse rulers of Dublin were reaching the zenith of their power. They took Waterford in 914 and Limerick in 920. Gradually, without quite abandoning piracy, the Vikings became traders in close association with the Irish, and their commercial towns became a new element in the life of the country. The decline of Norse power in the south began when they lost Limerick in 968 and was finally effected when the Scandinavian allies of the king of Dublin were defeated by High King Brian Boru at the Battle of Clontarf in 1014.

Though the Battle of Clontarf removed the prospect of Norse domination, it brought a period of political unsettlement. High kings ruled in Ireland but almost always "with opposition," meaning that they were not acknowledged by a minority of provincial kings. The Viking invasions had, in fact, shown the strength and the weakness of the Irish position. The fact that power had been preserved at a local level in Ireland enabled a maximum of resistance to be made; and although the invaders established maritime strongholds, they never achieved any domination comparable to their control of eastern England or northwest France. After Clontarf they remained largely in control of Ireland's commerce but came increasingly under the influence of neighbouring Irish kings.

In the 11th and 12th centuries the ecclesiastical reform movement of western Europe was extended into Ireland. And as the kings of Munster and Connaught, Leinster and Ulster struggled each to secure the dominant position that had once been held by Brian Boru, they came to realize the value to them of alliance with the forces of church reform. Thus, with the aid of provincial rulers, the reformers were able to set up in Ireland a system of dioceses, whose boundaries were coterminous with those of the chief petty kingdoms. At the head of this hierarchy was established the archbishopric of Armagh, in association with the province of Ulster dominated by the royal family of Uí Néill. But the victory of the reformers was not complete, for the parochial system was not introduced until after the Anglo-Norman Conquest. Moreover, the reformers sought to influence Irish conduct as well as church organization. The enormities of Irish moral behaviour were colourfully described by St. Bernard of Clairvaux in his life of his contemporary St. Malachy, the reforming bishop who introduced the Cistercian monks into Ireland. The reforming popes Adrian IV and Alexander III encouraged Henry II's invasion of Ireland, believing that it would further church reform in that country. In a remarkable account of the conquest, Gerald of Wales provided a lurid description of the archaic Irish civilization that the invaders encountered. The recognition of Henry II as lord of Ireland and the linking of the church to a foreign administration terminated the independence of Gaelic Ireland and reduced the country to a position of subordination for centuries to come.

Church reform

FIRST CENTURIES OF ENGLISH RULE (C. 1166–C. 1600)

The Anglo-Norman invasion. Before the arrival of Henry II in Ireland (October 1171), Anglo-Norman adventurers—including Richard de Clare, earl of Pembroke, subsequently known as "Strongbow," invited by Dermot MacMurrough, a king of Leinster who had been expelled by the high king, Rory O'Connor (Roderic)—had conquered a substantial part of eastern Ireland, including the kingdom of Leinster, the towns of Waterford, Wexford, and Dublin, and part of the kingdom of Meath. Partly to avert any chance of Ireland's becoming a rival Norman state, Henry took action to impose his rule there. He granted Leinster to de Clare and Meath to Hugh de Lacy, who had gone to Ireland in the King's army; but he kept the chief towns in his own hands, exacted forms of submission from the Irish kings, and secured from a church synod recognition of his overlordship. During subsequent years the Anglo-Norman sphere in Ireland was extended; and while all the Irish kings, except in the northwest, agreed to recognize his supremacy, Henry was obliged to acquiesce in the establishment of new Norman lordships in Ulster under John de Courci and in Munster under de Cogan, de Braose, and others. By the Treaty of Windsor (1175), O'Connor, the high king, accepted Henry as his overlord and restricted his own style to that of king of Connaught. But he was permitted to exercise some vague authority over the other Irish kings and was charged with collecting from them tribute to be paid to Henry. This arrangement was unsuccessful, for thereafter O'Connor encountered opposition even in his own province, and he was ultimately obliged to abdicate.

King John, who visited Ireland in 1210, established there a civil government independent of the feudal lords, and during the 13th century it became more fully organized. An Irish exchequer had been set up in 1200, and a chancery followed in 1232. The country was divided into counties for administrative purposes, English law was introduced, and serious attempts were made to reduce the feudal liberties of the Anglo-Norman baronage. (Counties were civil administration districts, whereas liberties were lands held in the personal control of aristocratic families and the church.) Parliament started in Ireland, as in England; in 1297 the peers and prelates were joined by representatives of counties, and in 1300 the towns also sent members. But these represented the Anglo-Irish only; the native Irish, to some extent resurgent in Ulster under the O'Neills and O'Donnells, and in southwest Munster under the MacCarthys, were aloof and unrepresented.

The medieval Irish Parliament

The 14th and 15th centuries. A brief threat to English control of Ireland, made by Edward Bruce, brother of King Robert I of Scotland, ended when Bruce was killed in battle at Faughart near Dundalk (1318). English control was reasserted and strengthened by the creation of three new Anglo-Irish earldoms, that of Kildare, given to the head of the Leinster Fitzgeralds; that of Desmond, given to the head of the Munster Fitzgeralds; and that of Ormonde, given to the head of the Butlers, who held lands around Tipperary. But the increased power and lands of



Administrative units of late medieval Ireland.

the Anglo-Irish brought about an inevitable reaction; and the remainder of the 14th century witnessed a remarkable revival of Irish political power, which was matched by a flowering of Irish language, law, and civilization. The Gaels recovered large parts of Ulster, the midlands, Connaught, and Leinster, while the Anglo-Irish became increasingly Irish, marrying Irish women and often adopting Gaelic customs.

The English government, which in any case, because of its aim to curtail feudal privileges, was always to some extent opposed by the Anglo-Norman aristocracy, made an effort to restore control but achieved little more than a definition of the status quo. Edward III's son, Lionel, duke of Clarence, as viceroy from 1361 to 1367, passed in the Irish Parliament the Statutes of Kilkenny (1366), which listed the "obedient" (English-controlled) lands as Louth, Meath, Trim, Dublin, Carlow, Kildare, Kilkenny, Wexford, Waterford, and Tipperary. Intermarriage or alliances with the Irish were forbidden. The independent Irish outside the Pale (the area of English control) were regarded as enemies and were assumed to possess their lands only by usurpation. In practice they were feared, and their attacks were often bought off by almost regular payments. Visits by King Richard II in 1394–95 and 1399 achieved nothing. During the first half of the 15th century Ireland was, in effect, ruled by the three great earls, of Desmond, Ormonde, and Kildare, who combined to dominate the Dublin government. Desmond had sway in Counties Limerick, Cork, Kerry, and Waterford; Ormonde in Tipperary and Kilkenny; and Kildare in Leinster. Although both the Gaels and the Anglo-Irish had supported the Yorkist side in the Wars of the Roses, the Yorkist king Edward IV found them no less easy to subjugate than had his Lancastrian predecessors. Succeeding (1468) in bringing about the attainder and execution for treason of Thomas, earl of Desmond, he was, nevertheless, obliged to yield to aristocratic power in Ireland. The earls of Kildare, who thereafter bore the title of lord deputy (for the English princes who were lords lieutenant), were, in effect, the actual rulers of Ireland until well into the 16th century.

The Kildare ascendancy. The substitution (1485) of Tudor for Yorkist rule in England had no apparent effect in Ireland, where the ascendancy of the Fitzgerald earls of Kildare, established when Thomas, 7th earl, was created lord deputy in 1471, had passed (1477) to his son Garret More (Great Gerald). The fiction of the king's power was preserved by appointing an absentee lieutenant, from whom Kildare held as deputy; in practice, any real power was exercised largely through dynastic alliances with the chief Gaelic and Anglo-Irish lords. Opposition to Kildare was negligible so long as the king was unable to maintain a permanent power to which his opponents might turn; an attempt to displace him was made when Kildare gave support (1487) to Lambert Simnel, a pretender to the English throne. After the advent of a more dangerous pretender, Perkin Warbeck, it was decided (1494) to remove Kildare and rule through an Englishman, Sir Edward Poynings. Poynings subdued Kildare, but he could not reconquer the northern Gaelic Irish. At Drogheda (1494–95) he induced a Parliament to pass an act that came to be known as "Poynings' Law"; it subjected the meetings and legislative drafts of the Irish Parliament to the control of the English king and council. But his administrative expenses were too great, and Henry VII decided in 1496 to restore Kildare.

Sir Edward Poynings

On Kildare's death (1513), the deputyship passed to his son Garret Oge (Young Gerald), 9th earl of Kildare, who continued, though less impressively, to dominate the country. But James, 10th earl of Desmond, intrigued with the emperor Charles V; and Henry VIII became convinced that Kildare had lost the power to keep Ireland neutral. Therefore, when the divorce (1533) of Catherine of Aragon made the danger of imperial intervention particularly acute, the King summoned Kildare to England (1534). There were, thereafter, no Irish-born viceroys for more than a century.

The Reformation period. Rumours that Kildare had been executed precipitated the rebellion of Kildare's son, Thomas Fitzgerald, Lord Offaly, called Silken Thomas. The rebellion facilitated the transition to the new system. Silken Thomas had opposed Henry VIII's breach with Rome; his execution (1537) caused a revival of the power of the Butlers of Ormonde; Piers Butler, earl of Ossory, helped to secure the enactment of the royal (instead of papal) ecclesiastical supremacy by the Dublin Parliament of 1536–37. As a further step in getting rid of the pope, in 1541 a complaisant Parliament recognized Henry VIII as king of Ireland (his predecessors had held the title of lord of Ireland). Confiscation of monastic property, as well as the lands of the rebels, met most of the costs of the expanded administration. This loss of land inevitably drove the religious orders and the Anglo-Irish into the arms of the Gaelic Irish, thus weakening the old racial rivalries of medieval Ireland.

Sir Anthony St. Leger, lord deputy in 1540–48 and again in 1550–56, then began a conciliatory policy by which outstanding lords were persuaded, in order to gain new titles and grants of lands, to renounce the pope and recognize the king's ecclesiastical supremacy. This policy, however, required a steady series of efficient governors and disciplined administrators; in fact, neither in Tudor nor in Stuart times did the English succeed in converting elective chiefs into hereditary nobles holding offices delegated by the crown. Moreover, even those who had recently submitted were often condemned for religious conservatism and deprived of their lands. St. Leger's personal success was all the more remarkable because the first Jesuit mission to Ireland arrived in the north in 1542.

Under Edward VI (1547–53), the Dublin authorities carried out a forward policy in religion as well as in politics, but Protestantism got no support except from English officials. The official restoration of Roman Catholicism under Queen Mary (1553–58) revealed the strength of resentment in Ireland against Protestantism. As in England, the papal jurisdiction was restored, but otherwise the Tudor regulations of authority were observed. The pope was induced to recognize the conversion of the Tudor Irish lordship into a kingdom. Finally, Mary gave statutory approval for the plantation, or resettlement of Irish lands by Englishmen, of Leix, Offaly, and other Irish lordships of the central plain. Her viceroy was Thomas Radcliffe, earl of Sussex, lord deputy (1556–59), who was soon, as lord lieutenant (1559–66) for Elizabeth I, to restore the state's authority over the church.

Ireland under Elizabeth I. The Acts of Supremacy and Uniformity, enforcing the Anglican Church settlement, were passed in Ireland in 1560, but fear of driving the inhabitants of the Pale into alliance with the Gaelic Irish and perhaps with the Spanish made the government lenient in enforcing the terms of the Acts. Political affairs continued to preoccupy the administrators, so that the new Protestant church was unequipped to resist the forces of the Counter-Reformation. This was inevitable in an Ireland only superficially conformed to royal obedience; but the seriousness of the situation was shown by the three great rebellions of the reign, those of Shane O'Neill (1559), of the Fitzgeralds of Desmond (1568–83), and of O'Neill (Tyrone) and O'Donnell (1594–1603).

The rebellion of Shane O'Neill

The first of these rebellions, that of Shane O'Neill, fully exposed the weakness and later the folly of the government. O'Neill's father, Con Bacach (the Lame), who as the "O'Neill" was head of a whole network of clans, had been made earl of Tyrone in 1541, the succession rights of his illegitimate son Feardorchadh (Matthew) being recognized. Shane, younger but the eldest legitimate son, was, nevertheless, elected O'Neill on his father's death (1559), and soon afterward Feardorchadh was killed. O'Neill then took the field against the Dublin government, demanding recognition according to the laws of primogeniture, and he insisted that neither of Feardorchadh's sons, Brian and Hugh, had claims to the earldom. Elizabeth invited O'Neill to London to negotiate, but the opportunity for a statesmanlike settlement was lost. O'Neill was to be "captain of Tyrone" and was encouraged to expel from Antrim the MacDonnell (MacDonald or MacConnell) migrants from Scotland. Returning to Ireland in May 1562, O'Neill routed the MacDonnells, as well as the loyal O'Donnells of the northwest, and attempted to secure support from Scotland and France. Eventually the government was

saved from a serious situation only through the defeat of O'Neill by the O'Donnells and his murder in 1567 by the MacDonnells.

The lands of the O'Neills and even of loyal Gaelic lords were declared forfeit in 1569, and in a wave of enthusiasm for colonization, various questionable adventurers were permitted to attempt substantial plantations in Munster, Leinster, and Ulster. The folly of this policy was seen when the government, despite its having declared the position of "O'Neill" extinguished, yet allowed the O'Neills to elect Shane's cousin, Turloch Luineach, as their chief. Butlers and Munster Fitzgeralds also combined forcibly to resist the plantations. The only gleam of statesmanship shown in these years by Henry Sidney, lord deputy (1565–71, 1575–78), was that he managed to avoid a major combination against the government's religious policy. The Butlers were induced to submit, the planters were given only limited support, and a head-on collision with Turloch Luineach was averted. When the Ulster plantation plans could not be carried out against Irish resistance, the Queen wisely decided they should be dropped. The pardon of the Butlers pacified Leinster, and although in Munster the Earl of Desmond's cousin James Fitzgerald, called "Fitzmaurice," attempted to make the war one of religion, he too was eventually pardoned.

The Desmond wars

Despite his pardon, Fitzmaurice in 1575 fled to the Continent, returning to Ireland in 1579 with papal approval for a Roman Catholic crusade against Queen Elizabeth. Although neither France nor Spain supported it, and Fitzmaurice was surprised and killed in August 1579, the government was extremely apprehensive. Gerald Fitzgerald, 14th earl of Desmond, then assumed direction of the enterprise. As a military commander he was wholly deficient, and his mediocrity may well have kept outstanding figures in the north and west out of the movement. The rebels were defeated, and in November 1580 a force of Italians and Spaniards was massacred at Dún an Óir (Golden Fort), Smerwick Harbour, County Kerry.

The end of the Desmond rebellion gave the government the opportunity to confiscate more than 300,000 acres (100,000 hectares) in Munster and initiate more stringent proceedings against Roman Catholics. But the plantation was not a success. A more statesmanlike attitude was displayed in regard to Connaught land titles. When Sir John Perrot was lord deputy (1584–88), a number of agreements were made with individual landowners and chieftains, by which their titles were officially recognized in return for fixed regular payments. This was a step in the process of converting a great part of the country to English tenures. Perrot was less successful in handling the 1585–86 Parliament, in which the government's anti-Catholic program was defeated by the opposition.

The Tyrone war

The origins of the third rebellion, the O'Neill (Tyrone) war, remain in doubt. Both Hugh Roe O'Donnell and Hugh, younger son of Feardorchadh O'Neill, for whom the earldom of Tyrone had been revived in 1585 and who had had himself elected O'Neill on Turloch Luineach's death in 1595, certainly resented the extension of the royal administration, but the religious issue was probably more important. For a generation exiled Roman Catholics had been trained as missionaries in the continental colleges of the Counter-Reformation, and the majority of those who returned to Ireland concluded that Roman Catholicism could only survive there if Elizabeth were defeated. The outbreak of hostilities in Ulster in 1594 was at first confined to the northwest, where O'Donnell and Maguire, lord of Fermanagh, tried to drive out the English troops. The intervention of Hugh O'Neill was expected, if not inevitable. His participation with his brother-in-law O'Donnell proved decisive in the north and west, and the English were defeated both in Ulster and in Connaught. A more intimidating combination thus threatened Dublin than even in Shane O'Neill's time. Even in the Pale, arbitrary exactions and exclusion from offices won Hugh much sympathy, and it was said that he knew of Dublin Castle decisions before they were known in the city. Resentful of O'Neill's alleged ingratitude, Elizabeth became impatient of negotiations with him and finally sent Robert Devereux, earl of Essex, to Ireland (1599) to subdue him. But Essex lost his reputation by his inglorious progress through the country and by the speed with which he returned to England after a private conversation with O'Neill. Before Charles Blount, Lord Mountjoy, arrived (1600) to replace Essex, the Irish leaders had gained the qualified support of Pope Clement VIII and of King Philip III of Spain. But Philip could afford to send only a minimal force to aid the Irish rebels. Its leader, Juan del Aguila, occupied Kinsale and was besieged (1601) by Mountjoy. O'Neill marched south to relieve Aguila, but a rash attempt to surprise the English lines by night proved disastrous (Dec. 24, 1601); the Irish were defeated and the Spaniards surrounded. O'Neill held out in Ulster for more than a year but finally submitted a few days after the Queen's death (March 1603).

Viewed generally, Elizabeth's Irish policy had the distinction of having reduced the country to obedience for the first time since the invasion of Henry II. But the cost was a serious one. The loyalty of the Irish was perennially strained over the religious issue, so that further rebellion was almost inevitable and virtually predictable in 1640 when the English government was embarrassed by the Second Bishops' War with Scotland. Economically, the towns and the countryside were needlessly exploited by the new administrators and planters, while the Queen's expenditure was substantially increased. Commitments in Ireland were at least partly responsible for the poverty of the crown, which was to become a serious factor in precipitating its 17th-century conflict with Parliament.

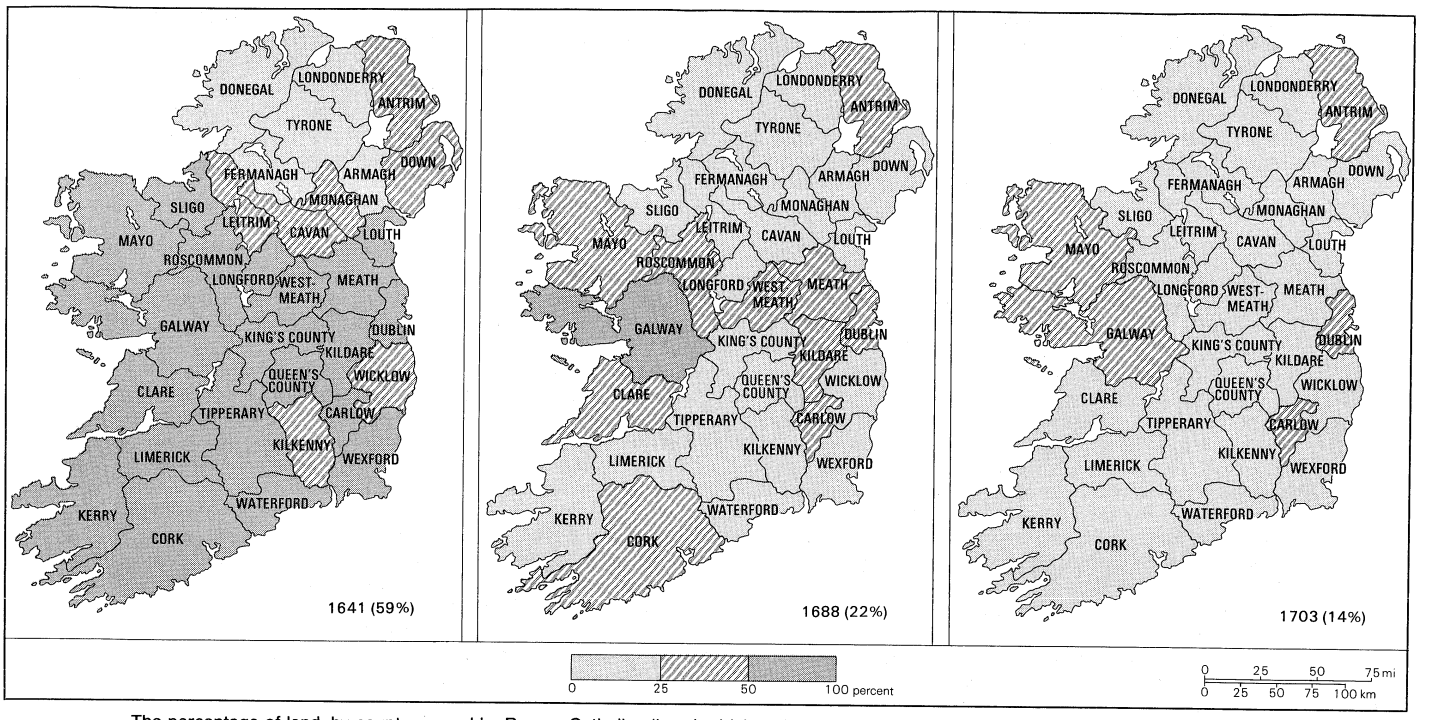
MODERN IRELAND UNDER BRITISH RULE

The 17th century. *James I (1603–25).* James VI of Scotland, who also became King James I of England and Ireland in 1603, might have pursued an Irish policy more enlightened than that of Elizabeth, who had been committed to war against the papacy and against Spain. He did make peace with Spain, but his policy of guarded religious toleration was nullified by the intransigence of the established Anglican Church and of the papacy. Unfortunately, James allowed Irish policy to be dominated by the interests of the English governing class and also sought to provide in Ireland opportunities for his countrymen. He thereby virtually continued Elizabethan policy, and as a result the steady exodus of Irish soldiers and churchmen to Roman Catholic countries on the Continent was unabated. On a short-term basis, their absence contributed to peace; but their influence abroad made the Irish question an international one. In Ireland, the overwhelming majority of the Gaelic Irish and of the old Anglo-Irish remained detached from government in attitude as well as in way of life.

Plantation of Ulster

As soon as James's policy became clear, the earls of Tyrone and of Tyrconnel and other Ulster Gaelic lords joined the flight from Ireland. Their departure opened the way for the plantation of Ulster by a new landowning class, which included Scots as well as Englishmen. This proved the most successful British settlement made in Ireland, mainly because British tenantry and labourers were introduced as well as landlords. The newcomers were mainly from the Scottish Lowlands, and at first the English feared them almost more than they feared the Irish. In an attempt to counter their influence, the city of London was involved in the Ulster plantation, its name being combined with that of the historic ecclesiastical settlement of Derry to provide the name of a plantation. The Presbyterianism of the Scottish immigrants was successfully kept at bay until the time of the English Civil War; the Anglican bishoprics in Ireland were well endowed and powerful, and it was not until 1643 that the first presbytery was established in Belfast.

In the Parliament of 1613–15, summoned to ratify the Ulster plantation, a small Protestant majority was achieved because many new boroughs had been created in the newly planted areas. But government was concerned more with the appearance than the reality of consent, and no Parliament was called again until 1633. In the last years of James's reign, pressure from his Spanish and French allies caused him to concede toleration to the Roman Catholics; and from 1618 a Roman Catholic hierarchy was in residence in Ireland.



The percentage of land, by county, owned by Roman Catholics (*i.e.*, the Irish natives) in 1641, 1688, and 1703. The average percentage for all of Ireland is indicated after the year identifying each map.

From A. Orme, *Ireland*

Charles I (1625–49) and the Commonwealth (1649–60). Charles I conceived the idea of raising armies and money in Ireland in return for religious concessions, known as "the Graces," by which Roman Catholics were allowed to engage in various public activities. But this policy was abandoned by Thomas Wentworth, later earl of Strafford, lord deputy from 1633 to 1640. He set himself to break the power of the great and of trade monopolists, both Irish and English, including the London city companies. He induced the Roman Catholic members of the Irish House of Commons to join in voting large subsidies in the hope of obtaining further concessions but then abolished most of the existing Graces. He thus seriously weakened the loyalty to the crown of the old landowning classes; and later all his enemies in Ireland joined with those in England in bringing about his execution (1641). His Irish Army was disbanded, and control of the Irish government passed to Puritan lords justices.

Ireland during the English Civil War and its aftermath

A general rising of the Irish in Ulster was almost inevitable. It took place in October 1641, and thousands of colonists were murdered or fled. A Roman Catholic confederacy was formed at Kilkenny in 1642, but it did not succeed in welding together the various groups of which it was composed. During the period of the English Civil War there were Irish confederate armies in Ulster and in Leinster; English parliamentary armies operated in the north and south; and Dublin was held by James, duke of Ormonde, commanding an army of Protestant royalists. Negotiations for peace between Ormonde and the confederates were difficult and protracted; and in 1646, when it was clear that Charles I's cause was lost, Ormonde surrendered Dublin to a parliamentary commander. The confederates in isolation could offer little resistance (1649–50) to Oliver Cromwell. By 1652 all Irish resistance was over.

During the Commonwealth and Protectorate, authority in Ireland was exercised by parliamentary commissioners and chief governors. A union of the three kingdoms of England, Scotland, and Ireland, effected in 1653, resulted in Irish representatives attending Parliaments held in London in 1654, 1656, and 1659. By an Act of Settlement, Ireland, regarded as conquered territory, was parceled out among soldiers and creditors of the Commonwealth, and only those Irish landowners able to prove their constant support of the Parliamentary cause escaped having their estates confiscated. Of these, those who were Roman Catholics were still obliged to exchange land owned to the northeast or south of the River Shannon for land in Connaught. Roman Catholics and Anglicans were forbidden to practice their religion, but the campaign against Irish Catholicism was not successful. After the Restoration (1660) Charles II personally favoured complete religious toleration, but the forces of militant Protestantism sometimes proved too strong for him. The Commonwealth parliamentary union was, after 1660, treated as null and void.

The Restoration period and the Jacobite war. Most significant of the events of the Restoration was the second Act of Settlement (1662), which enabled Protestant loyalists to recover their estates. The Act of Explanation (1665) obliged the Cromwellian settlers to surrender one-third of their grants, providing a reserve of land from which Roman Catholics were partially compensated for losses under the Commonwealth. This satisfied neither group. Roman Catholics were prevented from residing in towns, and local power, in both borough and county, became appropriated to the Protestant interest. But Protestantism itself became permanently split; as in England, the Presbyterians refused to conform to Episcopalian order and practice and, in association with the Presbyterians of Scotland, organized as a separate church.

Under James II, advantage was taken of the King's Catholicism to reverse the tendencies of the preceding reign. After his flight from England to France (1688), James crossed to Ireland, where, in a Parliament, the Acts of Settlement and Explanation were repealed and provision was made for the restoration of expropriated Roman Catholics. When William III landed in Ireland to oppose James, the country divided denominationally, but the real issue was not religion but the land. After his defeat at the Battle of the Boyne (1690) James fled to France, but his Roman Catholic supporters continued in arms until defeated at Aughrim and obliged to surrender (1691) at Limerick. There, however, they secured either the right to go overseas or, if they accepted William's regime, immunity from discriminatory laws. But civil articles to secure toleration for the Roman Catholics were not ratified, thus enabling later Irish leaders to denounce the "broken treaty" of Limerick. Immediately after Limerick, the Protestant position was secured by acts of the English Parliament declaring illegal the acts of King James's Parliament in Ireland and restricting to Protestants mem-

bership of future Irish Parliaments. The sale of the lands forfeited by James and some of his supporters further reduced the Roman Catholic landownership in the country; by 1703 it was less than 10 percent. On this foundation was established the Protestant Ascendancy.

The Protestant Ascendancy

The 18th century. The Protestant Ascendancy was a supremacy of that proportion of the population, about one-tenth, which alone belonged to the established Protestant Episcopalian Church. They celebrated their position as a ruling class by annual recollections of their victories over their hated popish enemies. Not only the Roman Catholic majority but also the Presbyterians and other Nonconformists, whose combined numbers exceeded those of the church establishment, were excluded from full political rights, notably by the Test Act of 1704, which made tenure of office dependent upon willingness to receive communion according to the Protestant Episcopalian (Church of Ireland) rite. Owing to their banishment from public life, the history of the Roman Catholic Irish in the 18th century is concerned almost exclusively with the activities of exiled soldiers and priests, many of whom distinguished themselves in the service of continental monarchs. Details of the lives of the unrecorded Roman Catholic majority in rural Ireland can be glimpsed only from ephemeral literature in English and from the Gaelic poetry of the four provinces.

The Protestant Ascendancy of 18th-century Ireland began in subordination to that of England but ended in asserting its independence. In the 1690s commercial jealousy compelled the Irish Parliament to destroy the Irish woolen export trade, and in 1720 the Declaratory Act affirmed the right of the British Parliament to legislate for Ireland and transferred to the British House of Lords the powers of a supreme court in Irish law cases. By the end of the first quarter of the 18th century, resentment at this subordination had grown sufficiently to enable the celebrated pamphleteer Jonathan Swift to whip up a storm of protest over the affair of "Wood's halfpence." William Wood, an English manufacturer, had been authorized to mint coins for Ireland; the outcry against alleged exploitation of the lesser country by arbitrary creation of a monopoly became so violent that it could be terminated only by withdrawing the concession from Wood.

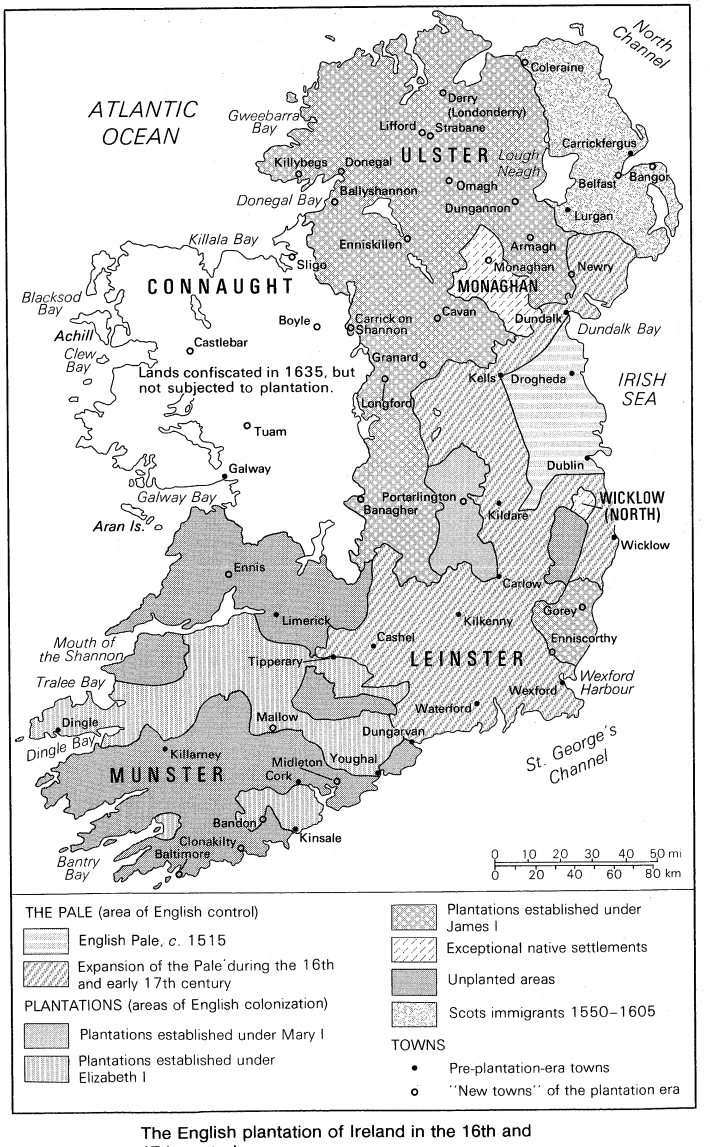
Nevertheless, it was another 30 years before a similar protest was made. In 1751 a group of patriots organized themselves to defeat government resolutions in the Irish Parliament appropriating a financial surplus as the English administrators rather than the Irish legislators saw fit. Although in 1768 the Irish Parliament was made more sensitive to public opinion by provision for fresh elections every eight years instead of merely at the beginning of a new reign, it remained sufficiently controlled by government to pass sympathetic resolutions on the revolt of the American colonies.

The American Revolution greatly influenced Irish politics, not least because it removed government troops from Ireland, and the Protestant Irish volunteer corps, spontaneously formed to defend the country against possible French attack, exercised a coercive influence for reform. A patriotic opposition, led by Henry Flood and Henry Grattan, began an agitation that led in 1782 to the repeal of the Declaratory Act of 1720, and to an amendment of Poynings' Law to give legislative initiative to the Irish Parliament. In this period many of the disabilities suffered by Roman Catholics in Ireland were abolished, and in 1793 the British government, seeking to win Roman Catholic loyalty on the outbreak of war against Revolutionary France, gave them the franchise and admission to most civil offices. The government further attempted to conciliate Roman Catholic opinion in 1795 by founding the seminary of Maynooth to provide education for the Roman Catholic clergy. But the Protestant Ascendancy had become concerned about its position and resisted efforts to make the Irish Parliament more representative. The outbreak of the French Revolution had effected a temporary alliance between an intellectual elite among the Presbyterians and leading middle-class Roman Catholics; these, under the inspiration of Wolfe Tone, founded societies of United Irishmen, a series of radical political clubs. After the outbreak of war, the societies, reinforced by agrarian malcontents, were driven underground. In despair they sought the military support of Revolutionary France, which, between 1796 and 1798, dispatched a series of abortive naval expeditions to Ireland. The United Irishmen were preparing for rebellion; it broke out in May 1798 but was widespread only in Ulster and in Wexford in the south. Although the rebellion was unsuccessful, it brought the Irish question forcibly to the attention of the British Cabinet, and the prime minister William Pitt planned and carried through an amalgamation of the British and Irish Parliaments, merging the kingdoms of Great Britain and Ireland into the United Kingdom. Despite substantial opposition in the Irish Parliament, the measure passed into law, having effect from Jan. 1, 1801. To Grattan and his supporters the Union seemed the end of the Irish nation; the last protest of the United Irishmen was made in Robert Emmet's abortive rebellion of 1803.

The Union with Great Britain

Social, economic, and cultural life in the 17th and 18th centuries. Although the late 16th century saw the destruction of Gaelic civilization in the upper levels of society, it was preserved for more than the next two centuries among the ordinary people of the northwest, west, and southwest, who continued to speak Irish and who maintained a way of life remote from that of the new landlord class.

The 17th-century confiscations made Ireland a land of great estates and, except for Dublin, of small towns decaying under the impact of British restrictions on trade.



The English plantation of Ireland in the 16th and 17th centuries.

Except on the Ulster plantations, the tenantry was relatively poor in comparison with that of England and employed inferior agricultural methods. Over large parts of the east and south, tillage farming had given way to pasturage. In the north of Ireland, a somewhat similar tendency, creating a decline in the demand for labour, led in the early 18th century to the migration of substantial numbers of Ulster Scots to North America. In Ulster there gradually emerged a tenantry who compelled their landlords to maintain them in their farms against the claims and bids of Roman Catholic competitors now once again legally entitled to hold land. This purpose immensely strengthened the Orange Order, founded in 1795 in defense of the Protestant Ascendancy. Increasingly it linked the Protestant gentry and farmers, while excluding Roman Catholics from breaking into this privileged ring. Tillage farming was maintained in Ulster more extensively than in the south and west, where there developed on the poorer lands a system of subdivision apparently necessitated by population increase. Little is known of the lives of the ordinary people, and even for those of the gentry the evidence, apart from that of estate records, is rarely extensive.

Little need be said of the culture of the Anglo-Irish in the same period, as it followed so closely the traditions of Britain and, very occasionally, those of the rest of Europe. Gradually, during the 18th century, the new landowning class developed some appreciation of the visual arts. But the really original achievement of the period was in literature, particularly in drama, where the rhetorical gifts of the people secured an audience. In this period there was a strong connection between rhetoric and the arts, as between oratory, themes of social decay, and the consoling power of language and form. Such works as Oliver Goldsmith's *Deserted Village* and *Traveller,* Edmund Burke's speeches, and the speeches and plays of Richard Brinsley Sheridan are manifestations of a rhetorical tradition central to Irish feelings.

The 19th and early 20th centuries. By the Act of Union it was provided that Ireland would have in the United Kingdom about one-fifth of the representation of Great Britain, with 100 members in the House of Commons. The union of the churches of England and Ireland as the established denominations of their respective countries was also effected, and the preeminent position in Ireland of Protestant Episcopalianism was further secured by the continuation of the British Test Act, which virtually excluded Nonconformists (both Roman Catholic and Protestant) from Parliament and from membership of municipal corporations. Not until 1828–29 did the repeal of the Test Act and the concession of Roman Catholic emancipation provide political equality for most purposes. It was also provided in the Union that there should be free trade between the two countries and that Irish merchandise would be admitted to British colonies on the same terms as British merchandise. But these advantages were not enough to offset the disastrous effect upon Ireland of exposure to the full impact of Britain's industrial revolution. Within half a century agricultural produce dropped in value and estate rentals declined, while the rural population increased substantially. When the potato, the staple food of rural Ireland, rotted in the ground through the onset of blight in the mid-19th century, perhaps as many as a million people died of starvation and fever in the Great Famine that ensued, and even more fled abroad.

At first, and perhaps for more than one-third of the 19th century, the auguries of success for the Union were favourable. After the defeat of Napoleon at Waterloo (1815), political discontent increased but became concentrated, so far as the Roman Catholics were concerned, on securing their emancipation. Until this was achieved there had not clearly emerged any notable difference in outlook between the Roman Catholics and Presbyterians; but the dramatic manner in which the Roman Catholic Daniel O'Connell was elected to a Parliamentary seat for County Clare (1828), subsequently sweeping the emancipation movement into victory, provoked a panic among timid Protestants and led to an alliance between the Presbyterians and their old oppressors, the Protestant Episcopalians.

Emergence of Daniel O'Connell

After emancipation, the middle-class Roman Catholics and Protestants drifted apart, the latter increasingly clinging to the Union, the former more slowly but at last decisively coming to seek its repeal.

O'Connell's adherence to the cause of repeal did not prevent him from participating actively in British politics. Lord Melbourne, the Whig prime minister, by a bargain known as the Lichfield House Compact (1835), secured O'Connell's support in return for a promise of "justice for Ireland." But meanwhile the Tories, led by Sir Robert Peel, exercised through their control of the House of Lords an effective restriction on promised social and economic reforms for Ireland, and when Peel returned to power in the early 1840s O'Connell, despairing of further concessions, began a massive campaign outside Parliament for repeal of the Union, notably by organizing large popular demonstrations. A climax was reached in October 1843 when troops and artillery were called out to suppress the mass meeting arranged at Clontarf outside Dublin. O'Connell's method of popular agitation within the law proved unavailing, however, and his influence thereafter rapidly declined.

Associated with O'Connell's repeal agitation was the Young Ireland movement, a group connected with a repeal weekly newspaper, *The Nation,* and led by its editor Charles Gavan Duffy, its chief contributor Thomas Osborne Davis, and its special land correspondent John Blake Dillon. But they became increasingly restless at O'Connell's cautious policy after Clontarf and in 1848 became involved in an abortive rising. Its failure, and the deportation or escape from Ireland of most of the Young Ireland leaders, destroyed the repeal movement.

For about 20 years after the Great Famine, political agitation was subdued, while emigration continued to reduce the population every year. The landowners had also suffered severely from inability to collect rents, and there was a wholesale transfer of estates to new owners. Evictions were widespread, and cottages were demolished at once by the landlords to prevent other impoverished tenants from occupying them. The flow of emigrants to the United States was encouraged by invitations from Irishmen already there. And in England, also, the new industrial cities and shipping centres attracted large settlements of hopeless and embittered exiles from Ireland.

Rise of Fenianism. Among the exiles both in the United States and in England, the Fenian movement spread widely. A secret revolutionary society named for the Fianna, the Irish armed force in legendary times, it aimed at securing Ireland's political freedom by exploiting every opportunity to injure English interests.

Imprisonment of Fenian leaders

In Ireland, Fenian ideals were propagated in the newspaper *The Irish People;* and in 1865 four Fenian leaders, Charles Kickham, John O'Leary, Thomas Clarke Luby, and Jeremiah O'Donovan Rossa were sentenced to long-term imprisonment for treasonable writing in it. During the next two years, plans gradually developed for a projected nationwide rising, financed largely by funds collected in the United States. It took place in March 1867 but was easily crushed and its leaders imprisoned. The prime minister, William Ewart Gladstone, at last recognized the necessity for drastic Irish reforms, disestablishing the Protestant Church in Ireland in 1869 and in 1870 introducing the first Irish Land Act, which conceded the principles of secure tenure and compensation for improvements made to property. He may also have been concerned at the cleavage between English and Irish public opinion caused by the execution at Manchester of William P. Allen, Michael Larkin, and Michael O'Brien for involvement in a Fenian prisoner-rescue operation that resulted in the shooting of a British police sergeant. To most Englishmen the "Manchester murderers" richly deserved their fate; to most Irishmen they were the "Manchester martyrs," celebrated in ballad and legend.

The Home Rule movement and the Land League. Soon afterward, in 1870, a constitutional movement, the Home Government Association (Home Rule League), was founded by Isaac Butt, a prominent Unionist lawyer interested in land reform. In the election of 1874 it returned about 60 members to Parliament. The movement was

tolerated rather than encouraged by the various groups of Irish nationalists, and it was not fully supported by the Roman Catholic clergy until the 1880s.

A return of bad harvests in 1879 brought new fears of famine, and Michael Davitt founded the Irish Land League, seeking to achieve for tenants security of tenure, fair rents, and freedom to sell property. A formidable agrarian agitation developed when Davitt was joined by Charles Stewart Parnell, a young landowner and member of Parliament in the Home Rule Party, which soon elected him as its leader in place of Butt. Parnell undertook a tour of North America to raise funds for the Land League; there he was influenced by two Irish Americans, John Devoy, a leading member of Clan na Gael, an effective American Fenian organization, and Pat Ford, whose New York paper *The Irish World* preached militant Socialism and hatred of England. At Westminster Parnell adopted a policy of persistent obstruction, which compelled attention to Irish needs by bringing parliamentary business to a standstill. Gladstone was forced to introduce his Land Act of 1881, conceding fixity of tenure, fair rents, and free sale of the tenant's interest.

Parnell's success was not achieved without serious difficulties, including the ultimate proscription of the Land League by the government and the imprisonment of its leaders. As a result, Parnell used his parliamentary party, then increased to 86, to defeat and thus dismiss from office Gladstone's Liberal government, already unpopular in England as a result of its failure to relieve General Gordon at Khartoum in 1884. For a while Conservatives and Liberals both negotiated with Parnell, but ultimately Gladstone became converted to Home Rule, introducing a bill to bring it into effect after he returned to office in 1886. The bill, however, was defeated by a combination of Conservative Unionists influenced by Irish Orangemen and splinter groups from the Liberal Party. There followed 20 years during which Irish Nationalist ambitions seemed frustrated, partly because Conservative-Unionists were mainly in power and partly because bitter internal rivalries discredited the Irish Nationalist Party after Parnell's involvement (1889) in a divorce suit. Meanwhile, Gladstone's second Home Rule Bill (1893) was rejected in the House of Lords. Only in 1900 was a Parnellite, John Redmond, able to reunite the Nationalists. In the last years of the century, partly in reaction to political frustrations, a cultural nationalist movement developed, led by Douglas Hyde and Eoin MacNeill. Through the Gaelic League much was done to revive interest in the speaking and study of Irish. These cultural movements were reinforced by others, such as that of the Sinn Féin ("Ourselves") movement led by Arthur Griffith, who preached a doctrine of political self-help. It subsequently emerged that a Fenian organization, the Irish Republican Brotherhood, had revived and was secretly recruiting membership through various cultural societies and through the Gaelic Athletic Association, founded to promote specifically Irish sports.

Gladstone's Home Rule bill

Around the close of the century the Conservatives initiated a policy designed to "kill Home Rule by kindness" by introducing constructive reforms in Ireland. Their most important achievement in this field was the Wyndham Land Purchase Act of 1903. This, by providing generous inducements to landlords to sell their estates, effected by government mediation the transfer of land ownership to the occupying tenants.

The 20th-century crisis. After the great Liberal victory of 1906, Redmond decided to force the Liberals to revive Home Rule, and when David Lloyd George's radical budget provoked a collision with the House of Lords in 1909, he seized his opportunity. He agreed to support the campaign of the prime minister, H.H. Asquith, against the Lords in return for the promise of a Home Rule bill. The reduction of the power of the Lords by the 1911 Parliament Act seemed to promise success for the third Home Rule bill, introduced in 1912. But in the meantime the Irish Unionists, under their colourful leader Sir Edward Carson, had mounted an effective countermovement, backed by most of the British Unionists. Thousands of Ulstermen signed the Ulster Covenant of resistance to Home Rule (1912), and Carson announced that a provisional government would be formed there. At first planning to reject Home Rule for all Ireland, the Unionists gradually fell back on a demand for Ulster to be excluded from its scope. Redmond's claim that there was "no Ulster question," implying that even among the Ulster members of Parliament there was a majority for Home Rule, hardened the Protestant and Unionist resistance in the areas around Belfast. Down, Antrim, Armagh, and Derry all contained Unionist majorities; Donegal, Cavan, and Monaghan had strong Home Rule majorities; and Tyrone and Fermanagh had small Home Rule majorities. The Ulster Volunteer Force was organized and boasted of active sympathy among army officers; their boasts became formidable when all the officers in the cavalry brigade at the Curragh suddenly announced (March 1914) that they would resign if ordered to move into Ulster. Meanwhile, to counter the Ulster Volunteers, a nationalist force, the Irish Volunteers, had been launched in Dublin in November 1913. Both forces gathered arms, and Ireland was on the verge of civil war when World War I broke out. Assured of Redmond's support in recruiting for the army, Asquith enacted Home Rule but followed this with a Suspensory Act, delaying implementation until the return of peace.

Meanwhile in Ireland, the revolutionary element gained support from those alienated by Redmond's pro-British attitude. Before the end of 1914 the Irish Republican Brotherhood had made full plans for a revolutionary outbreak. Sir Roger Casement went to Germany to solicit help, but he obtained only obsolete arms and was himself arrested on his return to Ireland on April 21, 1916. When the rising took place three days later, on Easter Monday, only about 1,000 of the small force available were actually engaged. A provisional Republican government was proclaimed. The General Post Office and other parts of Dublin were seized; street fighting continued for about a week until Patrick Pearse and other Republican leaders were forced to surrender. Their subsequent execution aroused Irish public opinion and led to the defeat and virtual extinction of Redmond's constitutional party at Westminster in the general election of December 1918. Their successful opponents, calling themselves Sinn Féin but supporting the Republican program announced in 1916, were led by Eamon de Valera, a surviving leader of the Rising. Again the Republicans set up their provisional government, elected by the Irish members of Parliament at a meeting in Dublin called Dáil Éireann, the "Irish Assembly." This provided a constitutional facade when British government was rapidly breaking down except in the northeastern counties. Simultaneously, the Irish Republican Army (IRA) was organized to resist British administration and to secure recognition for the Republican government. Its members soon engaged in widespread ambushes and attacks on barracks, while the government retaliated with ruthless reprisals. A large proportion of the Irish police resigned and were replaced by English recruits, known from their temporary uniforms as the Black and Tans.

The Easter Rising

In this condition of virtual civil war Britain gradually alienated Irish public opinion and was forced, partly under American influence, to pass the Government of Ireland Act (1920). By this measure Ireland was divided into two self-governing areas, Northern Ireland and Southern Ireland. Both were to enjoy, within the United Kingdom, limited powers of self-government. After a general election in Ireland, King George V opened the Parliament of Northern Ireland in Belfast (1921) and in his speech appealed for an end to fratricidal strife. The King's initiative forced the British prime minister Lloyd George to open negotiations with de Valera, but for some time progress proved impossible, because neither side would admit the other's legality. Ultimately, on Dec. 6, 1921, an Anglo-Irish treaty was signed on behalf of Great Britain by Lloyd George and leading members of his cabinet and on behalf of Ireland by Arthur Griffith, Michael Collins, and other members of the Republican Cabinet.

INDEPENDENT IRELAND

Establishment of the Irish Free State. The Anglo-Irish treaty provided that Ireland should in future have the

"same constitutional status in the community of Nations known as the British Empire as the Dominion of Canada, the Commonwealth of Australia, the Dominion of New Zealand, and the Union of South Africa with a parliament having powers to make laws for the peace, order and good government of Ireland and an Executive responsible to that parliament." The new dominion was to be known as the Irish Free State. This peace agreement, ratified by the British Parliament, became operative when it had also been passed (January 1922) by a meeting of the Dáil. The new state comprised only 26 of the 32 counties, the northeastern area, known as Northern Ireland, remaining separate.

But the terms of the treaty had been accepted by the Irish signatories only because Lloyd George had threatened war on Ireland if they were rejected. Particularly obnoxious were a prescribed oath of allegiance to the British crown and the provisions allowing Northern Ireland to remain outside the new state. De Valera and the Republicans immediately repudiated the treaty, and after its passage in the Dáil, de Valera resigned the presidency. Michael Collins, chairman of the provisional government set up according to the terms of the treaty, and Arthur Griffith, the new president, desired an immediate general election to obtain a verdict on the treaty; in the deteriorating conditions Collins eventually made with de Valera an agreement known as the Pact (May 20, 1922), in which it was settled that government and Republican candidates would not oppose each other and that de Valera would consider resuming office. But the Pact naturally could not bind other parties, and at the election (June 16) Republicans were ousted in favour of members of a labour party and a farmers' party and by independents, thus reducing the anti-treaty vote to a small minority. Before the Dáil could meet, civil war had broken out between the government and the Republicans, who were allegedly accessories to the assassination (June 22) in London of Field Marshal Sir Henry Wilson soon after his acceptance of the position of military adviser to the government of Northern Ireland. The Republicans in Dublin had occupied the Four Courts (central courts of justice) in April and in late June. Under pressure from Britain, which also provided military equipment, Collins ordered them to retire. Serious fighting ensued for a week, until the Courts were blown up, and Rory O'Connor, the Dublin Republican leader, surrendered. Meanwhile, de Valera, who had escaped to the southwest, was openly supporting the Republicans. Griffith and Collins decided that no further compromise was possible, and military operations were begun. The strain had told so heavily on Griffith that he died suddenly on August 12, while Collins, inspecting the military operations, was killed in an ambush on August 22.

Repudiation of the Anglo-Irish treaty

The provisional government had thus lost two of its most prominent leaders, and surviving ministers could not appear openly without armed protection. Moreover, there was urgency in that, by the terms of the treaty, the newly elected Dáil was required to frame its constitution before Dec. 6, 1922. It met on September 9, elected as the new president William Thomas Cosgrave, and, in the absence of the Republican deputies, quickly passed the clauses of the constitution defining the relations of the Free State with the British crown and outlining arrangements for imperial defense. Timothy Michael Healy, a veteran follower of Parnell who had later supported Sinn Féin, was then appointed governor general, and Cosgrave became president of the executive council. The new constitution was also ratified at Westminster.

The Cosgrave ministries. Both before and after the ratification of the constitution the government resorted to strong measures to quell disorder and violence. Its decision to execute those found in unauthorized possession of firearms embittered Irish politics for years afterward. Numerous Republican insurgents were also imprisoned, and 77 were executed. Although Republican opposition was at first more bitter than ever, it eventually became less well organized, and by May 1923, on de Valera's recommendation, armed resistance ended.

At the end of August 1923 the fourth Dáil was elected, on a basis of adult suffrage for men and women. De Valera retained his personal following, and his party won 44 seats out of 128. Cosgrave's party won less than half the total number of seats, but as the Republicans refused to sit in the new Dáil, he had a majority among those who did attend. The absence of any effective opposition party greatly strengthened the power of the new government, and in the following years it displayed great energy. Despite initial economic difficulties, it pursued an efficient farming policy and carried through important hydroelectric projects. Government was increasingly centralized, with the elimination of various corrupt borough corporations; Kevin O'Higgins, as minister for justice, carried through many judicial reforms, and an efficient civil service was organized, in which competence in the Irish language became an essential qualification for promotion.

The fourth Dáil

The Anglo-Irish treaty of 1921 had provided that, if Northern Ireland did not enter the Free State, a boundary commission must establish the frontier between the two countries. Two of the six excluded counties, Tyrone and Fermanagh, contained clear, though small, Nationalist majorities, and the southern portions of both Down and Armagh had for years returned Nationalist members. Despite Northern Ireland's reluctance, a commission was established and sat in secret session during 1925. But it only recommended minor changes, which all three governments rejected as less satisfactory than maintaining the status quo.

In the general election of June 1927 Cosgrave's support in the Dáil was further reduced, but he nevertheless formed a new ministry, in which O'Higgins became vice president of the Executive Council. O'Higgins' assassination on July 10 suddenly revived old feuds, and Cosgrave passed a Public Safety Act, declaring all revolutionary societies treasonable. He forced the Republicans to acknowledge allegiance to the crown before being seated in the Dáil, though de Valera decried the oath as an "empty political formula." Shortly thereafter the Republicans, allied with the Labour Party and the National League, almost defeated Cosgrave, who thereupon dissolved the Dáil. In new elections Cosgrave won 61 seats as compared to Fianna Fáil's 57 and again formed a ministry. In the depression of the early 1930s, unemployment and general discontent with the government led to its defeat in February 1932. Fianna Fáil won enough seats for de Valera, with Labour Party support, to be able to form a new government.

De Valera's governments (1932–48) and the establishment of Éire. De Valera entered office with a policy of encouraging industry and improving the social services. He abolished the oath of allegiance to the crown and also stopped payment to Britain of interest on the capital advanced under the Land Purchase Acts of the late 19th and early 20th centuries. This refusal led to a tariff war with Britain. The country endorsed his policies in January 1933 by returning him to the Dáil with 77 seats and the support of the Labour Party.

De Valera invoked the Public Safety Act against the Blue Shirts, a quasi-Fascist body organized (by Cosgrave's militant supporters) allegedly to protect Treatyites from Republican extremists at public meetings.

De Valera introduced proposals for a new constitution in 1937. The power of the crown was ended, and the office of governor general was replaced by that of a president elected by national suffrage. The first president was Douglas Hyde, a Celtic scholar who had been associated with the Gaelic revival since 1890. The new constitution did not proclaim an independent republic, but it replaced the title of the Irish Free State with the word Éire (Ireland). The new constitution was ratified by a plebiscite at the 1937 general election (in which de Valera was again victorious) and became operative on Dec. 29, 1937.

Establishment of Éire

An agreement in April 1938 ended British occupation of three naval bases that had been left in British hands by the Anglo-Irish treaty of 1921. The dispute over the land-purchase annuities was settled, and the economic warfare was abated. In this period government relations with the radical Republicans deteriorated; members of the IRA suspected of complicity in shootings were imprisoned, and some were executed.

At the outbreak of World War II, de Valera renewed his

statement, made in 1938, that Ireland would not become a base for attacks on Great Britain. His government, re-elected in 1943 and 1944, remained strictly neutral, despite German air raids on Dublin in 1941 and, after the United States entered the war in December 1941, pressure from U.S. president Franklin D. Roosevelt.

The Republic of Ireland. In the general election of 1948 Fianna Fáil won 68 of the 147 seats in the Dáil, but de Valera refused to enter a coalition. John A. Costello emerged as the leader of a bloc composed of his own party, Fine Gael, and several smaller groups. Out of office, de Valera toured the world advocating the unification and independence of Ireland. Fearful of de Valera's prestige, Costello introduced in the Dáil the Republic of Ireland Act, which ended the fiction of Commonwealth membership that had been maintained since 1937. Britain recognized the status of Ireland in 1949 but declared that cession of the six counties could not occur without consent of the Parliament of Northern Ireland. Economic difficulties and a controversy between the ministry and the Roman Catholic hierarchy over the Public Health Act weakened Costello's government, and after the general election of 1951, de Valera again became prime minister. His ministry aroused sufficient discontent for Costello to be returned to power in 1954, but economic troubles enabled Fianna Fáil to win a majority in 1957. This was to be de Valera's last administration. He retired as prime minister in 1959 and was elected to the presidency, serving until 1973.

(R.W.D.E./J.O'B.R./Ed.)

Economic growth. Sean Lemass, prime minister from 1959 to 1966, initiated measures to stimulate Ireland's seriously stagnating economy. Under the First Programme for Economic Expansion (1958–63), economic protection was dismantled and foreign investment was encouraged; a growth rate that was planned to reach 2 percent actually reached 4 percent. This prosperity brought profound social and cultural changes to what had been one of the poorest and most backward countries in Europe. Emigration substantially declined, consumer spending increased, and religious social teaching was challenged and often set aside by newly affluent Irish men and women.

Ireland, like Britain, suffered setbacks in attempting to join the European Economic Community (EEC), but both nations formally entered that organization on Jan. 1, 1973. In elections held later that year, the Fianna Fáil government of Jack Lynch (prime minister from 1966) was defeated by a Fine Gael–Labour coalition led by Liam Cosgrave. The oil crisis and recession of 1974–75 forced the imposition of deflationary economic policies, a wealth tax, and attempts to tax farmers' incomes. Jack Lynch returned to power in 1977, and the Fianna Fáil proposed an ambitious economic policy based on tax cuts and the creation of new enterprises by foreign borrowing.

Despite a brief boom, serious problems became evident in the Irish economy by 1980. These included declining agricultural prices, rising imports and rising prices for imported oil, only a small increase in output, and a rapidly growing population, nearly one-half of which was under 25 years of age. Foreign borrowing had increased, while unemployment and inflation rose steeply.

Unstable politics of the 1980s

The early 1980s proved politically volatile for the Irish republic. In the general election of 1981 no clear majority resulted, and Garret FitzGerald of the Fine Gael–Labour coalition took power, ousting Charles Haughey, who had succeeded Jack Lynch as Fianna Fáil prime minister in 1979. The major issues of the campaign were economic policy, including the imposition of a wealth tax, and the removal of a constitutional ban on divorce. The budget of the coalition government was defeated in January 1982, and another general election in February returned Fianna Fáil and Charles Haughey to power. The life of this government, too, proved short and uneasy. In the face of a budget deficit equal to 7 percent of the gross national product, a program of severe public spending cuts was introduced. The government was defeated on a no-confidence vote in November and another general election—the third in 18 months—followed. This time a Fine Gael–Labour coalition under the leadership of FitzGerald secured an overall majority of the vote.

By the mid-1980s the economy showed faint signs of improvement. Inflation was at its lowest level in nearly two decades, helped by lower oil prices. The budget deficit and high level of unemployment, however, continued to pose considerable problems. Emigration, a barometer of Irish economic health, again began to increase in the mid-1980s. The prolonged recession had once again brought to the surface doubts and anxieties about the future of the Irish state and its real independence.

Social changes. The close relationship between the Irish republic and the Roman Catholic Church was highlighted by the visit of Pope John Paul II to Ireland in 1979, the first visit there of a reigning pontiff. The relationship was to be tested in the 1980s, however, as attempts were made to alter Irish law in its relation to Roman Catholic doctrine. In 1983, after a campaign from Roman Catholic pressure groups, a referendum was held to approve a draft constitutional amendment reinforcing the republic's existing ban on abortion. After a divisive campaign less than 55 percent of the electorate participated in the referendum, and only 36 percent of the total electorate voted in favour of the amendment, which nonetheless was enough for approval.

In 1985 the Roman Catholic Church strenuously, but futilely, opposed the government's liberalization of legislation concerning contraception. Church–state relations were again tested the following year by a referendum to remove the constitutional ban against divorce. The government, which favoured removal of the ban, was defeated by 39 percent of the electorate.

Relations with Northern Ireland. During the late 1950s and early 1960s the Irish government had had to deal with IRA attacks on British Army posts along the Ulster border. An attempt to ease cross-border tensions was made in 1965, when Ireland's prime minister, Sean Lemass, and Northern Ireland's prime minister, Terence O'Neill, exchanged visits. In 1970, the republic's Prime Minister Lynch dismissed two Cabinet ministers following an attempt to import arms for use in Northern Ireland.

The Irish government was increasingly preoccupied by the situation in Northern Ireland during the 1970s. In 1973 Prime Minister Cosgrave participated in talks with Edward Heath, prime minister of Britain, and representatives of Northern Ireland, resulting in the Sunningdale Agreement. This accord recognized that the North's relationship with Britain could not be changed without agreement of a majority of the population in the North, and it provided for the establishment of a Council of Ireland composed of members from both the Dáil and the Northern Ireland Assembly. The agreement collapsed the next year.

Although the republic was little affected by the violence in Ulster, there were a number of serious terrorist incidents. The murder of the British ambassador in Dublin in 1976 led to a state of emergency and the unpopular measure of strengthening emergency powers legislation.

In 1981 Prime Minister FitzGerald launched a constitutional crusade to make the reunification of Ireland more attractive to Ulster Unionists. At the end of the year the Irish and British governments set up an Anglo-Irish intergovernmental council to discuss matters of common concern, especially security. The report of the New Ireland Forum—a discussion group that included representatives of the political parties in Ireland and Northern Ireland—in 1984 set out three possible frameworks for political development in Ireland, those of a unitary state, a federal state, and joint sovereignty. Fianna Fáil preferred a unitary state, while Fine Gael and Labour preferred the federal solution. In November 1985 at Hillsborough in Northern Ireland, Ireland and Britain again agreed that any change in the status of Northern Ireland would come about only with the consent of the majority of the people of Northern Ireland, and an intergovernmental conference was established to deal with political, security, and legal relations between the two parts of the island.

The Anglo-Irish Agreement

In February 1987 Fianna Fáil returned to power under Charles Haughey, but without an overall majority. The new Progressive Democrat Party, formed in 1985 by men and women who had split from Fianna Fáil, made a strong showing. Sinn Féin, the political wing of the

Provisional IRA—following a decision in November 1986 to drop its abstentionist policy and contest future Dáil elections, standing on a socialist and pro-IRA platform—failed to win a seat. Shortly after the election the leader of Fine Gael and former prime minister Garret FitzGerald resigned and was succeeded by Alan Dukes.

The new government embarked on a program of comprehensive public spending cuts. This secured the support of Fine Gael and the Progressive Democrats, but opinion within the Labour Party, traditionally committed to high public expenditure programs, was more critical.

In May 1987, following a Supreme Court decision, a constitutional referendum ratified the Single European Act and thus served to confirm Ireland's participation in the European Communities (EC). The act legislated for the harmonization of social and fiscal measures within the EC.

Despite Fianna Fáil's initial criticism of the 1985 Anglo-Irish Agreement, while in power the Haughey government maintained support for the agreement. Contacts between the Irish and British governments continued after February 1987 within the formal structure of the intergovernmental conference. Fears that the violence in Northern Ireland would spill into the republic as a consequence of closer Anglo-Irish cooperation in the wake of the agreement proved unfounded. (J.O'B.R.)

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

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 921, 923, 961, 963, and 972.

BIBLIOGRAPHY

Physical and human geography: (*General works*): A good introduction is IRELAND, DEPARTMENT OF FOREIGN AFFAIRS, *Facts About Ireland,* 6th ed. (1985); see also IRISH NATIONAL COMMITTEE FOR GEOGRAPHY, *Atlas of Ireland* (1979). Other basic works include BASIL CHUBB (ed.), *A Source Book of Irish Government*, rev. ed. (1983), a selection of official information; *Encyclopaedia of Ireland* (1968), with articles on a wide variety of subjects by experts in their fields; T.W. FREEMAN, *Ireland: A General and Regional Geography,* 4th ed. (1969, reprinted in 1972); JAMES MEENAN, *The Irish Economy Since 1922* (1970), a well-documented, comprehensive survey; and J.W. O'HAGAN (ed.), *The Economy of Ireland: Policy and Performance,* 3rd ed. (1981). Historical, archaeological, topographical, and other information, with illustrations, is found in such guidebooks as LORD KILLANIN and MICHAEL V. DUIGNAN, *Shell Guide to Ireland,* 2nd rev. ed. (1967); and IAN ROBERTSON (ed.), *Ireland,* 5th ed. (1987). Pictorial coverage is provided in EDNA O'BRIEN and RICHARD FITZGERALD, *Vanishing Ireland* (1987). (*National culture*): For coverage of Irish culture over the centuries, see BRIAN DE BREFFNY (ed.), *Ireland: A Cultural Encyclopaedia* (1983); ROBIN FLOWER, *The Irish Tradition* (1947, reprinted 1978), a review of Gaelic Ireland's contribution to western European culture; MYLES DILLON (ed.), *Early Irish Society* (1954, reprinted 1969), six short, authoritative essays; E. ESTYN EVANS, *Irish Folk Ways* (1957, reprinted 1979); MÁIRTÍN Ó MURCHÚ, *The Irish Language* (1985), a beautifully illustrated overview of the historical development of the language; CATHAIR Ó DOCHARTAIGH, "Irish," in PETER TRUDGILL (ed.), *Language in the British Isles* (1984); and KENNETH HUDSON and ANN NICHOLLS, *The Cambridge Guide to the Museums of Britain and Ireland* (1987).

History: (*Historical surveys*): Comprehensive overviews of Ireland's history can be found in CHRISTOPHER HAIGH (ed.), *The Cambridge Historical Encyclopedia of Great Britain and Ireland* (1985); GIOVANNI COSTIGAN, *A History of Modern Ireland: With a Sketch of Earlier Times* (1969); RUTH DUDLEY EDWARDS, *An Atlas of Irish History,* 2nd ed. (1981); SÉAN P. Ó RÍORDÁIN, *Antiquities of the Irish Countryside,* 5th rev. ed., revised by RUAIDHRÍ DE VALERA (1979); R.B. MCDOWELL, *Irish Public Opinion, 1750–1800* (1943, reprinted 1975), *Public Opinion and Government Policy in Ireland, 1801–1846* (1952, reprinted 1975), and *The Irish Administration, 1801–1914* (1964, reprinted 1976), covering two centuries of developments in political and social life; ROBERT DUDLEY EDWARDS, *A New History of Ireland* (1972); JOHN O'BEIRNE RANELAGH, *A Short History of Ireland* (1983); J.C. BECKETT, *The Making of Modern Ireland, 1603–1923,* new ed. (1981); and A.T.Q. STEWART, *The Narrow Ground: Aspects of Ulster, 1609–1969* (1977). The turbulent history of the relationship between church and state in Ireland is the subject of KATHLEEN HUGHES, *The Church in Early Irish Society* (1966, reprinted 1980); LUDWIG BIELER, *Ireland, Harbinger of the Middle Ages* (1963, originally published in German, 1961), on the Irish religious missions and their influence in Europe; ROBERT DUDLEY EDWARDS, *Church and State in Tudor Ireland: A History of Penal Laws Against Irish Catholics, 1534–1603* (1935, reprinted 1972); TOM INGLIS, *Moral Monopoly: The Catholic Church in Modern Irish Society* (1987); and ROBERT G. CRAWFORD, *Loyal to King Billy: A Portrait of the Ulster Protestants* (1987).

(*Early and medieval history*): For special period studies see JAMES F. KENNEY, *The Sources for the Early History of Ireland: Ecclesiastical: An Introduction and Guide,* rev. ed. (1966, reprinted 1979), a treatment of religious manuscripts to the end of the 12th century; and RICHARD J. HAYES (ed.), *Manuscript Sources for the History of Irish Civilization,* 11 vol. (1965), an extensive index of manuscript sources on history from the 5th to the 20th century. Detailed histories include T.G.E. POWELL, *The Celts,* new ed. (1980, reprinted 1985); EOIN MACNEILL, *Celtic Ireland* (1921, reissued 1981); JAMES CHARLES ROY, *The Road Wet, the Wind Close: Celtic Ireland* (1986); GODDARD HENRY ORPEN, *Ireland Under the Normans,* 4 vol. (1911–20, reprinted 1968), which remains indispensable for the period of 1169–1333; A.J. OTWAY-RUTHVEN, *A History of Medieval Ireland,* 2nd ed. (1980), essential for the study of the period prior to 1496; ART COSGROVE (ed.), *Medieval Ireland, 1160–1534* (1987), a comprehensive selection of essays, in the "New History of Ireland" series; H.G. RICHARDSON and G.O. SAYLES, *The Administration of Ireland, 1172–1377* (1963), and *The Irish Parliament in the Middle Ages* (1952, reprinted 1964); RICHARD BAGWELL, *Ireland Under the Tudors: With a Succinct Account of the Earlier History,* 3 vol. (1885–90, reprinted 1965), and *Ireland Under the Stuarts and During the Interregnum* (1909–16, reprinted 1963).

(*The 18th and 19th centuries*): For the history of Ireland in the 18th and 19th centuries, see T.W. MOODY and W.E. VAUGHAN (eds.), *Eighteenth-Century Ireland, 1691–1800* (1986), a comprehensive history of the period; K.H. CONNELL, *The Population of Ireland, 1750–1845* (1950, reprinted 1975), a scholarly account; CECIL WOODHAM-SMITH, *The Great Hunger: Ireland 1845–1849* (1962, reissued 1987), on one of the worst disasters that influenced the course of Irish history; ROBERT DUDLEY EDWARDS and T. DESMOND WILLIAMS (eds.), *The Great Famine: Studies in Irish History, 1845–52* (1956, reissued 1976), another important source on the subject; and THOMAS GALLAGHER, *Paddy's Lament: Ireland 1846–1847: Prelude to Hatred* (1982, reissued 1987), on the repercussions of the potato famine on Irish society and its relations with England. Discussions of 19th-century Irish politics may be found in NICHOLAS MANSERGH, *The Irish Question, 1840–1921: A Commentary on Anglo-Irish Relations and on Social and Political Forces in Ireland in the Age of Reform and Revolution,* 3rd ed. (1975), an interpretative study of the political situation; KEVIN B. NOWLAN, *The Politics of Repeal: A Study in the Relations Between Great Britain and Ireland, 1841–50* (1965, reprinted 1975), a briefer examination of the problem; J.L. HAMMOND, *Gladstone and the Irish Nation* (1938, reprinted 1974), a standard work on Gladstone's Irish policy in 1869–93; JAMES LOUGHLIN, *Gladstone, Home Rule, and the Ulster Question, 1882–93* (1987), a later study that reconsiders the period; ANGUS MACINTYRE, *The Liberator: Daniel O'Connell and the Irish Party, 1830–1847* (1965); CONOR CRUISE O'BRIEN, *Parnell and His Party, 1880–90* (1964); F.S.L. LYONS, *Charles Stewart Parnell* (1977); R.F. FOSTER, *Charles Stewart Parnell: The Man and His Family,* 2nd ed. (1979); CONOR CRUISE O'BRIEN (ed.), *The Shaping of Modern Ireland* (1960, reprinted 1970); D.J. HICKEY and J.E. DOHERTY, *A Dictionary of Irish History Since 1800* (1980); and F.S.L. LYONS and R.A.J. HAWKINS (eds.), *Ireland Under the Union: Varieties of Tension* (1980), a collection of survey analyses.

(*The 20th century*): Twentieth-century developments are considered in F.S.L. LYONS, *Ireland Since the Famine,* 2nd rev. ed. (1973); DOROTHY MACARDLE, *The Irish Republic: A Documented Chronicle of the Anglo-Irish Conflict and the Partitioning of Ireland, with a Detailed Account of the Period 1916–1923* (1937, reissued 1965), a fascinating though partisan study; DEIRDRE MCMAHON, *Republicans and Imperialists: Anglo-Irish Relations in the 1930s* (1984), an authoritative account; J. BOWYER BELL, *The Secret Army: The IRA, 1916–1979,* rev. ed. (1979, reprinted 1983); PATRICK BISHOP and EAMONN MALLIE, *The Provisional IRA* (1987); TIM PAT COOGAN, *Ireland Since the Rising* (1966, reprinted 1976), a narrative work; and TERENCE BROWN, *Ireland: A Social and Cultural History, 1922 to the Present* (1985), an analysis of factors contributing to the sense of national identity since independence. Introductory surveys include OLIVER MACDONAGH, *Ireland: The Union and Its Aftermath,* rev. ed. (1977); JOHN A. MURPHY, *Ireland in the Twentieth Century* (1975); and DERVLA MURPHY, *Ireland* (1985), a pictorial work.

(F.H.B./R.W.D.E./J.O'B.R.)